Analytica Chimica Acta 879 (2015) 10–23

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Tutorial

A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding

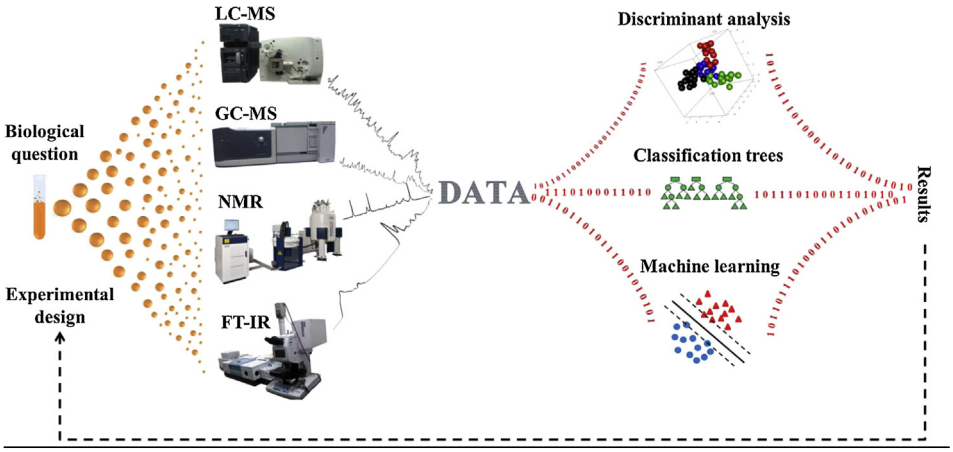
Piotr S. Gromski [a], Howbeer Muhamadali [a], David I. Ellis [a], Yun Xu [a], Elon Correa [a], Michael L. Turner [b], Royston Goodacre [a,*]

[a] School of Chemistry, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[b] School of Chemistry, Brunswick Street, The University of Manchester, Manchester M13 9PL, UK

HIGHLIGHTS

- PLS-DA, PC-DFA, SVM and RF analyses were compared for metabolomics analyses.
- Parsimonious models for feature selection and data reduction were presented.
- Comparisons include generally recognized pros along with specific caveats for each of the methods.
- Statistical models applied in the analysis of metabolomics data were shown.
- Pros and cons of common analytical techniques used in metabolomics studies are highlighted.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 11 October 2014
Received in revised form 3 February 2015
Accepted 6 February 2015
Available online 11 February 2015

Keywords:
Metabolomics
Chemometrics
Partial least squares-discriminant analysis
Principal component-discriminant function analysis
Support vector machines
Random forests

ABSTRACT

The predominance of partial least squares-discriminant analysis (PLS-DA) used to analyze metabolomics datasets (indeed, it is the most well-known tool to perform classification and regression in metabolomics), can be said to have led to the point that not all researchers are fully aware of alternative multivariate classification algorithms. This may in part be due to the widespread availability of PLS-DA in most of the well-known statistical software packages, where its implementation is very easy if the default settings are used. In addition, one of the perceived advantages of PLS-DA is that it has the ability to analyze highly collinear and noisy data. Furthermore, the calibration model is known to provide a variety of useful statistics, such as prediction accuracy as well as scores and loadings plots. However, this method may provide misleading results, largely due to a lack of suitable statistical validation, when used by non-experts who are not aware of its potential limitations when used in conjunction with metabolomics. This tutorial review aims to provide an introductory overview to several straightforward statistical methods such as principal component-discriminant function analysis (PC-DFA), support vector machines (SVM) and random forests (RF), which could very easily be used either to augment PLS or as alternative supervised learning methods to PLS-DA. These methods can be said to be particularly appropriate for the analysis of large, highly-complex data sets which are common output(s) in metabolomics studies where the numbers of variables often far exceed the number of samples. In addition, these alternative

* Corresponding author. Tel.: +44 0 161 306 4480.
E-mail address: roy.goodacre@manchester.ac.uk (R. Goodacre).

http://dx.doi.org/10.1016/j.aca.2015.02.012
0003-2670/© 2015 Elsevier B.V. All rights reserved.

techniques may be useful tools for generating parsimonious models through feature selection and data reduction, as well as providing more propitious results. We sincerely hope that the general reader is left with little doubt that there are several promising and readily available alternatives to PLS-DA, to analyze large and highly complex data sets.

© 2015 Elsevier B.V. All rights reserved.

