Multivariate Analysis of Multiple Datasets: a Practical Guide for Chemical Ecology

Article *in* Journal of Chemical Ecology · March 2018

DOI: 10.1007/s10886-018-0932-6

3 authors, including:

Maxime Hervé
Université de Rennes 1
77 PUBLICATIONS **556** CITATIONS
SEE PROFILE

Florence Nicolè
Université Jean Monnet
44 PUBLICATIONS **879** CITATIONS
SEE PROFILE

Journal of Chemical Ecology
https://doi.org/10.1007/s10886-018-0932-6

Multivariate Analysis of Multiple Datasets: a Practical Guide for Chemical Ecology

Maxime R. Hervé[1] · Florence Nicolè[2] · Kim-Anh Lê Cao[3]

Received: 28 July 2017 / Revised: 3 February 2018 / Accepted: 4 February 2018

Abstract
Chemical ecology has strong links with metabolomics, the large-scale study of all metabolites detectable in a biological sample. Consequently, chemical ecologists are often challenged by the statistical analyses of such large datasets. This holds especially true when the purpose is to integrate multiple datasets to obtain a holistic view and a better understanding of a biological system under study. The present article provides a comprehensive resource to analyze such complex datasets using multivariate methods. It starts from the necessary pre-treatment of data including data transformations and distance calculations, to the application of both gold standard and novel multivariate methods for the integration of different omics data. We illustrate the process of analysis along with detailed results interpretations for six issues representative of the different types of biological questions encountered by chemical ecologists. We provide the necessary knowledge and tools with reproducible R codes and chemical-ecological datasets to practice and teach multivariate methods.

Keywords Discriminant analyses · Distance-based analyses · Integrative analyses · Metabolomics · Multi-block methods · Ordination methods

Introduction

Chemical ecology promotes an ecological and evolutionary understanding of the origin and the function of chemicals mediating interactions within and between organisms. There are a vast number of questions concerning the interactions between living organisms and their biotic and abiotic environment. These include the identification of plant compounds providing resistance to herbivores (e.g. Gatehouse 2002) or attracting pollinators (e.g. Raguso 2008), pheromones mediating sexual attraction (e.g. Archunan 2009), cuticular hydrocarbons mediating social interactions (e.g. Howard and Blomquist 2005) or secondary metabolites involved in allelopathy (e.g. Bais *et al*. 2006). Knowledge on chemical mediators may also help to improve our understanding of the mechanisms underlying tolerance to toxic biologically-produced xenobiotics (e.g. Després *et al*. 2007), of the impact of abiotic factors on ecological interactions through modifications of metabolite production/emission (e.g. Reudler and Elzinga 2015), or of the use of chemical profiles as a taxonomic tool (e.g. Volkman *et al*. 1998).

To answer these biological questions, various statistical analyses can be performed. The aim of the analysis can be straightforward, such as assessing the differences in chemical profiles between distinct groups (species, populations, treatments etc.). They may also be more challenging, such as evaluating the influence of a multifactorial experimental design on chemical variation. Other complex questions include the identification of regulating compounds in interacting partners, or metabolic pathways activated as a response to environmental stressors. Chemical data are often associated with other relevant data such as genetic, phylogenetic, phenotypic, geographic, environmental or genomics data. Due to the abundance and complexity of these different variables, choosing and applying the appropriate statistical method to answer a specific biological question can be challenging. It is

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10886-018-0932-6) contains supplementary material, which is available to authorized users.

✉ Maxime R. Hervé
maxime.herve@univ–rennes1.fr

[1] University of Rennes, Inra, Agrocampus Ouest, IGEPP - UMR-A 1349, F-35000 Rennes, France

[2] University of Lyon, UJM-Saint-Etienne, CNRS, LBVpam FRE 3727, EA 3061, F-42023 Saint-Etienne, France

[3] Melbourne Integrative Genomics and School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia

Published online: 25 February 2018

particularly difficult with the recent advent of metabolomics, the large-scale study of all detectable metabolites in a sample.

In the present article, we provide a practical guide for multi-dataset analyses in chemical ecology, with a primary focus on multivariate statistical methods. Multivariate methods have numerous appealing properties for data integration, but can be challenging to apply for the non-specialist. These methods are complementary to univariate statistical tests, such as the *t*-test or ANOVA. We refer the reader to the review from Saccenti *et al.* (2014) for a detailed explanation of the complementarity between univariate and multivariate methods, and the appropriate use of univariate tests. Our article does not aim to give an extensive introduction of all multivariate techniques proposed so far. It rather aims to introduce key concepts, along with useful methods deemed appropriate for answering the main questions encountered by chemical ecologists as well as in other fields of research where the analysis of quantitative multivariate data is required. We particularly focus on methods that are implemented and freely available in the R software (R Core Team 2016).

We first address the pre-treatment of the different types of chemical data for multivariate analyses. Next, we give a detailed description of the application, validation and interpretation of the different methods in relation to specific biological questions in chemical ecology. Each method is illustrated on a real dataset, and we provide fully reproducible R scripts using R markdown (Allaire *et al.* 2017).

Pre-Treatment of Chemical Data

Chemical or metabolomic data generally first need to be 'pre-processed' to obtain the final matrix with samples in lines, and compounds in columns. This pre-processing includes several steps such as peak picking, alignment, integration and normalization and results in 'clean' data ready for statistical analysis, or 'post-processing'. Pre-processing is the object of extensive literature reviewed by Engel *et al.* (2013) and is out of the scope of this paper. However, we detail the first step of post-processing, which takes place prior to statistical analysis and that we call 'pre-treatment'. Pre-treatment consists of data formatting, transformations and scaling. Depending on the chemical extraction method, apparatus and experimenter, chemical data can be expressed in three different formats: presence/absence of the compound (binary data), concentration or emission rate of the compound ('concentration-like' data), and relative proportion of the compound in the chemical blend (compositional data). These three types of data aim to address different biological questions and are therefore analyzed with different statistical methods. They also require different pre-treatment steps which we detail below and are summarized in Fig. 1.

Binary Chemical Data Fingerprinting binary data encode chemical composition in the presence or absence of compounds. This type of data is often generated by certain extraction techniques that do not faithfully reflect the concentration of the compounds in the original sample, e.g. Solid Phase MicroExtraction (SPME) in uncontrolled environment (Tholl *et al.* 2006). Binary data can be used for biomarker discovery but most often they are used to address questions on global similarity. To do so, all compounds are combined into a unique sample-to-sample similarity measure. In chemotaxonomy for example, metabolic fingerprinting is used to discriminate different species (Ivanišević *et al.* 2011).

A technical point that is not restricted to binary data but is of particular importance here is the 'double-zero problem' (Legendre and Legendre 2012), where a given compound is absent from two samples. Are these samples similar because they lack this compound? Among the many similarity indexes that exist for binary variables (Gower and Legendre 1986; Legendre and Legendre 2012), similarity indexes (or distances) that consider a double absence as a similarity are called 'symmetric'. The most popular symmetric similarity index is the simple matching coefficient (Sokal and Michener 1958). It consists in calculating the proportion of compounds that are present or absent in both samples. The most popular asymmetric index is the Jaccard's community coefficient (Jaccard 1901). It calculates the number of compounds present in both samples divided by the number of compounds present in at least one of the two samples. The choice of the type of similarity depends on the data and the biological question.

A similarity index results in a similarity matrix, whereas multivariate methods require a distance matrix as an input. The similarity matrix is thus transformed into a distance matrix using the basic formula: distance = 1 − similarity, or a derived version of this formula.

'Concentration-Like' Chemical Data We consider absolute concentrations or any measure directly proportional to absolute concentrations, e.g. peak area, 'concentration-like' chemical data. Transformation of such data is not always necessary, but it can improve the outputs from the multivariate analyses. This is especially the case for skewed distributions that are transformed into symmetric distributions. Transformations are also applied to convert multiplicative relationships between compounds into additive relationships to assess the effect of each compound independently (van den Berg *et al.* 2006). The logarithmic transformation is most popular, but the fourth root transformation is preferable when data include zero values (Wold *et al.* 2001).

Concentration-like data can also be centered and/or scaled. Centering is a recommended transformation in most situations (Wold *et al.* 2001). It results in a data set in which all compounds have the same zero mean, enhancing the focus on the differences between samples. Multivariate analyses give more

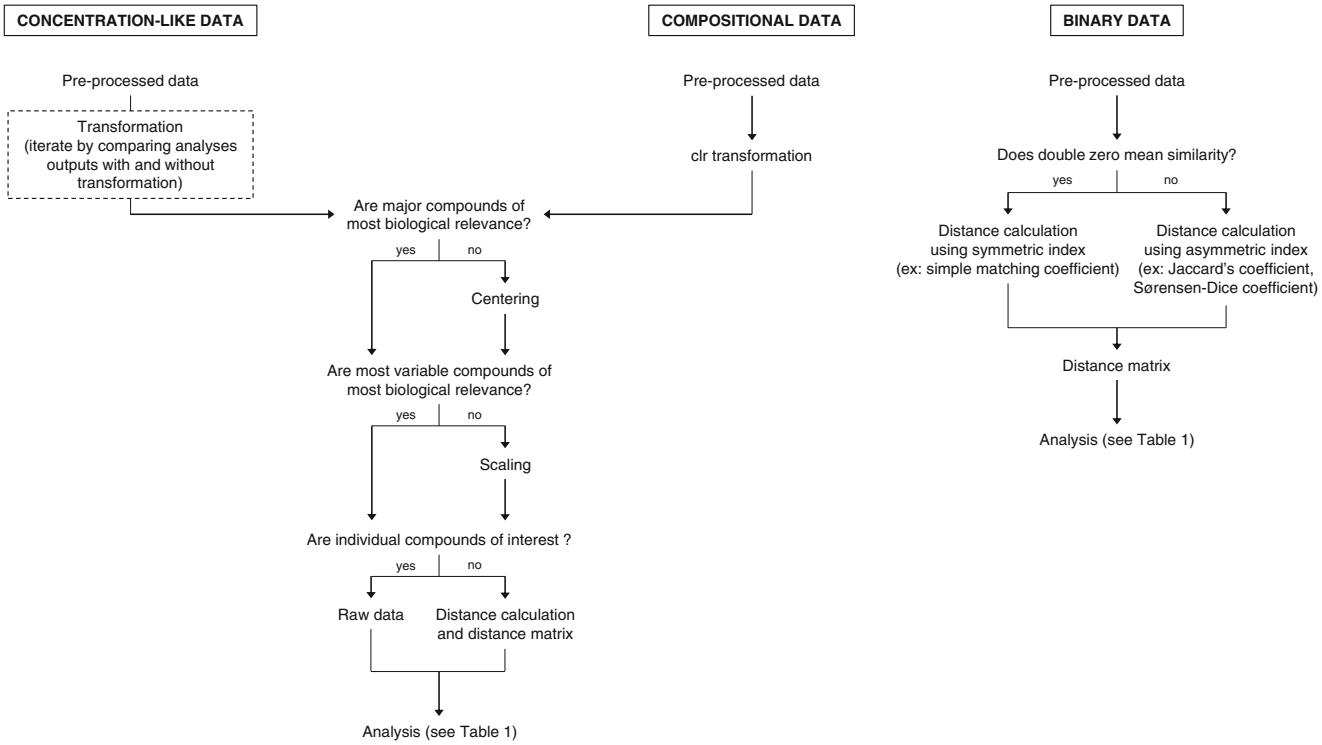


Fig. 1 Decision tree for how to pre-treat pre-processed data for statistical analysis

weight to compounds with high variance. However, this may not reflect biological hypotheses that variable compounds might be of higher importance. Scaling aims to exempt from this hypothesis by giving similar weights to all compounds in the analysis. Autoscaling is the recommended scaling method when no prior hypothesis is made about the relationship between compound variation and biological importance (van den Berg *et al*. 2006; Wold *et al*. 2001). The procedure will give the same weight to each of the compounds across samples by adjusting their variance to 1. The disadvantage of autoscaling is that it increases the variance, thus the weight of non-informative compounds with small variance. Finally, for distance-based statistical analyses the Euclidean distance is a natural choice for concentration-like data. As compounds with different orders of magnitude or variance have different weights in distance calculation, we recommend centering and scaling prior to distance calculation to consider all compounds with the same weight.

Compositional Chemical Data Compositional data refers to data where the abundance of one compound is a relative proportion of the total blend of all compounds. As a result, all compounds from a given sample sum up to 1 or 100%. Compositional data must be analyzed with caution. Indeed, the term 'relative proportions' does not explicitly express that proportions should be calculated relative to the sum of all compounds in the given sample. For example, transforming the peak area relative to the highest peak of the blend can lead to serious misinterpretations as illustrated in Fig. S1. Moreover, compositional data are only relevant in situations where the biological hypothesis explicitly relies on ratios of compounds in a blend. Indeed, relative proportions do not inform about concentrations or emission rates (Fig. S2). Finally, the analysis of compositional data faces two main challenges (see Brückner and Heethoff (2017) for a detailed explanation). First, variables are not independent from each other, since the proportion of a given compound in a blend is necessarily modified when the proportion of other compounds change. This may lead to spurious correlations (Pearson 1896). Second, proportional data are constraint because all variables sum to 1. Most statistical methods assume an unbounded data space (Aitchison 1983). The solution proposed by Aitchison (1986) is to transform compositional data using log ratio transformations. Usually, centered log ratio 'clr', or isometric log ratio are used (Filzmoser *et al*. 2009). Similar to concentration-like data, composition data generally benefit from centering and scaling.

For distance-based analyses, distance measures appropriate for compositional data must be used such as the Aitchison's distance (Aitchison *et al*. 2000; Palarea-Albaladejo *et al*. 2012; Pierotti and Martín-Fernández 2011). This distance is simply the Euclidean distance based on clr-transformed data. If we consider all compounds to have the same weight in the distance calculation, data should first be clr-transformed, then centered and scaled before calculating the distance.

Biological Questions and Appropriate Statistical Analyses

The choice of the statistical method primarily depends on the biological question. The first concern is the choice between raw-data methods, i.e. based on samples × compounds matrices, and distance methods, i.e. based on samples × samples distance matrices. Distance methods are appropriate when the question focuses on the overall sample similarity, while neglecting the role of specific compounds. This approach is typically used when testing for the 'isolation-by-distance' hypothesis. It can also be used to study the similarity between compounds while neglecting the role of individual samples. In the first case, all information about individual compounds is lost during the distance calculation process. As a consequence, distance-based multivariate analyses only produce sample score plots that highlight similarities between samples (Box 1). Therefore, if samples appear to be grouped in the score plot, it is not directly possible to identify the compounds that drive the clustering. On the other hand, if the biological question is related to the study of individual compounds, statistical analyses have to be based on raw data instead of a distance matrix. For all types of analyses, the multivariate nature of chemical data naturally leads to the use of ordination methods. These result in the representation of high-dimensional data in a lower dimensional space, thus summarizing most of the information. A large number of ordination methods have been proposed to study datasets either individually, or simultaneously.

The present manuscript introduces some of the most powerful multivariate ordination methods for chemical ecology, which are illustrated in four practical examples. Methods were chosen for their ability to answer typical questions in chemical ecology and for their ease of use (Table 1).

Example 1. ***Fundamental Routine Tools in Chemical Ecology*** The objective is to identify biologically meaningful clusters in the data with principal component analysis (PCA) and principal coordinate analysis (PCoA), and to discriminate known groups of samples with powered partial least squares – discriminant analysis (PPLS-DA).

Example 2. ***Analysis of Chemical Data to Explain Variation by one or Several Controlled Factors*** The objective is to identify which factors significantly influence the observed chemical variation in a controlled experimental design with redundancy analysis (RDA), as well as distance-based RDA when chemical data are formatted as a distance matrix.

Example 3. ***Co-analysis of Chemical Data and Another Dataset Measured on the Same Biological Entities*** **(Transcriptomic, Genomic, Proteomic, Phenotypic, Geographic Data Etc.)** The objective is to reveal the common information shared by the two datasets. Each dataset is first ordinated before superimposing the two ordinations with co-inertia analysis (CIA) and Procrustean co-inertia analysis (PCIA).

Example 4. ***Integrative Analysis of Chemical Data with One or Several Other Multivariate Datasets Measured on the Same Biological Entities*** The objective includes biomarker identification and obtaining a comprehensive understanding of the biological pathways involved at multiple functional levels. We illustrate the recent DIABLO method (Singh *et al*. 2016) that provides a comprehensive and discriminant framework to analyze multiple datasets that are complex and correlated.

Box 1 – How to interpret score plots and correlation circle plots

Most multivariate ordination methods can be interpreted using score plots and correlation circle plots.

Score plots display samples in a low-dimensional space defined by two components chosen by the data analyst. The coordinates of the samples are indicated in the component scores or values. Figure 2a depicts the first component (horizontal axis) and the second component (vertical axis). The plot is interpreted as follows: samples that appear close to each other (e.g. S26, S27 and S29) are chemically similar, whereas samples that are far from each other (e.g. S6 and S31) are chemically different. The score plots represent the samples projected in the space spanned by the two components.

Correlation circle plots display compounds represented as arrows in a circle of radius 1 on two chosen components. Each arrow represents a compound. The coordinates of each arrow tip are defined as the correlation coefficient between the original data and each component. Thus, long arrows represent compounds that contribute importantly to either one or two components (e.g. C1, C2, C3, C7 in Fig. 2b), and short arrows compounds that are poorly correlated with the displayed components (e.g. C4). Compounds with short arrows should not be used for interpretation as they are poorly represented in those dimensions. In addition, the correlation circle plot indicates the correlation structure between compounds by looking at the cosine angle value between two arrows. Considering compounds well represented with long arrows: a

n angle of 0° indicates a positive correlation of 1 (e.g. C1 and C6), an angle of 90° indicates independency (e.g. C1 and C7), and an angle of 180° indicates a negative correlation of −1 (e.g. C1 and C2).

By combining both score and correlation circle plots we can interpret the relationship between samples and compounds. For example S6 has a higher value of C2 than the set of samples S26, S27 and S29, and S31 has a low value of C2. C1 and C2 are negatively correlated with low value of C1 for S6 but high value for S31.

Practical Cases

Example 1: Fundamental Routine Tools in Chemical Ecology

Methods Background Principal Component Analysis (PCA) is a classical ordination method for raw data, while Principal Coordinate Analysis (PCoA) applies to a distance matrix.

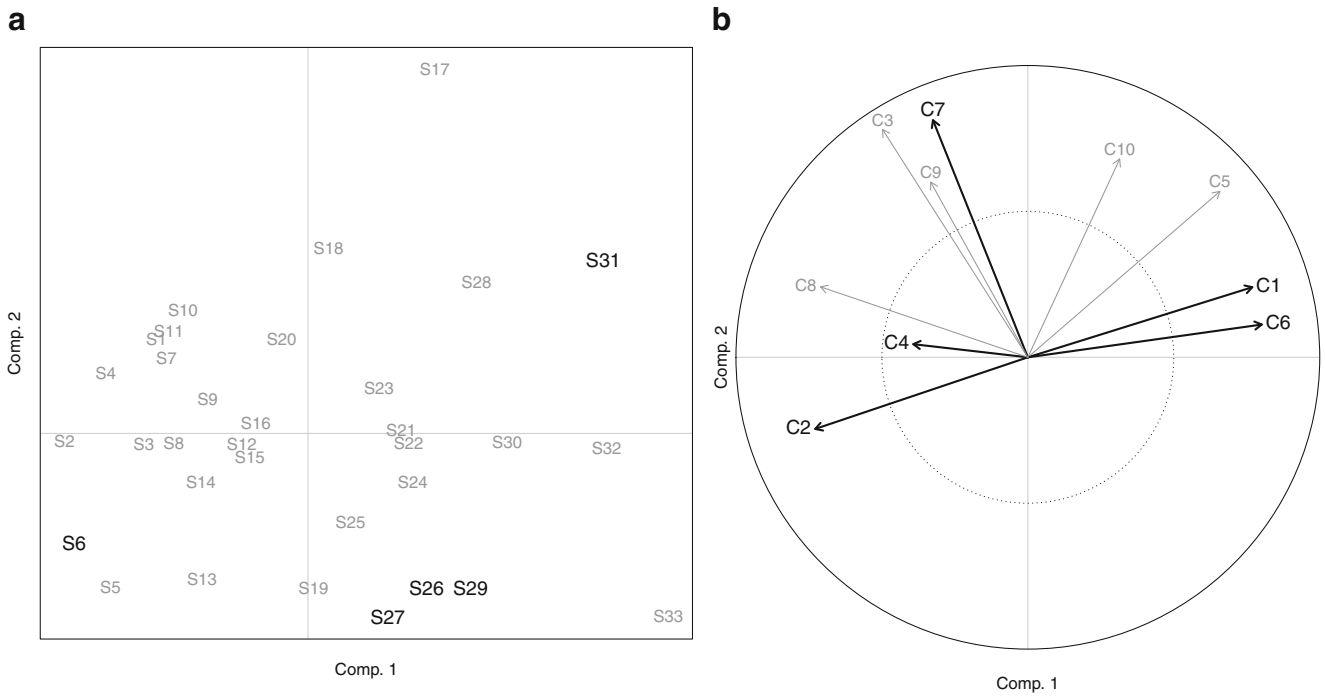


Fig. 2 Example of (**a**) a score plot to represent samples and (**b**) a correlation circle plot to represent compounds projected into a two dimensional space

Both analyses create components – also called axes or dimensions – that summarize the original information of the data in a reduced space. Powered Partial Least Squares – Discriminant Analysis (PPLS-DA) is then presented, that seeks for components that best discriminate between groups of samples that are known *a priori*. As any other discriminant method, PPLS-DA can also be used to assign unknown samples to a group (Box 2).

Data Dormont *et al.* (2014) compared the odor bouquet emitted by two color morphs (white and purple) of each of three orchid species. The aims of the study were to assess the relationship between flower color and flower-emitted volatile compounds, and to assess differences between species in relation to their pollination strategies. In this example, we focus on the plant species *Orchis simia*. The dataset is composed of 17 samples (8–9 samples per color morph) and 53 volatile compounds expressed as relative proportions. The R script of the analysis is provided as Supplementary Information Script 1. The three methods introduced earlier were performed to illustrate the similarity of the two color morphs using PCA and PPLS-DA on compounds' relative proportions and PCoA on the presence/absence distance matrix.

Pre-Treatment For PCA and PPLS-DA, the compositional data were first clr-transformed. Since the data include zeroes, a small constant was added to all values to be of one order of magnitude smaller than the smallest non-zero value (i.e. 0.01 if the smallest non-zero value is 0.1). Data were then autoscaled. For PCoA, the distance matrix was computed from binary data using a distance based on the simple matching coefficient (a symmetric index). Indeed, double-zeroes are considered as similarities in this case. When performing PCoA, it is important to check if the distance measure has Euclidean properties otherwise a special correction is needed beforehand (Box 3). Here the distance measure has Euclidean properties.

Performance of the Analysis and Validation For PCA and PCoA, the performance is classically evaluated through the proportion of (total) variance explained by the first few components. In our example, components 1 and 2 explain 46% of variance in the PCA, 54% in the PCoA. This is satisfying, but low percentages do not necessarily mean that the analysis is pointless (Box 4).