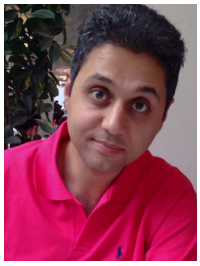
Contents

1. Introduction . . . 12
2. Technologies used for generation of metabolomics data . . . 12
3. Data analysis applied in metabolomics . . . 14
   3.1. Partial least squares-discriminant analysis (PLS-DA) . . . 14
   3.2. Principal components-discriminant function analysis (PC-DFA) . . . 15
   3.3. Support vector machines (SVM) . . . 17
   3.4. Random forests (RF) . . . 18
   3.5. Parameters selection . . . 19
4. Conclusions . . . 20
   Acknowledgments . . . 20
   References . . . 20

Piotr S. Gromski is a Research Associate in the Department of Pure and Applied Chemistry in the Faculty of Science at University of Strathclyde Glasgow, UK. He obtained an M.Sc degree from Kazimierz Pułaski Technical University of Radom (now Kazimierz Pułaski University of Technology and Humanities in Radom) in 2007. In 2011 he joined Prof. Roy Goodacre's group at the School of Chemistry, University of Manchester, UK as a Ph.D student to study "Application of chemometrics for the robust analysis of chemical and biochemical data". His main interest covers chemometrics and bioinformatics.

David Ellis was educated on the Welsh coast at the University of Wales, Aberystwyth, obtaining a B.Sc in Environmental Science and a Ph.D in Analytical Biotechnology/Microbiology. His research involving the rapid and quantitative detection of foodborne bacteria using FT-IR and machine learning has been widely publicised, featuring on BBC TV and radio, the Science Museum in London, as well as the national and international press (*e.g.*, WIRED). He now works at the Manchester Institute of Biotechnology (MIB), as a Senior Experimental Officer undertaking research and managing the labs of Roy Goodacre (biospec.net) and Doug Kell (dbkgroup.org), in the School of Chemistry, University of Manchester, UK.

Elon Correa is a Medical Statistician at the Faculty of Medical & Human Sciences Institute of Human Development, University of Manchester, UK. He gained his first degree in Mathematics at the Pontifical Catholic University of Parana, Brazil followed by his M.Sc in Numerical Methods in Engineering at the Federal University of Parana, Brazil and his Ph.D in Computer Science which was carried out at the University of Manchester, UK. His main research area is on the development of models of health outcomes involving complex biological systems from large datasets as well as the study of advanced hybrid bio-statistical and machine learning approaches to identify latent classes of disease risk and health outcomes.



Howbeer Muhamadali received a first class honours in Microbiology at the Manchester Metropolitan University in 2010. In 2011, he successfully completed an M.Phil in geomicrobiology at the University of Manchester under the supervision of Prof. Jonathan Lloyd, working on the optimisation and scale-up of a batch culture magnetite nanoparticle production bioprocess. He is currently in his third year of Ph.D (Biotechnology/Metabolomics), in Prof. Roy Goodacre's group at the University of Manchester. His project involves the metabolomics investigations of different microbial bioprocesses, such as *E. coli*,*Streptomyces* and *Geobacter* species, using various analytical techniques such as FT-IR, GC–MS, DIMS and multivariate statistical analysis approaches.

Yun Xu was educated at Tongji Medical University (now Tongji Medical College of Huazhong University of Science and Technology) and obtained a B.Sc in Pharmaceutical Analytical Chemistry, he then obtained his Ph.D in chemometrics at University of Bristol and is a post-doctoral researcher in Professor Roy Goodacre's group (http://www.biospec.net) at University of Manchester. His area of research is involved with multivariate statistics, regression, pattern recognition and machine learning with applications to metabolomics studies.

Mike Turner is Professor of Materials Chemistry (www.omec.org.uk/MLTurner/) and Director of the Organic Materials Innovation Centre (www.omic.org.uk) within the School of Chemistry, University of Manchester, UK. His principal research interests concern the synthesis of organic semiconductors and the use of these molecules in organic electronic and electrooptical devices such as organic transistors, organic light emitting diodes, sensors and solar cells. He is principal investigator for the Knowledge Centre for Materials Chemistry (www.materialschemistry.org) at the University of Manchester.

Roy Goodacre is Professor of Biological Chemistry at the School of Chemistry, The University of Manchester, UK. His group's main areas of research (http://www.biospec.net/) are broadly within analytical biotechnology, metabolomics and systems biology. His expertise involves mass spectrometry, FT-IR and Raman spectroscopy, as well as advanced chemometrics and machine learning. He is Editor-in-Chief of the journal *Metabolomics*, and on the Editorial Advisory Boards of *Analyst* and *Journal of Analytical and Applied Pyrolysis*. He is also a founding director of the *Metabolomics Society* and a director of the *Metabolic Profiling Forum*.

1. Introduction

Partial least squares-discriminant analysis (PLS-DA) [1] is one of the most well-known classification procedures in chemometrics [2]. This approach has also been extensively used in "omics" related fields, for example: metabolomics [3], proteomics [4], genomics [5,6] and in many other fields which generates large amounts of data such as spectroscopy, as described elsewhere [7]. The increasing interest in PLS-DA, especially in the field of metabolomics, is mainly due to its availability in most of the common statistical software packages such as: *R*, S-Plus, SAS, SPSS, MATLAB [7–13]. However, as recently reported by Brereton and Lloyd this method can often be misused by non-experts [2], largely due to its prevalence in commercially available packages where the non-cognoscenti can use the default setting for model construction and validation which may not be appropriate. Therefore, this review is inspired by the growing popularity of PLS-DA in metabolomics as evident from the rapidly increasing number of publications within the field that are almost proportionate over the last 10 years with the number of articles published in metabolomics (Fig. 1). This has been estimated using keywords such as: metabolomics, PLS and PLS-DA. However, a large disproportion can be observed between PLS-DA and metabolomics in terms of the number of published papers. Similar statistics were observed nearly a decade ago by Broadhurst and Kell, where these authors highlighted the fact that, in most cases, the statistical approaches that have been employed are not described in sufficient detail, nor are the models appropriately validated [14]. This is despite the fact that certain recommendations have been provided by the metabolomics standards initiative (MSI) [15–17], in terms of reporting research studies.

As recently reported the global market for bioinformatics (*e.g.*, genomics, molecular phylogenetics, metabolomics, proteomics, chemoinformatics and drug design) is expected to reach 12.48 billion USD by 2020, according to a new study by Grand View Research, Inc. [18]. In addition, bioinformatics-based metabolomics applications are expected to grow at the highest compound annual growth rate (CAGR) of over 23%. Therefore, the objective of this review is to introduce the reader to some of the alternative chemometrics approaches to PLS-DA. The algorithms presented here are recognized as robust supervised learning techniques which have been applied in various scientific fields and therefore can be successfully applied for the analysis of metabolomics data as shown in this study. Since metabolomics deals with large and highly complex datasets [19] here, we deliver what we would like to consider an introductory and essential explanation targeted, in some degree, toward researchers working in the exciting field of metabolomics, as well as others working with large and highly complex datasets (*e.g.*, spectroscopy, multi-readout sensors, and any data where the number of collected variables is large).

There are a variety of chemometrics methods available that can be used in metabolomics and many of them can be found in the literature [20–26]. In this article we will refer to the following examples: principal component-discriminant function analysis (PC-DFA) [27], support vector machines (SVM) [28,29] and random forests (RF) [30], as these are starting to be used within the metabolomics field. As these methods have been comprehensively described elsewhere [27,30–41], here we shall avoid in-depth and complex descriptions of the mathematics behind these classification models and instead provide simple explanations which include both the advantages and caveats as well as practical application within the field of metabolomics. We shall start by first introducing the different types of data and data structures that are generated by typical metabolomics experiments.

2. Technologies used for generation of metabolomics data

The term metabolome was first described by Oliver *et al.* [42] as the complete set of low molecular weight compounds present in a cell that are required for its maintenance, growth and normal function and contributes to the metabolic reactions of a cell in a particular physiological or developmental stage [43]. As metabolites are down-stream products of gene transcription and translation (proteins), compared to genomic and proteomics, metabolomics approaches can provide a clearer picture of the phenotype of a biological system. However, compared to the genome and proteome the metabolome is much more complex, for example the plant kingdom as a whole is estimated to contain 200,000 or more metabolites and phytochemicals [43]; moreover, there are currently 41,815 metabolite entries on the human metabolome database [44], a figure which is increasing and which does not include many metabolites found in humans (in particular the plethora of lipids within the lipidome; http://www.lipidmaps.org) and which may not as yet been registered in such databases [45].

The specific physicochemical properties of different groups of metabolites further adds to the complexity of metabolomics studies, which have been the driving force behind the development of various protocols, as well as the application of a wide-range of analytical platforms [46]. However, due to the complexity of the metabolome one method does not fit all [21] and covering this wide-range of metabolites requires the use of multiple protocols and analytical instruments [47–49].

The high precision of mass spectrometry (MS) and the reproducibility of nuclear magnetic resonance (NMR) spectroscopy combined with their ability to elucidate chemical structures [50–52] have resulted in these being the most common analytical technologies employed for metabolomics studies. However, vibrational spectroscopic techniques such as Raman and Fourier transform infrared (FT-IR) have recently gained in popularity in metabolomics [53–55], used for disease diagnostics [56–59],