For PPLS-DA, validation is mandatory. Indeed, discriminant analyses can be prone to 'overfitting the data'. Overfitting means that the underlying statistical model is too adjusted to the data and becomes ungeneralizable. In other words, while the score plots highlight a good separation of the sample groups, the generalizability of the results on an external dataset is poor. Graphically, samples groups look well discriminated on score plots but there is no real significant difference between these groups. Overfitting can occur in any large dataset as any combination of variables can discriminate sample groups but are in fact 'noise' that lead to non-reproducible results. It is a topical issue in high-throughput experiments where there are many more compounds than samples. Therefore, the performance of the model must be assessed before any interpretation of score plots. This

Table 1 Overview of multivariate statistical analyses presented in this paper and alternative methods available in the literature

Aim / biological question	No. datasets	Input	Methods presented	Alternative methods
Identify patterns in chemical variation	1	Raw data	**Principal Component Analysis** (PCA) (Hotelling 1933; Pearson 1901)	
Do samples structure into groups or along gradients? Do groups of compounds covary?		Distance matrix	**Principal Coordinate Analysis** (PCoA or PCO) = Metric Multidimensional Scaling (MDS) = Classical Scaling (Gower 1966) PCoA preserves distances in a multidimensional space and is the preferred choice when seeking for common information between two distance matrices (see Example 3a)	**Non Metric Multidimensional Scaling** (NMDS) (Kruskal 1964a, b) NMDS is based on ranks of distances, thus the absolute distance between two samples in a score plot cannot be interpreted as a real distance
Explain chemical variation by controlled variables (quantitative and/or qualitative) Does chemical composition differ between known sample groups? Which variables or interactions influence chemical variation in a multifactorial experimental design?	2	Raw data	**Powered Partial Least Squares - Discriminant Analysis** (PPLS-DA) (Liland and Indahl 2009) PLS ('Partial Least Squares' or 'Projection to Latent Structures') families of methods are better suited for datasets with a large number of variables (metabolites) relative to the number of individuals (samples)	**Linear Discriminant Analysis** (LDA) = Canonical Variate Analysis (CVA) = Canonical Discriminant Analysis (CDA) (Fisher 1936) LDA faces computational challenges when the number of variables (metabolites) are nearly equal to, or greater than, the number of individuals (samples), or when variables are collinear, i.e. correlated with each other **Other PLS-based discriminant methods** (e.g. Orthogonal PLS-DA, OPLS-DA or O-PLS-DA)
			The term 'PLS-DA' can refer to three different types of methods: 1/ a particular case of PLS Regression (PLSR; Wold *et al.* 1983) called PLSR-DA (Ståhle and Wold 1987) where the group information is coded as an indicator matrix (Table 2), 2/ an improved discriminant version of PLSR-DA (Barker and Rayens 2003; Indahl *et al.* 2007; Nocairi *et al.* 2005) called PLS-DA, and 3/ a particular case of a Canonical Powered PLS (CPPLS) framework (Indahl *et al.* 2009; Liland and Indahl 2009) called PPLS-DA. Our article focuses on PPLS-DA, as it has been shown to achieve better discrimination performances on fewer components than the other versions (Liland and Indahl 2009). Other PLS-based discriminant analyses exist such as the common OPLS-DA® (Bylesjö *et al.* 2006), which was reported to have the same discriminant power as PLSR-DA (Tapp and Kemsley 2009). All these methods are suited for class prediction of unknown samples (see Box 2)	
		Raw data	**Redundancy Analysis** (RDA) = Principal Component Analysis with Respect to Instrumental Variables (PCAIV) (Rao 1964; Van Den Wollenberg 1977) RDA decomposes the explained and unexplained variation from the experimental design and enables to interpret the significant effects in a single ordination	**ANOVA-Principal Component Analysis** (ANOVA-PCA) (Harrington *et al.* 2005) **ANOVA-Simultaneous Component Analysis** (ASCA) (Jansen *et al.* 2005; Smilde *et al.* 2005) **ANOVA-Target Projection** (ANOVA-TP) (Marini *et al.* 2015) **regularized Multivariate ANOVA** (rMANOVA) (Engel *et al.* 2015)
			These methods assess chemical variation while taking into account an explicit experimental design. It decomposes chemical variation with respect to the individual effect of each controlled variable and interaction term. When the design includes more than one factor, these methods are preferable to discriminant analyses. In the case of one-factor design, discriminant analyses are equivalent and can be used	
		Distance matrix	**distance-based Redundancy Analysis** (db-RDA) (Legendre and Anderson 1999) For multiple factors analysis, db-RDA is preferable to DPLS as it explicitly considers the experimental design	**Dissimilarity Partial Least Squares** (DPLS) (Zerzucha et al. 2012)
			These methods extend RDA and PLSR for distance matrices instead of raw data. For experimental designs with one factor, they are mostly exchangeable (in such situation, DPLS is called DPLS-DA)	

Table 1 (continued)

Aim / biological question	No. datasets	Input	Methods presented	Alternative methods
Identify or summarize common information between chemical data and other datasets Is chemical variation related to the variation in other datasets? Are common patterns of variation explained by differences between known sample groups?	2	Raw data	**Co-Inertia Analysis** (CIA) (Dolédec and Chessel 1994) CIA is more flexible than 2B–PLS and O2PLS as it is not restricted to quantitative datasets only (see Box 7).	**Canonical Correlation Analysis** (CCorA or CCA) (Hotelling 1936) CCorA is intractable when the number of variables in one dataset exceed the number of individuals **two-Block Partial Least Squares** (2B–PLS) (Sampson et al. 1989) **orthogonal two-block Partial Least Squares** (O2PLS) (Trygg 2002; Trygg and Wold 2003)
			2B–PLS and CIA are covariance-based methods (the covariance is the non-standardized version of the correlation where values are not bounded between −1 and +1) that seek for a good agreement representation of both datasets if a correlation structure between datasets exist. They do not suffer computational problems encountered by CCorA. Chemometricians extended 2B–PLS to O2PLS to improve interpretability	
		Distance matrix	**Procrustean Co-Inertia Analysis** (PCIA) (Dray *et al*. 2003a)	
	≥ 2	Raw data	**DIABLO** (Singh *et al*. 2016) DIABLO is an extension and combination of RGCCA and PLS-DA for multiple integration of omics datasets in a discriminative framework. It is particularly useful in chemical ecology since chemical ecology is fundamentally a comparative science	**Multiple Co-Inertia Analysis** (MCIA) (Chessel and Hanafi 1996) **orthogonal multiblock Partial Least Squares** (OnPLS) (Löfstedt and Trygg 2011; Löfstedt *et al*. 2012, 2013) **Regularized Generalized Canonical Correlation Analysis** (RGCCA) (Tenenhaus and Tenenhaus 2011)
			MCIA and OnPLS extend CIA and O2PLS to integrate more than two datasets. RGCCA extends CCorA for large datasets with a regularization parameter (Leurgans *et al*. 1993; Vinod 1976) and includes a PLS – Path Modeling framework (Lohmöller 1989; Wold 1985) to explicitly state the connection between datasets (e.g. an expected causal, or symmetric relationship)	

assessment is often achieved by evaluating the number of misclassifications (or error rate), i.e. the number of samples that do not belong to the group predicted by the model. The classification error rate is computed using cross-validation (Brereton and Lloyd 2014; Kjeldahl and Bro 2010; Westerhuis *et al*. 2008; Worley and Powers 2013; see also Box 5). A permutation test based on the classification error rate is used to conclude about the significance of the differences between groups (Westerhuis *et al*. 2008). The score plots and correlation circle plots can only be interpreted when at least two groups are declared significantly different. In our example, the classification error rate was ~ 7% using cross model validation (outer loop: 7-fold CV, inner loop: 6-fold CV). This satisfying performance was confirmed by rejecting the null hypothesis that there was no difference between samples groups ($P = 0.001$). In other words, there was a significant difference between the two color morphs as modelled by PPLS-DA. The interpretation of the score plots can then follow. In Example S1 that considers the two other orchid species studied in Dormont *et al*. (2014), we show an example where

Table 2 Transformation of a three-class sample grouping information factor into an indicator matrix. In reality, only two columns are needed since the third column can be built from the first two

Group		Group.G1	Group.G2	Group.G3
G1	→	1	0	0
G1		1	0	0
G2		0	1	0
G2		0	1	0
G3		0	0	1
G3		0	0	1

Box 2 – Prediction with PPLS-DA

Powered Partial Least Squares – Discriminant Analysis (PPLS-DA), Partial Least Squares Regression – Discriminant Analysis (PLSR-DA) and Partial Least Squares – Discriminant Analysis (PLS-DA) (see Table 1) can be used to predict the group of new samples for which only chemical data are available. However, PLS-based discriminant analyses are not classification tools, i.e. they can discriminate groups very well but they lack an objective and relevant decision rule to affect a sample to a group (Indahl *et al*. 2007). Several options are still possible for prediction, such as applying PPLS-DA first to discriminate the groups then applying a classification method such as LDA or Quadratic Discriminant Analysis (QDA; e.g. Hastie *et al*. 2001) on the PPLS-DA scores for the prediction (Indahl *et al*. 2007). QDA is more flexible than LDA but less robust when the number of individuals per group decreases (Liland and Indahl 2009). Such a strategy has been classically used with PCA scores instead of PPLS-DA scores (Bertrand *et al*. 1990; Jombart *et al*. 2010). However, Kemsley (1996) confirmed that using PLS-DA scores rather than PCA scores outperformed the classification results. The other solution is to apply a prediction matrix on the predicted indicator matrix values in PLSR-DA (Lê Cao *et al*. (2011) implemented in Rohart *et al*. (2017)).

score plots of a PPLS-DA show a strong discrimination between groups that is not significant.

Interpretation All three analyses showed a clear difference between the two color morph groups (Fig. 3a–c). Such result indicates that the inter-morph variation is the first source of variation in the data (PCA). Moreover, it shows that the information about the presence/absence of the compounds in odor bouquets is sufficient to detect differences between morphs (PCoA). The correlation circle of the PPLS-DA (Fig. 3d) indicates that many compounds are driving the discrimination. First, C2, C18, C19, C28, C35, C36 and C37 are negatively correlated with component 1, meaning their relative proportion is higher in odor bouquets of the purple morph of *O. simia*. A second subset of C6, C14, C31, C43 and C53 is positively correlated with component 1, meaning they are relatively more present in the white morph.

Box 3 – PCoA and Euclidean properties of the distance measure

Euclidean properties on a distance measure include: a minimum value 0, positiveness, symmetry, triangle inequality and embedding in an Euclidean space (Legendre and Legendre 2012). Principal Coordinate Analysis (PCoA) can be applied to any Euclidean or non-Euclidean distances. For the latter however, negative percentages of variance explained may occur because of negative eigenvalues. Fortunately, many distances other than the Euclidean distance have Euclidean properties, in their basic form or after a square-root transformation, as listed in Gower and Legendre (1986). If this is not the case, PCoA can still be used but with the inclusion of an offset in the distance matrix beforehand. This offset is equal to the absolute value of the largest negative eigenvalue (Gower and Legendre 1986; Legendre and Anderson 1999). The offset correction is available as an option in most R PCoA functions.

Box 4 – PCA, PCoA and percentages of variance explained

Multivariate analyses including Principal Component Analysis (PCA) and Principal Coordinate Analysis (PCoA), aim to summarize a certain type of information, for example the total variance in PCA and PCoA. Since only the first 2–3 components are generally used for interpretation, the performance of the analysis is often evaluated through the percentage of variance explained by these axes. When dealing with a large and noisy dataset such as in metabolomics, it is common that only a small fraction of the total variation is explained by the first components. This may seem to limit the ability to draw meaningful conclusions. However, a good representation can be achieved even with a small proportion of explained variance (Legendre and Legendre 2012), as long as the representation of the inter-sample distances in the new ordination space is well respected. In other words, the correlation between the real distances and the distances displayed in the reduced space spanned by the components should be substantial. The relationship between both real and projected distances in the reduced space is classically assessed with a Shepard diagram (Shepard 1962), see Example 3b.

Example 2. Analysis of Chemical Data to Explain Variation by one or Several Controlled Factors

Methods Background In this context, chemical variation (concentration-like or compositional data) is studied as a function of an experimental design. The controlled variables can be quantitative or qualitative, such as temperature, density of competitors or natural enemies, treatment, genotype, or geographical origin. In addition, the controlled variables can be treated as in classical

Box 5 – Cross-validation in discriminant analyses

Cross-validation is a multi-step process in which the dataset is randomly split into a training set and a test set. A discriminant analysis model, such as Linear Discriminant Analysis (LDA) or Partial Least Squares – Discriminant Analysis (PLS-DA), is fit on the training set. In a second round, this model is used to predict the sample group from the test set. The process is repeated M times and at each step a different $\frac{1}{M}$ fraction of the whole dataset is considered as the test set. At the end of the process, each sample is assigned a predicted group and the predictions are compared to the real groups by calculating the classification error rate. The process is called 'M-fold cross-validation'. It includes the special case of 'leave-one-out cross-validation', LOO-CV, where each test set contains only one sample (M = the number of samples in the whole dataset).

For PLS-based discriminant analyses, this cross-validation procedure has been shown to be biased or overoptimistic (Westerhuis *et al*. 2008). One way to circumvent this problem is to use cross model validation ('2CV'), which includes a second stratum of cross-validation within the training set. Therefore, there is an 'outer loop' with M-fold cross-validation, and, at each step of the outer loop, a complete 'inner loop' with its own M-fold cross-validation procedure (Smit *et al*. 2007; Szymańska *et al*. 2012; Westerhuis *et al*. 2008). This procedure is however unsuitable for small sample numbers. As a rule of thumb, we recommend using LOO-CV when there are less than 5 samples in the test set.

Note that except for LOO-CV, the CV procedure should be repeated several times to ensure an accurate evaluation performance, as the CV test sets are randomly assigned.

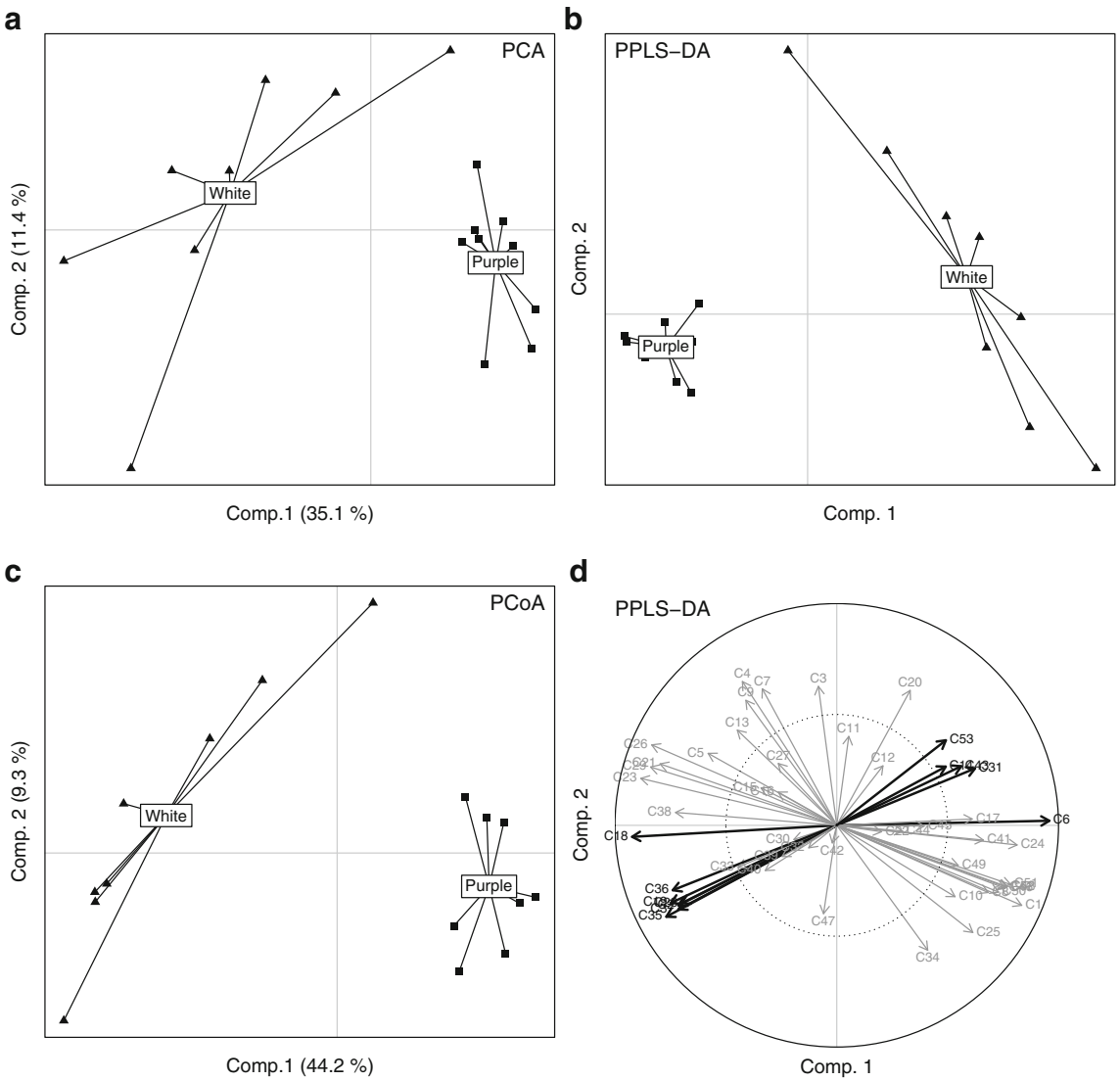


Fig. 3 Graphical plots from Example 1: **a** Principal Component Analysis (PCA), (**b**) Powered Partial Least Squares – Discriminant Analysis (PPLS-DA) and (**c**) Principal Coordinate Analysis (PCoA) score plots; (**d**) PPLS-DA correlation circle, where compounds in dark color are those mentioned in the text

Box 6 – Repeated measurements experimental designs

Experimental designs often include repeated measurements, where a series of samples from the same biological material are studied under different treatments or conditions. Such experiment leads to 'intra-subject' (or 'within-subject') and 'inter-subject' (or 'between-subject') variation. In a univariate framework, repeated measurements are classically analyzed using paired *t*-tests or linear mixed-effect models where the subject is considered as a random-effect factor. In a multivariate framework, repeated measurements can be dealt with by removing the between-subject variability first before analyzing the remaining variation in the data with classical multivariate analyses. This approach has been named 'multilevel analysis' and can be performed with the methods presented in this article (e.g. Liquet *et al.* 2012; van Velzen *et al.* 2008; Westerhuis *et al.* 2010). In the particular case of Redundancy Analysis (RDA), partial RDA (Legendre and Legendre 2012) was proposed to control for between-subject variation.

univariate linear models: crossed or nested factors, interactions, random effects etc. (see also Box 6). The aim of the analysis is similar to classical univariate models such as linear regression or ANOVA. In the multivariate framework, we consider Redundancy Analysis (RDA). The analysis consists of two steps. First we fit a multivariate linear regression between the chemical data and the controlled variables. Second we perform two PCAs. The 'constrained PCA' is applied on the fitted values of the regression and summarizes the variation of the chemical data explained by the controlled variables. The 'unconstrained PCA' is applied on the residuals of the regression and summarizes the variation that is not related to the controlled variables. When there is only one qualitative controlled variable, both RDA and PPLS-DA can be employed.

RDA was further extended to distance-based RDA (db-RDA) for a response dataset transformed to a distance matrix. db-RDA includes two steps. First, the distance matrix is ordinated through PCoA to obtain the sample coordinates on PCoA components. Second, these coordinates are input as response variables in a classic RDA.

The example below illustrates RDA. A similar procedure applies for db-RDA, and while it is not included in the example, a short example based on data from Greff *et al*. (2017) is provided as Example S2 (with Supplementary Information Script 3).

Data Conchou *et al*. (2014) studied the odor bouquet emitted by leaves and fruits of two fig tree species (*Ficus septica* and *F. nota*), twice a day (morning and noon). The aim of the study was to identify whether the differences in volatile bouquets could be explained by one or several variables (organ, species, and sampling time). Ultimately, the objective was to relate differences in odor bouquets with attractiveness to pollinators. The dataset includes 40 samples (5 samples per species/organ/time) and 70 volatile compounds expressed as relative proportions. The R script of the analysis is provided as Supplementary Information Script 2.

Pre-Treatment Similar to Example 1, a small constant was added to the compositional data to remove zeroes before being clr transformed and autoscaled.

Model The controlled variables are three factors with two levels each: the fig tree species (*F. septica* or *F. nota*), the organ (leaf or fig) and the time of the day (morning or noon). All two and three-way interactions between these factors were included in the model as they are biologically relevant.

Validation To assess the effect of the controlled variables on the chemical variation, two complementary strategies can be adopted. The first strategy is to measure the total percentage of the chemical variation explained by the controlled variables. This percentage is estimated by a canonical R^2 called 'bimultivariate redundancy statistic' (Miller and Farr 1971) and is calculated as proposed by Peres-Neto *et al*. (2006). In our example, the three factors and their interactions explain 39% of the total variation. This indicates that the remaining 61% of the variation could be due to noise in the data, or unknown factors as volatiles were sampled in the field. The second strategy is to test whether controlled variables explain a significant proportion of the chemical data, using a permutation *F*-test based on the canonical R^2 (Legendre and Legendre 2012). In our example the test was declared significant ($P = 0.001$), implying that the controlled variables explain a significant part of the chemical variation. Then, interpreting the correlation circle plots will be meaningful.

Individual Effects of the Controlled Variables Similar to classical univariate analyses, we are interested in testing the individual effect of each controlled variable and their interactions. This can be achieved using a permutation *F*-test. In our example (Table 3), the odor bouquet was significantly influenced by two interactions. The Species × Organ significant interaction indicated that the effect of the fig tree species is different for figs and leaves. Pairwise comparisons showed that figs of *F. nota* significantly differed from the three other groups (leaves of *F. nota*, figs and leaves of *F. septica*). The volatile bouquets of the latter three groups greatly overlapped and did not significantly differ from each other (Fig. 4c, upper table). Therefore, the fig tree species differ in their chemical profile for figs only. The significant interaction Organ × Moment indicated that the chemical profile of the organs changes in a different way between morning and noon. Pairwise comparisons showed that all four groups were different (Fig. 4c, lower table).

Interpretation Figures 4a,b display two identical constrained PCA score plots with different symbols highlighting the two significant interactions Species × Organ and Organ × Moment. Note that, by definition, the unconstrained PCA cannot assess the influence of the controlled variables. A total of 74% of the constrained variance was explained by the first and the second components of the PCA (45% and 29% respectively).

For the Species × Organ interaction (Fig. 4a), as previously shown by the pairwise comparisons, figs of *F. nota* were clearly separated from the three other groups, where samples overlapped. The discrimination between figs of *F. nota* and the other three groups is visualized on the first component. Figure 4d displays the correlation circle of the constrained PCA. It highlights many discriminative compounds such as e.g. C49 and C53, which are negatively correlated with component 1. In contrast, C27, C36 and C67 are positively correlated with component 1. The relative proportion of compounds C49 and C53 in odor bouquets of figs of *F. nota* is higher than in the three other groups, and we observe the opposite pattern for C27, C36 and C67. For the Organ × Moment interaction (Fig. 4b), the discrimination between the two time points can be observed on the second component. Samples collected in the morning are in the upper part of the plot whereas samples collected at noon are projected in the lower part. The interaction effect can also be observed in the plot. Although the contrast between morning and noon goes in the same vertical direction for figs and leaves, its amplitude is much higher for figs than leaves. The correlation circle (Fig. 4d) highlights the compounds that most likely explain the difference between morning and noon. The relative proportion of compounds C6, C8, C15, C32 and C33 in odor bouquets emitted at noon is higher than in the odor bouquets emitted in the morning, while the pattern is opposite for C18.