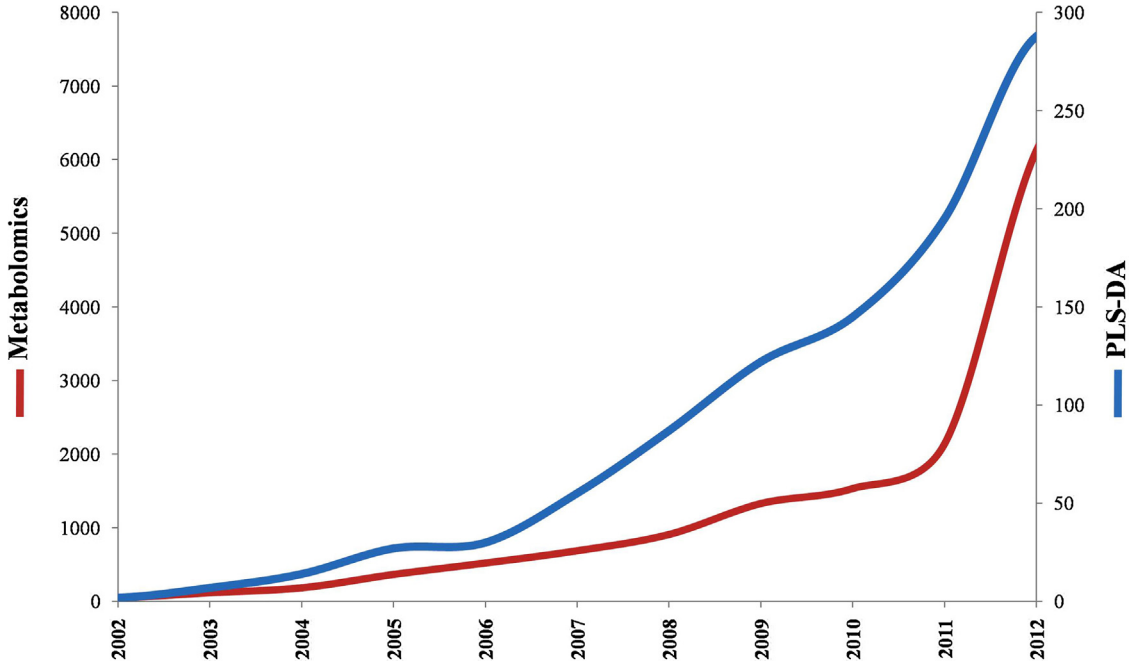


Fig. 1. Number of publications in the field of metabolomics per year (red line) *versus* the number of publication that include PLS-DA as a tool for the analysis of metabolomics data (blue line) (from Thomas Reuters' ISI Web of Science® using the keywords [metabolomics; PLS* and PLS-DA*]). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and these are able to effect spatial measurements without destroying the sample [60,61]. Table 1 describes the advantages and disadvantages of some of the most common analytical techniques used in metabolomics approaches; for a more detailed discussion on instrumentation and their applications the reader is directed to the following papers [46,49,62–65].

Depending on the complexity of the samples, chromatographic separation techniques such as liquid (LC) or gas chromatography (GC) can be coupled to MS to enhance resolution [62]. A recent study [66] employed GC–MS, LC–MS and NMR for the analysis of human blood resulting in the quantification, identification and validation of more than 4000 serum and plasma metabolites, followed by an extensive survey of the literature yielding an extra 665 complementary metabolites. The collected information was combined into a single repository forming the most comprehensive electronically accessible database containing 4229 serum/plasma compounds and their reported links to different diseases.

Bearing in mind the intricate features of the metabolome combined with the ever improving capabilities of metabolomics technologies and platforms and the development of more efficient protocols (recovery and detection of even more metabolites), not even considering the different experimental conditions, one can imagine the vast amount of data generated by metabolomics studies. Another example is the recent study of over 1000 healthy

Table 1
Typical advantages and disadvantages of common analytical techniques used in metabolomics studies.

Techniques	Advantages	Disadvantages
GC–MS	• High resolving power and accuracy • Reproducible retention time • Applicable to volatile and semi-volatile compounds (through derivatization) • low cost compared to LC–MS and NMR • Comprehensive databases available	• Volatility can be a restriction • Heat sensitive compounds cannot generally be analysed • Derivatization may complicate sample preparation and identification (due to additives and multiple derivative products)
LC–MS	• High accuracy, resolving power, sensitivity and specificity • Sample preparation is minimal compared to GC–MS (generally no derivatization required) • Applicable to complex mixtures, polar and non-polar compounds	• Without comprehensive MS–MS or MS^n structural information is limited • Matrix effects • Formation of multiple adducts
NMR	• Robust and highly reproducible • Provides very specific structural information • Non-destructive (samples can be recovered) • Minimal sample preparation • Highly quantitative	• Very expensive • Spectral interpretation is time consuming • Low sensitivity (micromolar range) compared to MS (picomolar range)
FT-IR	• Rapid and high throughput • Relatively inexpensive • Provides information rich data	• Water is an issue in mid IR (samples must be dried), although wet samples can be analysed using attenuated total reflectance (ATR) • Mixtures may complicate data interpretation • Not all compounds can be detected • Low chemical specificity

individuals using GC–MS and LC–MS analysis of the human serum metabolome with the goal of starting to define the molecular phenotype of healthy populations in the UK [67]. Therefore, visualization of such complex data sets requires the aid of robust statistical analysis techniques and strategies for its transformation to interpretable biological knowledge which can be linked to metabolism [20,21].