Table 3 Example 2: permutation *F*-tests of the factors included in Redundancy Analysis (RDA) (999 permutations). Significant *P*-values are indicated in bold

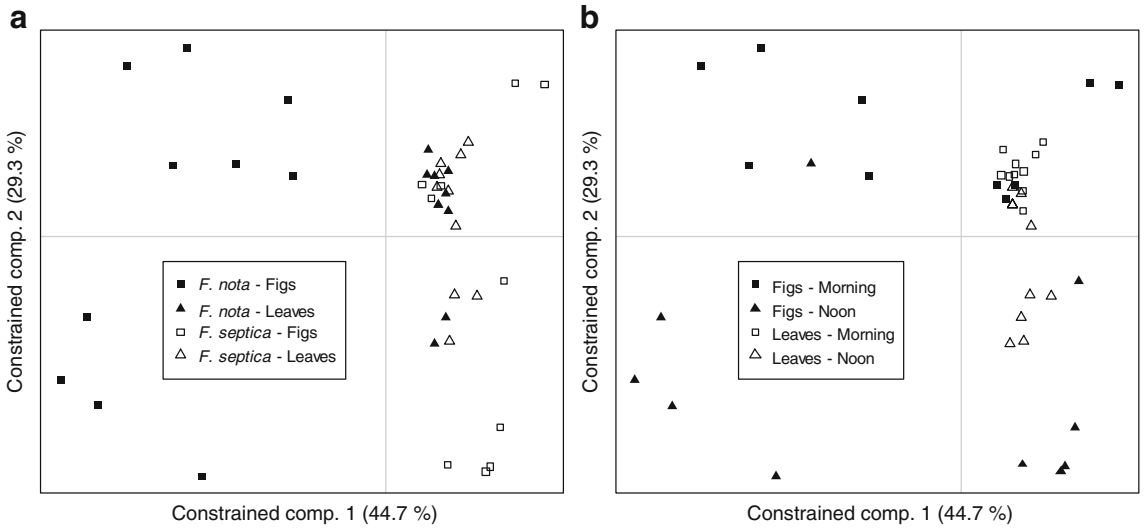
	F	*P*
Species	4.230	**0.001**
Organ	3.596	**0.001**
Moment	4.725	**0.001**
Species × Organ	3.560	**0.001**
Species × Moment	1.196	0.227
Organ × Moment	1.842	**0.035**
Species × Organ × Moment	1.007	0.414

Example 3. Co-Analysis of Chemical Data and another Dataset Measured on the Same Biological Entities

We illustrate two examples that typically arise in chemical ecology, where datasets are either distance matrices or raw data. The methods we introduce superimpose the two datasets to maximize their covariance, and reveal common patterns if they exist, in a symmetric analysis.

Example 3a – Coupling two Distance Matrices

Methods Background Procrustean Co-Inertia Analysis (PCIA) combines the advantages of Procrustes analysis (Gower 1971) and Co-Inertia Analysis (CIA). The superimposition between



c

	Fn – Figs	*Fn* – Leaves	*Fs* – Figs
Fn – Leaves	**0.002**	-	-
Fs – Figs	**0.002**	0.247	-
Fs – Leaves	**0.002**	0.731	0.247

	Figs – M	Figs – N	Leaves – M
Figs – N	**0.002**	-	-
Leaves – M	**0.035**	**0.002**	-
Leaves – N	**0.002**	**0.004**	**0.003**

Fn: *F. nota, Fs*: *F. septica,* M: morning, N: noon
999 permutations
P-value adjustment method: fdr

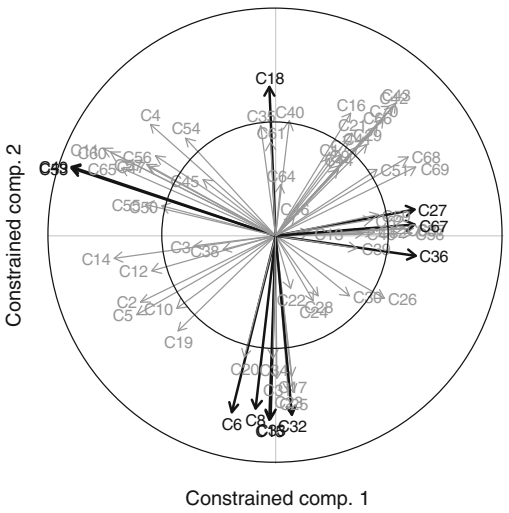


Fig. 4 Example 2 outputs. Redundancy Analysis (RDA) score plot from the constrained Principal Component Analysis (PCA) with symbols indicating (**a**) fig tree species and organ, and (**b**) organ and time; (**c**) permutation *F*-tests to assess pairwise comparisons for fig tree species and organ (upper table) and organ and time (lower table); (**d**) RDA correlation circle where compounds in dark color are those mentioned in the text

two sets of samples is achieved with Procrustes analysis. This translates, rotates and scales the configuration of one set of points in a multidimensional space to fit the configuration of another set of points. Then CIA is applied on the Procrustes analysis results to improve interpretation by providing a simultaneous representation of the information contained in both datasets.

Data Bonelli *et al.* (2015) studied whether the chemical composition of the cuticle of the social wasp *Polistes biglumis* could vary with the wasp population. The aim was to test the 'isolation-by-distance hypothesis'. This states that an increase in geographical distance between populations leads to an increase in difference between their cuticular compounds. We analyzed the following datasets: a geographic distance matrix (9 populations × 9 populations) calculated from longitudinal, latitudinal and altitudinal coordinates, and a chemical distance matrix corresponding to the same nine populations. The chemical distance matrix was based on 31 cuticular hydrocarbons expressed as relative proportions in 4–7 foundresses per population. The chemical compositional data were clr-transformed and autoscaled. The distance between each pair of populations was computed as the mean Euclidean distance between two individuals of each population. The R script of the analysis is provided as Supplementary Information Script 4.

Pre-Treatment PCIA can only be used with individuals × variables matrices and cannot be directly applied on distance matrices. Thus, to deal with distance matrices, a preliminary step is to ordinate the samples with PCoA and use the components of the ordination as input variables in PCIA. Since the geographic distance matrix did not have Euclidean properties, a correction was added in the 'geographic PCoA' (Box 3). No correction was needed for the 'chemical PCoA'. The most structuring components from each PCoA were then retained for PCIA: two components for the geographic PCoA (100% of variance explained), five components for the chemical PCoA (75%). Note that there is no criterion that have been proposed to assess how many components should be retained in PCoA.

Test of Concordance The concordance of the two PCoAs, then of the two distance matrices, is evaluated by a statistic called m^2 which varies between 0 (no concordance) and 1 (perfect concordance). The significance of this statistic, i.e. the significance of the correlation between the distance matrices, is tested using a test named PROTEST (Dray *et al.* 2003a; Jackson 1995). In our example, we obtained $m^2 = 0.584$ and $P = 0.027$ indicating a significant concordance between chemical and geographic data. Therefore, the increased spatial distance between two wasp populations leads to an increased difference between the composition of their cuticle (in terms of hydrocarbons' relative proportions).

Interpretation PCIA displays samples similarity for each distance matrix (Fig. 5a,b) as well as the concordance between the two matrices (Fig. 5c). In the latter plot, each sample is represented by an arrow. The start of the arrow indicates the position of the population for the geographic PCoA. The tip of the arrow gives the position of the same population for the chemical PCoA. Therefore, short arrows indicate a strong concordance between the two distance matrices. Figures 5a, b show that populations are more divergent for the geographic data than for the chemical data. However, there is some concordance between the two distance measures. Populations that are the most geographically distant, such as P6 *vs*. P8, or P1 *vs*. P7 are also the most chemically distant.

Example 3b – Coupling Two Raw Data Tables

Data Sacristán-Soriano *et al.* (2011) studied the production of brominated alkaloids, secondary metabolites in the sponge *Aplysina aerophoba*, in relation to the bacterial communities associated with this sponge. Since only microorganisms were thought to be able to produce such brominated compounds, the aim was to find out whether the bacteria found in sponge cells might be the true producers of these metabolites. The first goal was to investigate relationships between metabolite concentration in the sponge and abundance of specific bacteria in contact with sponge cells. The first dataset includes the absolute concentration of four metabolites measured on 32 samples. The second dataset includes the relative proportion of 24 bacterial Operational Taxonomic Units (OTUs) in the bacterial community on the same samples. The R script of the analysis is provided as Supplementary Information Script 5.

Pre-Treatment Chemical data were log-transformed then autoscaled. Since bacterial data were compositional and contained zero values, a small constant was added before clr transformation and autoscaling. A mandatory step prior to CIA is to ordinate each dataset separately to identify synthetic variables structuring the information contained in the original datasets (Dolédec and Chessel 1994; Dray *et al.* 2003b and Box 7). Here as both datasets are quantitative variables, a 'chemical PCA' and a 'bacterial PCA' were performed. The first two components of the chemical PCA explained 91% of the chemical variation, while the first two components of the bacterial PCA explained 40% of the bacterial variation. The latter may seem unsatisfying. As

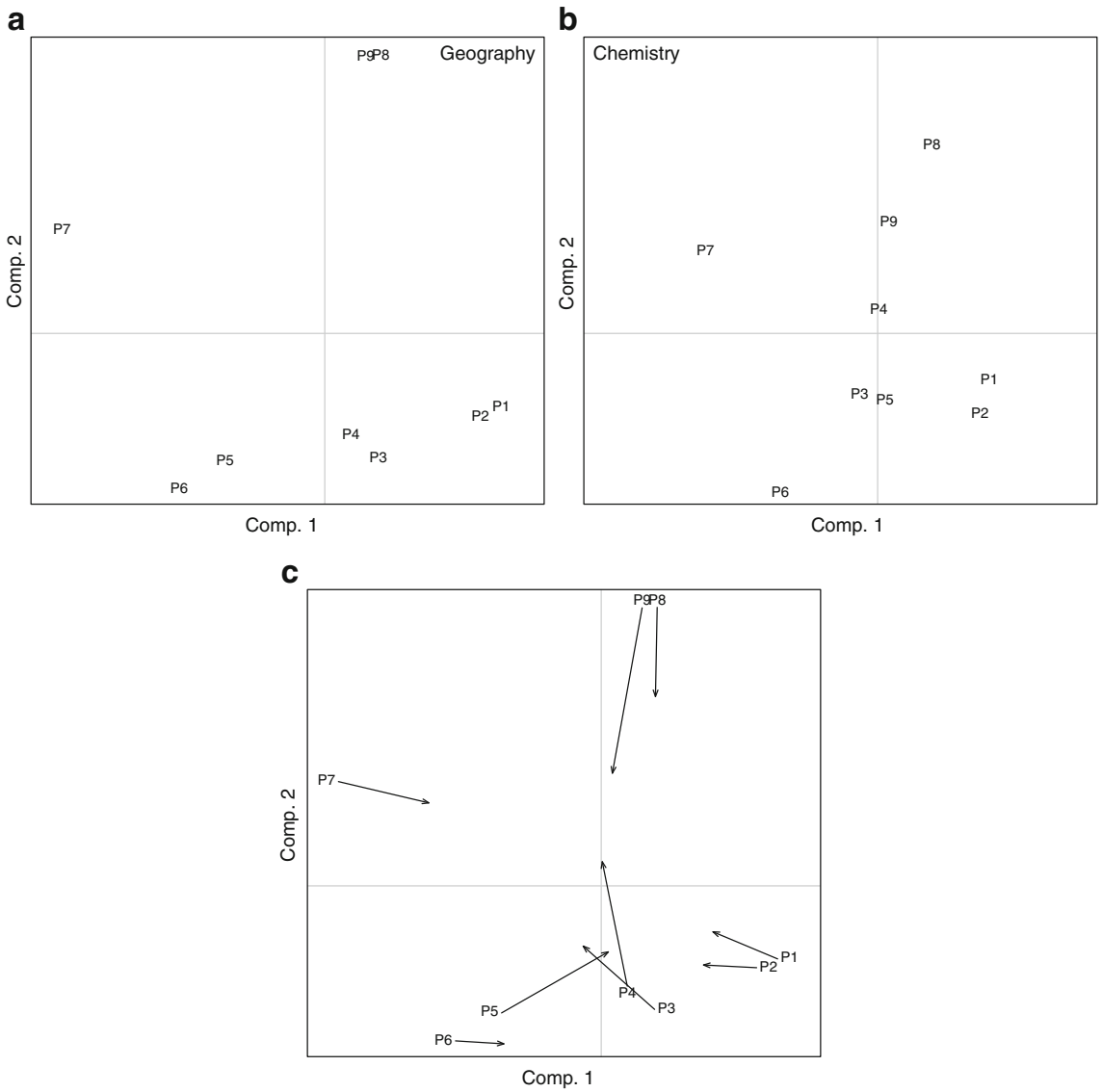


Fig. 5 Example 3a: Procrustean Co-Inertia Analysis (PCIA) score plots. Population coordinates in the PCIA space based on (**a**) geographic distances; (**b**) chemical distances; (**c**) both distances (arrow-starting point: geography, arrow-ending point: chemistry)

mentioned in Box 4 low percentages of explained variance are common and do not inhibit interpretation of the biological relevance. In this example, the inter-sample distances were well preserved by using only the first two components of the bacterial PCA (Fig. S3). Since both datasets can be summarized by a few components, we could proceed with the CIA to identify potential common structures or co-structures between the two datasets.

Test of Concordance The concordance of the patterns in the spaces of the two PCAs is evaluated by the *RV*-coefficient (Escoufier 1973; Robert and Escoufier 1976). This multivariate statistic is analogous to the squared Pearson correlation coefficient that ranges between 0 (no concordance) and 1 (perfect concordance). The significance of the concordance between the two datasets is assessed using a permutation test (Heo and Gabriel 1998). Here $RV = 0.277$ and $P = 0.004$ which indicate a significant concordance between the concentration of brominated alkaloids and the composition of the

Box 7 – CIA is a flexible co-analysis technique

As Co-Inertia Analysis (CIA) cannot be applied on raw data, unconstrained ordinations are to be performed first on each dataset separately. While this may seem cumbersome, it makes CIA a very flexible method. Indeed, different ordination methods can be used depending on the nature of the data. For example, quantitative variables are handled with PCA, qualitative variables with multiple correspondence analysis (Tenenhaus and Young 1985), and a mix of both quantitative and qualitative variables with a Hill and Smith analysis (Hill and Smith 1976). Thus, CIA enables the coupling of chemical data to any other type of datasets (Dray *et al.* 2003b).

sponge-associated bacterial community. Therefore, we can expect that the relative proportions of some OTUs are correlated with the concentration of some metabolites.

Interpretation The score plots in CIA represent one multivariate reduced space per dataset that correspond to the rotations from the two PCAs. Each pair of co-inertia components is defined at a time. In our example, most of the co-inertia, i.e. the common information shared by the two datasets, was explained by the first pair of components (66%) which were highly positively correlated ($r = 0.66$). This means that the co-structure of the two datasets can be mostly depicted in one dimension. The first pair of co-inertia components highlight a set of highly positively correlated compounds (C1, C2 and C4) that were mostly positively correlated to OTUs 6 and 24 (Fig. 6a,b). This indicates that the higher the relative proportion of these two OTUs in the bacterial community, the higher the concentration of compounds C1, C2 and C4 in the sponge. The second pair of co-inertia components explained almost all of the remaining co-inertia (29%) and were positively correlated ($r = 0.70$). This pair of components showed that compound C3 varied independently from the three other compounds and was positively correlated to OTU 17 (Fig. 6c, d). Therefore, the higher the relative proportion

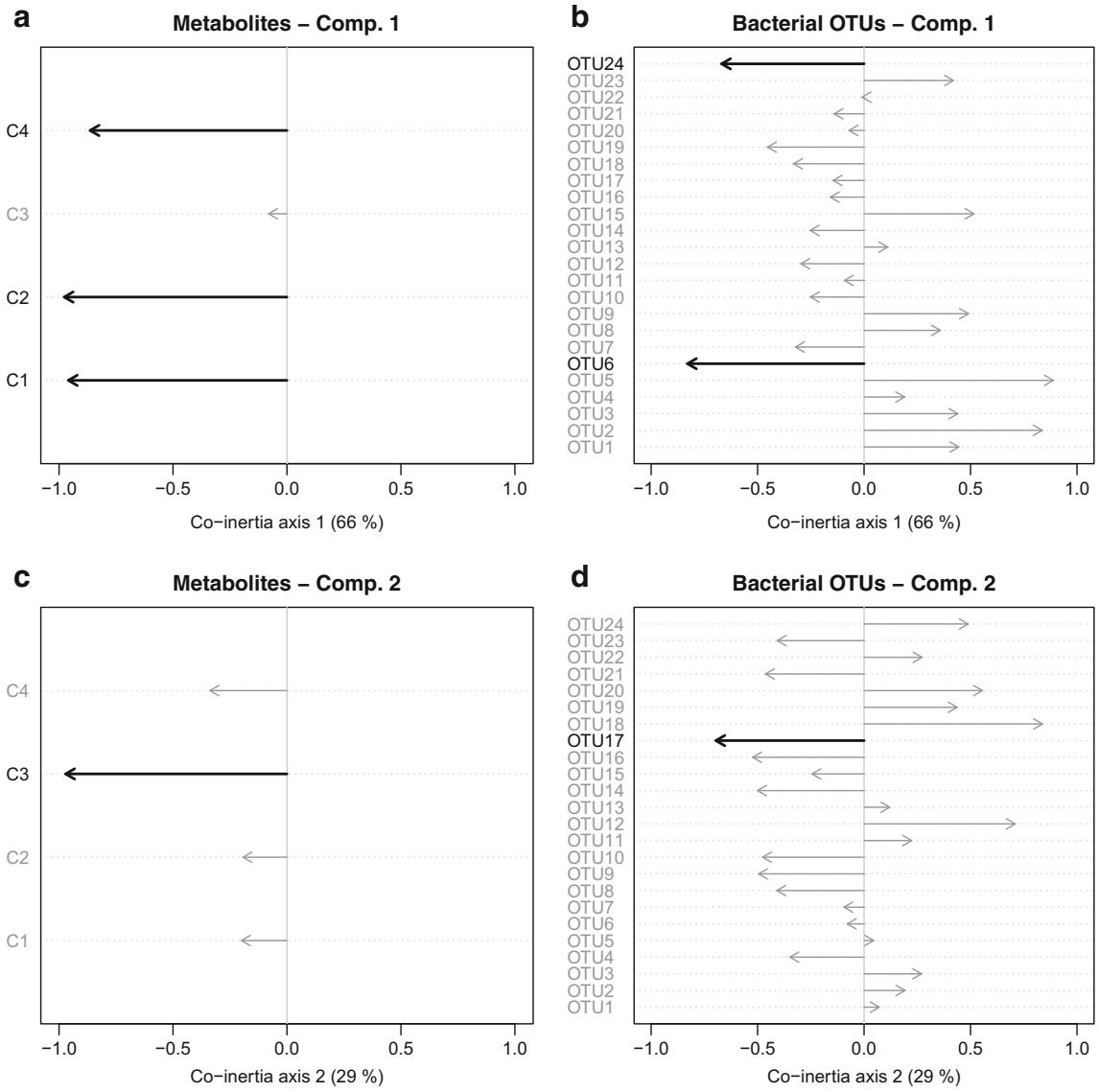


Fig. 6 Example 3b: Co-Inertia (CIA) univariate plots. Correlation between (**a**) metabolites and the first 'chemical' co-inertia component; (**b**) bacterial OTUs and the first 'bacterial' co-inertia component; (**c**) metabolites and the second chemical co-inertia component; (**d**) bacterial and the second bacterial co-inertia component. Compounds and OTUs in dark color are those mentioned in the text

of OTU 17 in the bacterial community, the higher the concentration of compound C3. We conclude from the analysis that the three OTUs 6, 17 and 24, appear as good candidates to be producers of the brominated alkaloids found in this sponge species.

Example 4: Integrative Analysis of Chemical Data with One or Several Other Multivariate Datasets Measured on the Same Biological Entities

Methods Background Integrative statistical analyses use information from several datasets simultaneously, which makes the analysis more powerful than separate independent analyses (Tseng *et al*. 2015; Zhang *et al*. 2010). Several integrative multivariate methods were developed to identify reliable omics signatures that can predict a given phenotype (Tieri *et al*. 2015; Tseng *et al*. 2015). Amongst these methods, DIABLO might be particularly relevant for chemical ecology problems. DIABLO extends existing multivariate methods to achieve the following goals: the identification of compounds that are (i) highly correlated between several datasets and (ii) discriminative between *a priori* known groups of samples. A typical situation where DIABLO is appropriate is to integrate metabolomics and transcriptomics datasets to identify highly correlated metabolites and transcripts that discriminate several treatments. Such analyses can give direct insights into metabolic pathways and genes involved, for example, in the response of an organism when interacting with another organism.

Data Hervé *et al*. (2014) compared several genotypes of oilseed rape for the chemical content of their flower buds. These genotypes are known to be differently attacked by an insect herbivore feeding from the pollen inside the buds. Targeted classes of primary and secondary metabolites were quantified in two bud tissues that are fed upon by the insect. These tissues were expected to be chemically different, since they are different plant structures: the perianth, i.e. the 'bud wall', and the anthers, which contain the pollen. The aim of the study was to investigate whether, despite different chemical compositions, the two tissues could be characterized by a genotype-specific signature. Our example here was considered based on three genotypes. The two sample-matching datasets (5 samples per genotype, 15 samples in total) include the absolute concentration of 40 metabolites in the perianths and the absolute concentration of 43 metabolites in the anthers. The R script of the analysis is provided as Supplementary Information Script 6.

Pre-Treatment Chemical data of each dataset were autoscaled. A fourth root-transformation was considered, but did not improve the discrimination of the data and was not included in the analysis.

Validation Since DIABLO builds on PLS-DA, validation of its discriminant power is as important as with PPLS-DA (see Example 1). The additional complexity resides in integrating several datasets (blocks) to discriminate the samples groups. Each block is summarized by a set of components which lead to their own predictions which are then combined. In our example, the classification error rate was ~ 1%. This high classification performance was confirmed with a cross-validation-based permutation test ($P = 0.001$), which indicates a significant difference between the three genotypes. Further downstream analysis can then follow for interpretation.

Similar to CIA, the DIABLO score plots represent one reduce space per dataset. The covariance between each series of components from each dataset is maximized, as well as their discrimination with regards to the groups of interest, in this case the genotypes. Here, both the first and the second pair of components were highly positively correlated ($r = 0.96$ and 0.95, respectively (Fig. S4)). This indicates that the DIABLO model could highlight common information between the two datasets.

Interpretation DIABLO's results can be interpreted with classical score and correlation circle plots (Fig. 7, see also Fig. S4). The score plots showed similar patterns (Fig. 7a, b). For example, genotype 3 is characterized by (i) compounds C33, C34, C36 and C39 that are overexpressed compared to the other genotypes, and compounds C13 and C22 that are underrepresented in perianths (Fig. 7c); (ii) compounds C33 to C36 (C32 and C37) that are overexpressed (underexpressed) in anthers (Fig. 7d). All these compounds are highly correlated, within and between floral tissues.

More sophisticated graphical outputs were also developed for integrative analyses, and are constructed directly from the components (see details in González *et al*. 2012). Clustered image maps (CIMs) show the level of expression of variables from all datasets, ordered through unsupervised clustering on both samples and compounds simultaneously. Here the CIM reveals clear global patterns (Fig. 8). First, samples were naturally grouped by their genotype, confirming that these genotypes have different chemical signatures. Second, variables clustered into three groups, each containing compounds from both datasets. Third, each cluster of metabolites was associated with one particular genotype: genotype 1 is characterized by an overexpression of metabolites from cluster 2, genotype 2 by an overexpression of metabolites from cluster 1 and genotype 3 by an overexpression of metabolites from cluster 3. Other useful graphical representations include circos plots (Fig. S5, see Singh *et al*. 2016) and relevance networks (Fig. S6, see González *et al*. (2012)).