3. Data analysis applied in metabolomics

The main steps in metabolomics data analysis comprise pre-processing, pre-treatment, processing, post-processing, validation and finally, interpretation. It would be remiss of us not to mention the need for adequate statistical design for sample collection before data are even collected. This is often overlooked until after the data are collected! And is of course necessary to ensure that the only non-random effect is the feature (*e.g.*, disease *versus* healthy) that one is trying to predict [14]. Here we will focus on the 'processing' step in the analysis of metabolomics data as the other steps are reviewed in more detail elsewhere [7,22,68–76]. A typical metabolomics study begins with a biological question which, in the case of clinical studies for example, usually encompasses two groups of individuals, often called case and control (e.g., disease/healthy) [14]. This experimental design (Fig. 2) is followed by the selection and employment of appropriate techniques that will extract useful information from the samples, and numerous statistical techniques can be used. These may include discriminant analysis, classification trees and machine learning. However, each of these techniques has related advantages and disadvantages. Thus, a combination of different analytical technologies and statistical tools must be used to gain a comprehensive interpretation of the results. This may of course include traditional univariate statistical testing as detailed in [14] and we shall not focus on these methods here.

The analysis of metabolomics data is perhaps the most challenging and time consuming step in the processing pipeline, therefore careful selection of appropriate statistical techniques needs to be considered. This subject has been repeatedly reported in the literature [14,21–24,68,69] with classification difficulties in metabolomic analyses being mainly a result of high data dimensionality, where the generated data contains many variables and insufficient number of samples. In addition, one of the main challenges that data analysts encounter when analyzing metabolomics data is the over-fitting of the model, meaning that the chosen statistical approach classifies the training data too well, but subsequent samples are very often incorrectly classified [77]; that is to say the model is unable to 'generalize' because it has learnt the training data perfectly. Other hurdles may be related to the analytical technology used to generate the data, as well as statistical methods used for the analysis. Therefore, robust statistical analyses that allow for the generation of accurate results are required.

3.1. Partial least squares-discriminant analysis (PLS-DA)

PLS-DA is a chemometrics technique used to optimise separation between different groups of samples, which is accomplished by linking two data matrices *X* (*i.e.*, raw data) and *Y* (*i.e.*, groups, class membership *etc.*). The method is in fact an extension of PLS1 which handles single dependent continues variable whereas PLS2 (called PLS-DA) can handle multiple dependent categorical variables [1,37]. This approach aims to maximize the covariance between the independent variables *X* (sample readings; that is to say the metabolomics data) and the corresponding dependent variable *Y* (classes, groups; that is to say the targets that one wants to predict) of highly multidimensional data by finding a linear subspace of the explanatory variables. This new subspace permits the prediction of the *Y* variable based on a reduced number of factors (PLS components, or what are also known as latent variables). These factors describe the behavior of dependent variables *Y* and they span the subspace onto which the independent variables *X* are projected. For example, if we consider the division of samples into different classes, such as case-control study (Fig. 3A), then the variable *Y* will comprise of a single vector which will have an entry 0 for all samples in the first class and an entry 1 for all samples in the second class (or *vice versa*, it does not matter). When the data contain three classes then the three groups will be binary encoded in 3 variables with the *Y* matrix as [1 0 0] for all samples from class 1, [0 1 0] for samples from class 2, and finally [0 0 1] for samples from class 3, as is illustrated in Fig. 3B [1,37].

The main advantage of this PLS-DA approach is the availability and handling of highly collinear and noisy data, which are very common outputs from metabolomics experiments [11]. In addition, this provides several statistics such as loading weight, variable importance on projection (VIP) and regression coefficient