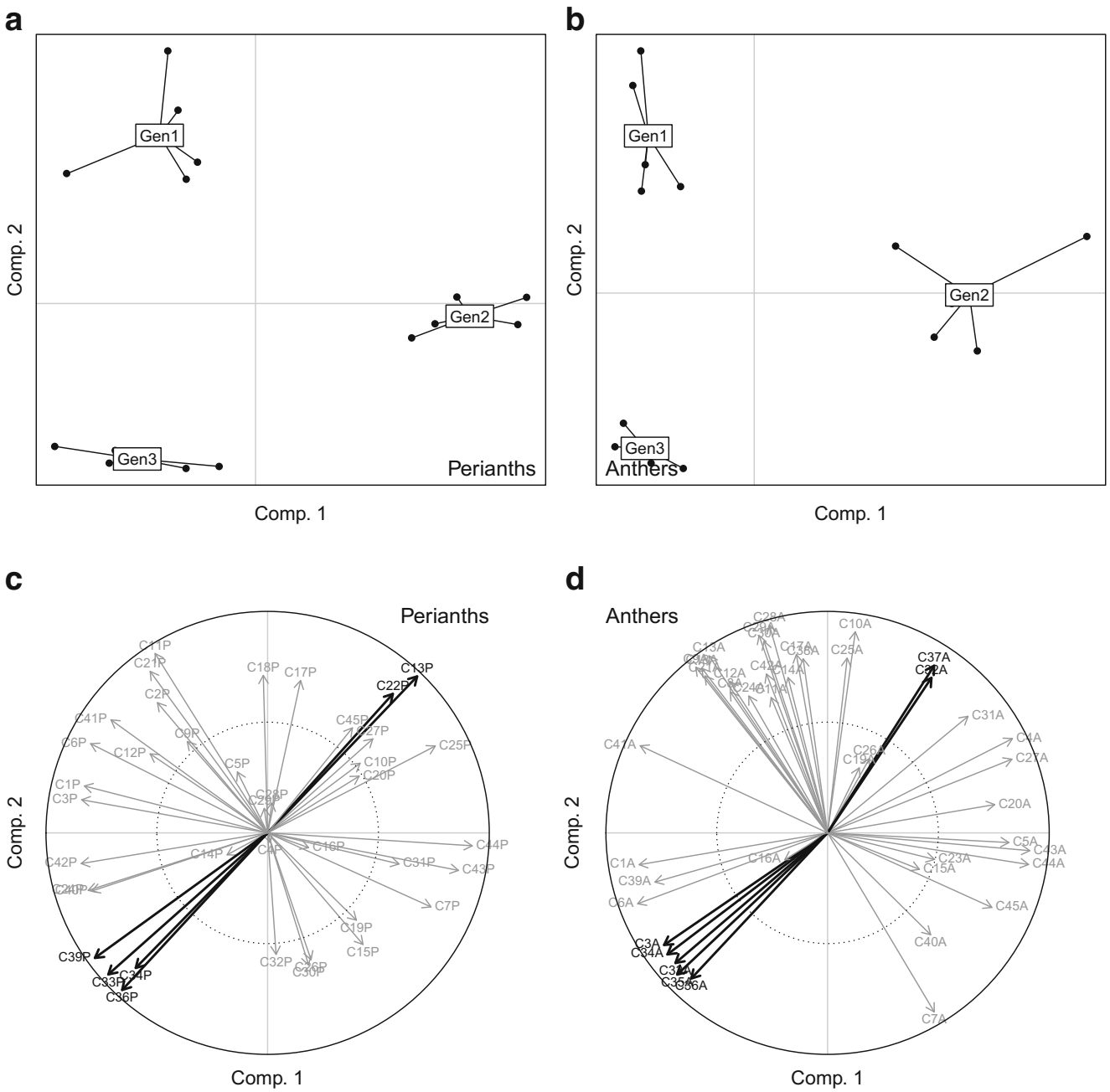


Fig. 7 Example 4: DIABLO graphical outputs. (**a**-**b**): Score plot of each dataset, (**c**-**d**): correlation circles where compounds in dark color are those mentioned in the text

Concluding Remarks

The literature about multivariate analyses is both rich and fast-growing (Meng *et al.* 2016; Zhang *et al.* 2010). The present article introduced a selection of methods that we deemed of particular interest for the field of chemical ecology, whether or not those methods were specifically developed for this particular research field.

The amount of data generated by metabolomics and other omics platforms is drastically increasing with the advances in technology. Each of those datasets monitor thousands of variables. All methods presented in this article are scalable to such large numbers of omics variables. However, the high dimension of these datasets also leads to increased background noise that can mask biologically relevant signals. Several techniques were developed to automatically select variables that are most important for prediction (see the extensive review of Mehmood *et al.* (2012)). Among those techniques, 'sparse' methods were proposed to set a specific number of variables during the modelling, such as sparse PCA (sPCA; Shen and Huang 2008), sPLS-DA (Lê Cao *et al.* 2011), sparse Canonical Correlation Analysis (sCCA; Witten

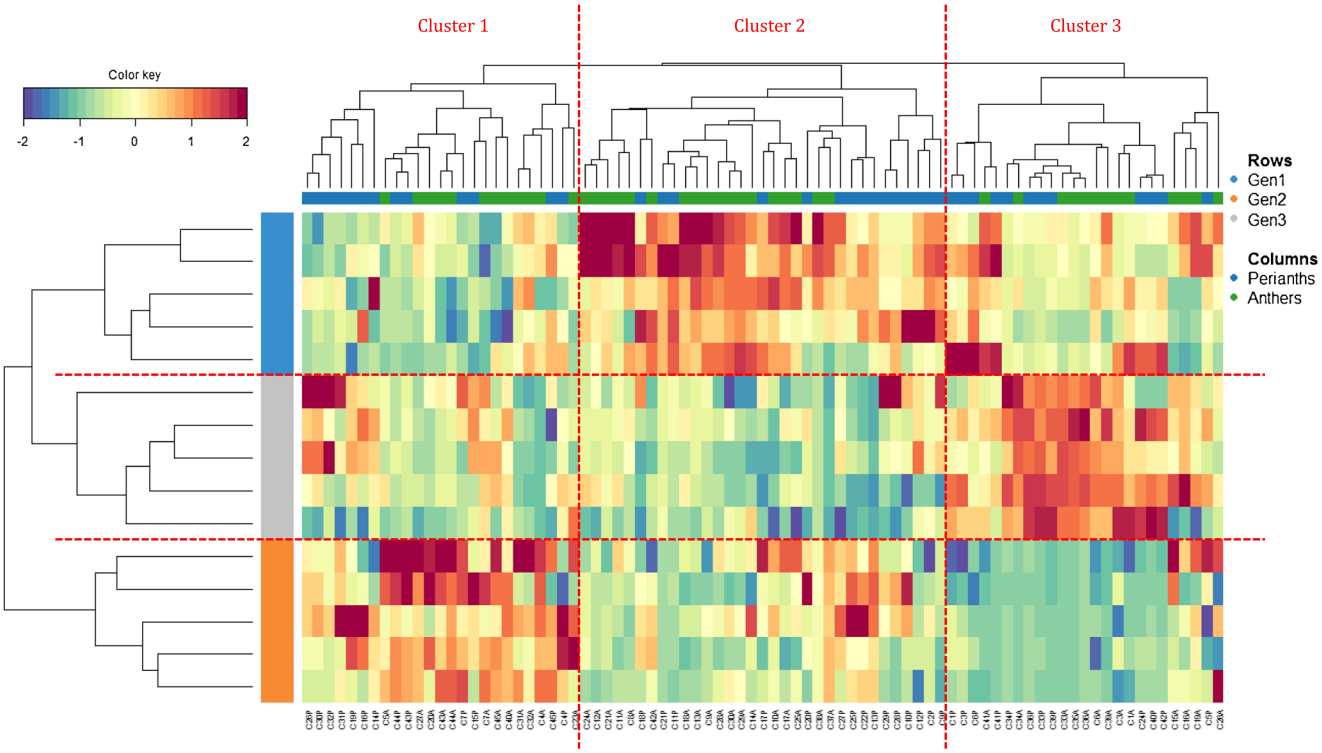


Fig. 8 Example 4: DIABLO Clustered image map. Added dotted lines delimit clusters of samples (horizontal) and compounds (vertical)

and Tibshirani 2009; Witten *et al*. 2009), and sparse Generalized Canonical Correlation Analysis (SGCCA; Tenenhaus *et al*. 2014). DIABLO is also a sparse method that is derived from both Regularized Generalized Canonical Correlation Analysis (RGCCA) and SGCCA – a property that we chose not to illustrate in this article.

Another challenge to overcome in chemical ecology is to adopt a systems biology approach, whereby chemical interactions are studied from multiple perspectives. These can be at different molecular and functional levels (e.g. transcriptomic, proteomic and metabolomic) of a given organism, in several interacting organisms, at different functional levels of interacting organisms etc. In this context, multi-block methods based on Partial Least Squares – Path Modeling (PLS-PM) are useful to define relevant connections between blocks, where integrative analyses are driven by both biological knowledge and hypotheses. Several other methods we presented or mentioned can integrate more than two datasets, such as DIABLO, RGCCA, SGCCA and orthogonal multiblock Partial Least Squares (OnPLS). In addition to ordination techniques, we encourage researchers to consider network analyses, which are complementary, powerful and intuitive tools to represent and understand complex systems. We refer to Tieri *et al*. (2015) for a good starting point and many examples of network-based integrative analyses. However, biologists and chemists should keep in mind that integrating multiple datasets in the same analysis is a very complex task that must be performed with strong biological questions in mind. It should also be considered that they are primarily meant to generate novel hypotheses. Experimental design and, most importantly, validating new analytical algorithms is a challenge that is accentuated by the much larger number of variables than samples in 'omics datasets. Close interactions between biologists, chemists and biostatisticians is therefore required more than ever.

Acknowledgments We are very grateful to Bernard Banaigs, Lucie Conchou, Laurent Dormont, Stéphane Greff, Maria Cristina Lorenzi, Thierry Pérez, Bertrand Schatz, Oriol Sacristán-Soriano and Olivier Thomas who kindly provided their data to illustrate the examples, Stéphane Dray and Denis Poinsot for their insightful comments on the manuscript and Zoe Welham for proof reading of the manuscript.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

Aitchison J (1983) Principal component analysis of compositional data. Biometrika 70:57

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall Ltd, London

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distance. Math Geol 32:271–275

Allaire J, Cheng J, Xie Y, McPherson J, Chang W, Allen J, Wickham H, Atkins A, Hyndman R, Arslan R (2017) Rmarkdown: dynamic documents for R. R package version 1.6. https://CRAN.R-project.org/package=rmarkdown

Archunan G (2009) Vertebrate pheromones and their biological importance. J Exp Zool India 12:227–239

Bais HP, Weir TL, Perry LG, Gilroy S, Vivanco JM (2006) The role of root exudates in rhizosphere interactions with plants and other organisms. Annu Rev Plant Biol 57:233–266

Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemom 17:166–173

Bertrand D, Courcoux P, Autran J-C, Meritan R, Robert P (1990) Stepwise canonical discriminant analysis of continuous digitalized signals: application to chromatograms of wheat proteins. J Chemom 4:413–427

Bonelli M, Lorenzi MC, Christidès J-P, Dupont S, Bagnères A-G (2015) Population diversity in Cuticular hydrocarbons and mtDNA in a mountain social wasp. J Chem Ecol 41:22–31

Brereton RG, Lloyd GR (2014) Partial least squares discriminant analysis: taking the magic away. J Chemom 28:213–225

Brückner A, Heethoff M (2017) A chemo-ecologists' practical guide to compositional data analysis. Chemoecology 27:33–46

Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom 20:341–351

Chessel D, Hanafi M (1996) Analyses de la co-inertie de K nuages de points. Rev Stat Appliquée 44:35–60

Conchou L, Cabioch L, Rodriguez LJV, Kjellberg F (2014) Daily rhythm of mutualistic pollinator activity and scent emission in Ficus Septica: ecological differentiation between co-occurring pollinators and potential consequences for chemical communication and facilitation of host speciation. PLoS One 9:e103581