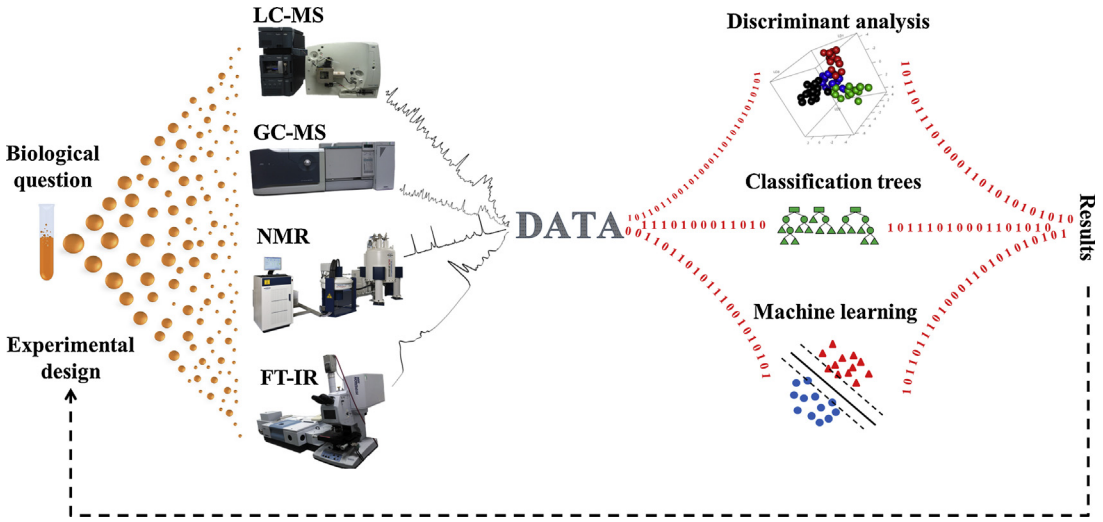


Fig. 2. A graphical representation of the different analytical approaches and informatics techniques employed in metabolomics studies. A standard experimental design in a metabolomics study includes sample preparation and data acquisition using for instance LC–MS, GC–MS, NMR or FT-IR. Once the data have been generated, several types of univariate statistical (not depicted here) and multivariate algorithms such as: discriminant analysis, classification trees and machine learning can be applied to reduce the dimensionality of the data and facilitate biological interpretation, and which could also improve the design of subsequent or future experiments.

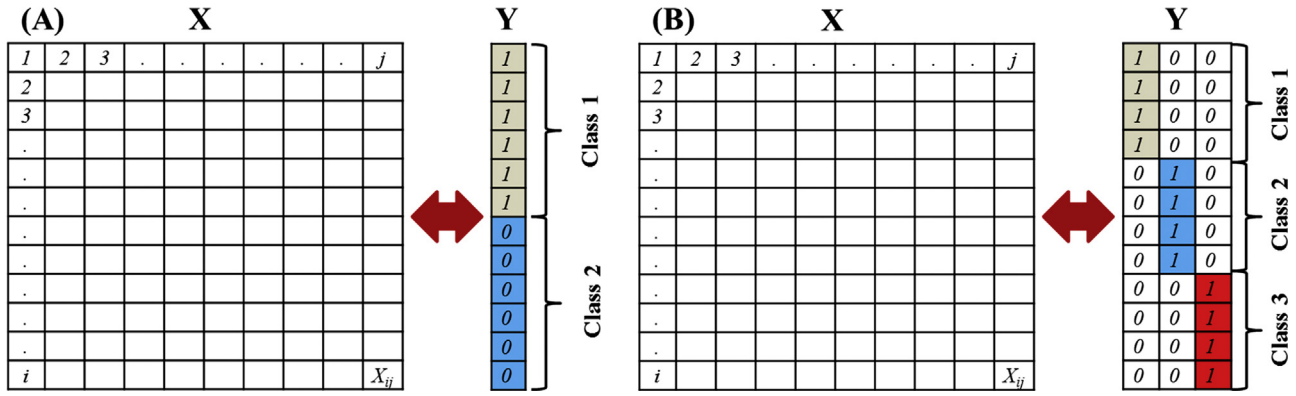


Fig. 3. An illustration of partial least squares-discriminant analysis (PLS-DA) for models that includes (A) two classes and (B) three classes. When two classes are considered the *Y* corresponds to a vector else to the matrix where binary encoding is used on the *Y* vectors. In the matrices show for the *X* input data there are *i* samples and *j* metabolites and the *Y* output will have n samples (where *n* is the number of targets to predict) and will be of length *i* (*i.e.*, it is the same as the number of samples in *X*).

that can be used to identify the most important variables [78–80]. This technique provides a visual interpretation of complex datasets through a low-dimensional, easily interpretable scores plot that illustrates the separation between different groups [81]. Comparison of loadings and scores plot supports investigations in terms of the relationship between important variables that can be specific to the group of interest [1,37,82]. However, more recently scores plots have been questioned in terms of their strength [38].

To date, there are several caveats associated with PLS-DA that need to be considered when using this model, which have been reported in the literature [38,39,77,83,84]. These include: (1) difficulties in the identification of small numbers of variables that are responsible for the separation between two or more groups (classes) and therefore a larger number of variables are required to achieve a good prediction accuracy; (2) in many cases when the number of variables significantly exceeds the number of samples, the model is likely to lead to good classifications by chance (tendency to over-fit); (3) quoting R^2 and Q^2 are questionable statistics to describe the predictive ability of classification models as these two statistics were designed for regression models and not categorical models [85]; and finally, (4) Westerhuis *et al.* advocate against the use of PLS-DA scores plots for inference of class differences as it provides an over-optimistic understanding of the separation between two or more classes, whilst similar results can be achieved when random data are classified [38]. Paradoxically, many researchers have ignored the *Y* variable and these target vectors should be inspected as they can be used to assess the accuracy of the predictions. That is to say, if one wanted to predict a case-control (1 *versus* 0; Fig. 3A) then predicted values from validation data should be close to 0 and 1 for controls and cases respectively, if predictions are close to 0.5 then the model has insufficient discriminatory power and should not be used for predictions. However, special attention needs to be given to the latest publication by Nuzzo, where the author highlighted the fact of overuse of *p*-value as a tool to solve most of the problems, and suggests applying the methods as supportive techniques along with other multivariate approaches and therefore the same should apply to PLS-DA [86].