Després L, David J-P, Gallet C (2007) The evolutionary ecology of insect resistance to plant chemicals. Trends Ecol Evol 22:298–307

Dolédec S, Chessel D (1994) Co-inertia analysis: an alternative method for studying species–environment relationships. Freshw Biol 31: 277–294

Dormont L, Delle-Vedove R, Bessière J-M, Schatz B (2014) Floral scent emitted by white and coloured morphs in orchids. Phytochemistry 100:51–59

Dray S, Chessel D, Thioulouse J (2003a) Procrustean co-inertia analysis for the linking of multivariate datasets. Écoscience 10:110–119

Dray S, Chessel D, Thioulouse J (2003b) Co-inertia analysis and the linking of ecological data tables. Ecology 84:3078–3089

Engel J, Gerretzen J, Szymańska E, Jansen JJ, Downey G, Blanchet L, Buydens LMC (2013) Breaking with trends in pre-processing? TrAC Trends Anal Chem 50:96–106

Engel J, Blanchet L, Bloemen B, van den Heuvel LP, Engelke UHF, Wevers RA, Buydens LMC (2015) Regularized MANOVA (rMANOVA) in untargeted metabolomics. Anal Chim Acta 899: 1–12

Escoufier Y (1973) Le Traitement des Variables Vectorielles. Biometrics 29:751

Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. Environmetrics 20:621–632

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugenics 7:179–188

Gatehouse JA (2002) Plant resistance towards insect herbivores: a dynamic interaction. New Phytol 156:145–169

González I, Lê Cao K-A, Davis MJ, Déjean S (2012) Visualising associations between paired "omics" data sets. BioData Min 5:19

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338

Gower JC (1971) Statistical methods of comparing different multivariate analyses of the same data. In: Tautu P (ed) Mathematics in the archaeological and historical sciences. Edinburgh University Press, Edinburgh, pp 138–149

Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. J Classif 3:5–48

Greff S, Aires T, Serrão EA, Engelen AH, Thomas OP, Pérez T (2017) The interaction between the proliferating macroalga Asparagopsis Taxiformis and the coral Astroides Calycularis induces changes in microbiome and metabolomic fingerprints. Sci Rep 7:42625

Harrington P d B, Vieira NE, Espinoza J, Nien JK, Romero R, Yergey AL (2005) Analysis of variance–principal component analysis: a soft tool for proteomic discovery. Anal Chim Acta 544:118–127

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Spinger, New York

Heo M, Gabriel KR (1998) A permutation test of association between configurations by means of the rv coefficient. Commun Stat Simul Comput 27:843–856

Hervé MR, Delourme R, Gravot A, Marnet N, Berardocco S, Cortesero AM (2014) Manipulating feeding stimulation to protect crops against insect pests? J Chem Ecol 40:1220–1231

Hill MO, Smith AJE (1976) Principal component analysis of taxonomic data with multi-state discrete characters. Taxon 25:249

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Educ Psychol 24:417–441

Hotelling H (1936) Relations between two sets of variates. Biometrika 28(377):321

Howard RW, Blomquist GJ (2005) Ecological, behavioral, and biochemical aspects of insect hydrocarbons. Annu Rev Entomol 50:371–393

Indahl UG, Martens H, Næs T (2007) From dummy regression to prior probabilities in PLS-DA. J Chemom 21:529–536

Indahl UG, Liland KH, Naes T (2009) Canonical partial least squares-a unified PLS approach to classification and regression problems. J Chemom 23:495–504

Ivanišević J, Thomas OP, Lejeusne C, Chevaldonné P, Pérez T (2011) Metabolic fingerprinting as an indicator of biodiversity: towards understanding inter-specific relationships among Homoscleromorpha sponges. Metabolomics 7:289–304

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bull Soc Vaud Sci Nat 37:547–579

Jackson DA (1995) PROTEST: a PROcrustean randomization TEST of community environment concordance. Écoscience 2:297–303

Jansen JJ, Hoefsloot HCJ, van der Greef J, Timmerman ME, Westerhuis JA, Smilde AK (2005) ASCA: analysis of multivariate data obtained from an experimental design. J Chemom 19:469–481

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94

Kemsley EK (1996) Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. Chemom Intell Lab Syst 33:47–61

Kjeldahl K, Bro R (2010) Some common misunderstandings in chemometrics. J Chemom 24:558–564

Kruskal JB (1964a) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27

Kruskal JB (1964b) Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:115–129

Lê Cao K-A, Boitard S, Besse P (2011) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinf 12:253

Legendre P, Anderson MJ (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecol Monogr 69(1)

Legendre P, Legendre L (2012) Numerical Ecology. Elsevier, Amsterdam

Leurgans SE, Moyeed RA, Silverman BW (1993) Canonical correlation analysis when the data are curves. J R Stat Soc Ser B Methodol 55: 725–740

Liland KH, Indahl UG (2009) Powered partial least squares discriminant analysis. J Chemom 23:7–18
Liquet B, Lê Cao K-A, Hocini H, Thiébaut R (2012) A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. BMC Bioinformatics 13:325
Löfstedt T, Trygg J (2011) OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation. J Chemom 25: 441–455
Löfstedt T, Hanafi M, Mazerolles G, Trygg J (2012) OnPLS path modelling. Chemom Intell Lab Syst 118:139–149
Löfstedt T, Hoffman D, Trygg J (2013) Global, local and unique decompositions in OnPLS for multiblock data analysis. Anal Chim Acta 791:13–24
Lohmöller J (1989) Latent variables path modeling with partial least squares. Physica-Verlag, Heidelberg
Marini F, de Beer D, Joubert E, Walczak B (2015) Analysis of variance of designed chromatographic data sets: the analysis of variance-target projection approach. J Chromatogr A 1405:94–102
Mehmood T, Liland KH, Snipen L, Sæbø S (2012) A review of variable selection methods in partial least squares regression. Chemom Intell Lab Syst 118:62–69
Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform 17:628–641
Miller J, Farr S (1971) Bimultivariate redundancy: a comprehensive measure of interbattery relationship. Multivar Behav Res 6:313–324
Nocairi H, Mostafa Qannari E, Vigneau E, Bertrand D (2005) Discrimination on latent components with respect to patterns. Application to multicollinear data. Comput Stat Data Anal 48: 139–147
Palarea-Albaladejo J, Martín-Fernández JA, Soto JA (2012) Dealing with distances and transformations for fuzzy C-means clustering of compositional data. J Classif 29:144–169
Pearson K (1896) Mathematical contributions to the theory of evolution - on a form of spurious correlation which may Arise when indices are used in the measurement of organs. Proc R Soc Lond 60:489–498
Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2:559–572
Peres-Neto PR, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614–2625
Pierotti MER, Martín-Fernández JA (2011) Compositional analysis in behavioural and evolutionary ecology. In: Pawloswky-Glahn V, Buccianti A (eds) Compositional data analysis: theory and applications. John Wiley & Sons, Ltd, Hoboken, pp 218–234
R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Raguso RA (2008) Wake up and smell the roses: the ecology and evolution of floral scent. Annu Rev Ecol Evol Syst 39:549–569
Rao CR (1964) The use and interpretation of principal component analysis in applied research. Sankhyā Indian J Stat Ser A 329–358
Reudler JH, Elzinga JA (2015) Photoperiod-induced geographic variation in plant defense chemistry. J Chem Ecol 41:139–148
Robert P, Escoufier Y (1976) A unifying tool for linear multivariate statistical methods: the RV- coefficient. Appl Stat 25:257
Rohart F, Gautier B, Singh A, Le Cao K-A (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 13(11):e1005752
Saccenti E, Hoefsloot HCJ, Smilde AK, Westerhuis JA, Hendriks MMWB (2014) Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics 10:361–374
Sacristán-Soriano O, Banaigs B, Casamayor EO, Becerro MA (2011) Exploring the links between natural products and bacterial assemblages in the sponge Aplysina aerophoba. Appl Environ Microbiol 77:862–870
Sampson PD, Streissguth AP, Barr HM, Bookstein FL (1989) Neurobehavioral effects of prenatal alcohol: part II. Partial least squares analysis. Neurotoxicol Teratol 11:477–491
Shen H, Huang JZ (2008) Sparse principal component analysis via regularized low rank matrix approximation. J Multivar Anal 99:1015–1034
Shepard RN (1962) The analysis of proximities: multidimensional scaling with an unknown distance function. II. Psychometrika 27:219–246
Singh A, Gautier B, Shannon CP, Vacher M, Rohart F, Tebutt SJ, Le Cao K-A (2016) DIABLO-an integrative, multi-omics, multivariate method for multi-group classification. BioRxiv 067611. https://doi.org/10.1101/067611
Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers R-JAN, van der Greef J, Timmerman ME (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics 21:3043–3048
Smit S, van Breemen MJ, Hoefsloot HCJ, Smilde AK, Aerts JMFG, de Koster CG (2007) Assessing the statistical validity of proteomics based biomarkers. Anal Chim Acta 592:210–217
Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409–1438
Ståhle L, Wold S (1987) Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. J Chemom 1:185–196
Szymańska E, Saccenti E, Smilde AK, Westerhuis JA (2012) Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics 8:3–16
Tapp HS, Kemsley EK (2009) Notes on the practical utility of OPLS. TrAC Trends Anal Chem 28:1322–1327
Tenenhaus A, Tenenhaus M (2011) Regularized generalized canonical correlation analysis. Psychometrika 76:257–284
Tenenhaus M, Young FW (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50:91–119
Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V (2014) Variable selection for generalized canonical correlation analysis. Biostatistics 15:569–583
Tholl D, Boland W, Hansel A, Loreto F, Röse USR, Schnitzler J-P (2006) Practical approaches to plant volatile analysis. Plant J 45:540–560
Tieri P, Nardini C, Dent JE (2015) Multi-omic data integration. Frontiers Media SA, Lausanne
Trygg J (2002) O2-PLS for qualitative and quantitative analysis in multivariate calibration. J Chemom 16:283–293
Trygg J, Wold S (2003) O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. J Chemom 17: 53–64
Tseng G, Ghosh D, Zhou X (2015) Integrating omics data. Cambridge University Press, Cambridge
van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7:142
Van Den Wollenberg AL (1977) Redundancy analysis an alternative for canonical correlation analysis. Psychometrika 42:207–219
van Velzen EJJ, Westerhuis JA, van Duynhoven JPM, van Dorsten FA, Hoefsloot HCJ, Jacobs DM, Smit S, Draijer R, Kroner CI, Smilde AK (2008) Multilevel data analysis of a crossover designed human nutritional intervention study. J Proteome Res 7:4483–4491
Vinod HD (1976) Canonical ridge and econometrics of joint production. J Econ 4:147–166
Volkman JK, Barrett SM, Blackburn SI, Mansour MP, Sikes EL, Gelin F (1998) Microalgal biomarkers: a review of recent research developments. Org Geochem 29:1163–1179

Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven JPM, van Dorsten FA (2008) Assessment of PLSDA cross validation. Metabolomics 4:81–89

Westerhuis JA, van Velzen EJJ, Hoefsloot HCJ, Smilde AK (2010) Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics 6:119–128

Witten DM, Tibshirani RJ (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol 8:1–27

Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10:515–534

Wold H (1985) Partial least squares. In: Kotz S, Johnson N (eds) Encyclopedia of statistical sciences. Wiley, New York, pp 581–591

Wold S, Martens H, Wold H (1983) The multivariate calibration problem in chemistry solved by the PLS method. In Matrix Pencils, (Springer), pp. 286–293

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130

Worley B, Powers R (2013) Multivariate analysis in metabolomics. Curr Metabolomics 1:92–107

Zerzucha P, Daszykowski M, Walczak B (2012) Dissimilarity partial least squares applied to non-linear modeling problems. Chemom Intell Lab Syst 110:156–162

Zhang W, Li F, Nie L (2010) Integrating multiple “omics” analysis for microbial biology: application and methodologies. Microbiology 156:287–301