As an example of the use of PLS-DA, a very recent paper by Gromski *et al.* [87] compared modern feature selection and classification approaches for the analysis of mass spectrometry data. In this study PLS-DA was compared to a number of other approaches including LDA, SVM and RF in terms of variable selection and classification. This study emphasized the limited ability of PLS-DA compared to other approaches with regard to prediction accuracy, when used for data sets with a low number of variables, whilst prediction accuracy significantly improved when larger numbers of variables were included in the classification model. Elsewhere, Szymanska *et al.* [3] investigated how different validation approaches influence the outcomes of PLS-DA when applied to metabolomics data. Their findings showed that the commonly used parameter Q^2, which is used to assess the goodness of prediction [88] for optimization and performance of PLS-DA, had been outperformed by other approaches such as number of misclassifications (NMC) and the area under the receiver operating characteristic (AUROC) [85,89]. Therefore, if one wants significantly to reduce the large number of features and use the same model for classification, then PLS-DA is not the best approach. We do not wish to be over critical and of course there are plenty of examples that have successfully and judiciously applied PLS-DA in the analysis of metabolomics data. These include: an investigation of lung cancer metabolic signatures in urine [90]; the analysis of liver and serum metabolites of obese and lean mice fed on high fat or normal diets [91], and comprehensive metabolomic profiling and pathways of large biological data sets [92].

3.2. Principal components-discriminant function analysis (PC-DFA)

DFA, also known as canonical variates analysis (CVA) [27,93–95], is based on linear discriminant analysis (LDA) and fits a straight line into variable hyperspace effectively dividing that space into relevant regions [27]. This line corresponds to a vector which best discriminates between two or more classes. This vector is calculated based on a generalized distance matrix. In other words, for a given number of classes (groups) which correspond to dependent variable *Y*, DFA searches for a linear arrangement in the hyperspace constructed on the independent variable *X* that produces the largest mean difference between analysed classes, thereby maximizing between-class distance while minimizing within-class distance (directly related to the Fischer ratio). However, as this method is very sensitive to collinearity in the *X* matrix (as described in [96]) the method is very often preceded by principal component analysis (PCA) to overcome this limitation, hence PC-DFA (Fig. 4 shows a pictorial explanation of this algorithm). PCA is a commonly used technique for the dimensionality reduction of multivariate data whilst preserving most of the variance [97]; which is achieved without *a priori* knowledge of the groupings within samples in the original data set. Therefore, only small numbers of new uncorrelated latent variables (principal components) that best describe natural relationships between samples are applied to DFA. This allows for the reduction of noise in the data without reducing relevant information from the original data set.

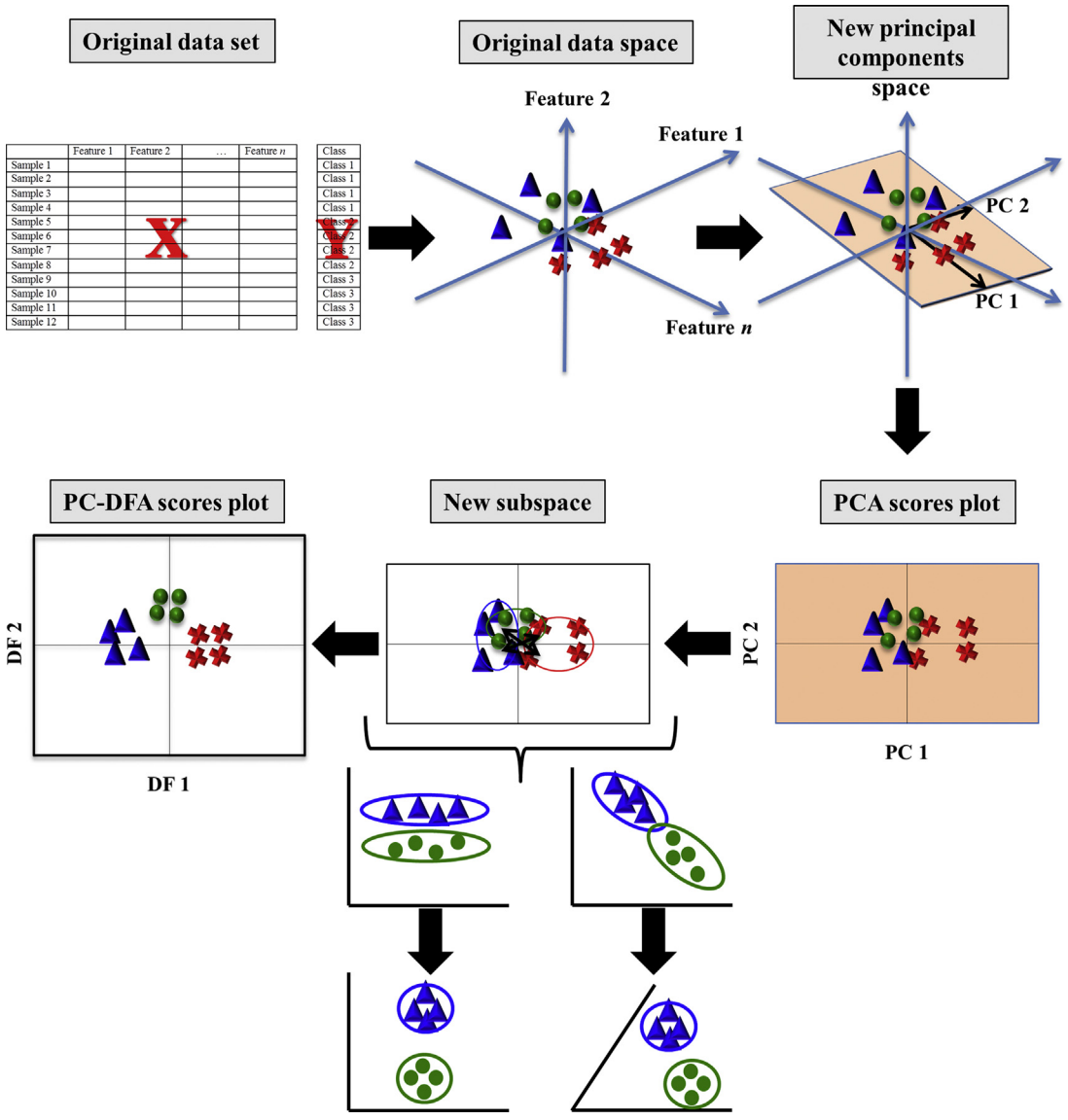


Fig. 4. PC-DFA outline. Initially the unsupervised approach of PCA is applied only to the *X* input data (independent variables) taken from the "original data set" (top left). The algorithm rotates the original data space (top middle) into a "new principal components space" (top right) such as the new axes; here PC1 and PC2 correspond to direction of highest variance as calculated from the original data (*X*). In this example, "PCA scores plot" (bottom right) represents two-dimensional plane that usually allow drawing qualitative conclusions about separation of the samples. In the next step the certain numbers of components that cover most of the variance from the original data set are used for DFA. Here, DFA allows separating different classes (*Y*) by searching a linear combination in a "new subspace" that maximize distance amongst two or more classes and instantaneously minimizing within-class distance (bottom middle). This allows obtaining a two-dimensional (bottom left) representation of the data that permits drawing valid conclusions.

To date, several advantages of this method have been reported extensively in the literature that describes variation of this approach and the interested reader is directed to the following Refs. [27,98–106]. The main advantages of this approach include simplicity, robustness, and statistics. Moreover, this technique identifies a linear separation between two or more groups without loss of information. Plotting of the first two PC-DFA scores recovered from all of the samples in many cases allows for visual interpretation, and other advantages such as χ^2 confidence intervals can be plotted around the different groups allowing the robustness of the clustering to be assessed [101,102]. Likewise, the method seeks to reduce dimensionality whilst optimizing class separation. Finally, PC-DFA loadings can be used to determine which variable discriminates between two or more classes, and of course PC-DFA can be used like all supervised learning methods to derive a classification model for predicting the group membership of new observations.

We note that under certain circumstances PC-DFA is similar to PLS-DA, as shown by Barker and Rayens [37]. However, as recently demonstrated by Brereton and Lloyd researchers encounter more hurdles with the latter [2]. Other caveats that need to be considered when using this method may be related to the fact that the method assumes unimodality of a Gaussian probability. Finally, this approach generates $C-1$ feature projections, where C corresponds to the number of classes therefore plotting of the scores plot for two class problems is rather abstract [27,77,99,100,102]; although this can be largely overcome by generating frequency histograms where the abscissa represents the distance in the first discriminant function and the ordinate contains information on the number of samples that appear in that area [107].

The PC-DFA technique was explored by Broadhurst and Kell where the authors investigated this approach in comparison to PLS-DA in terms of false discoveries in metabolomics and related experiments [14]. The authors emphasized the importance of

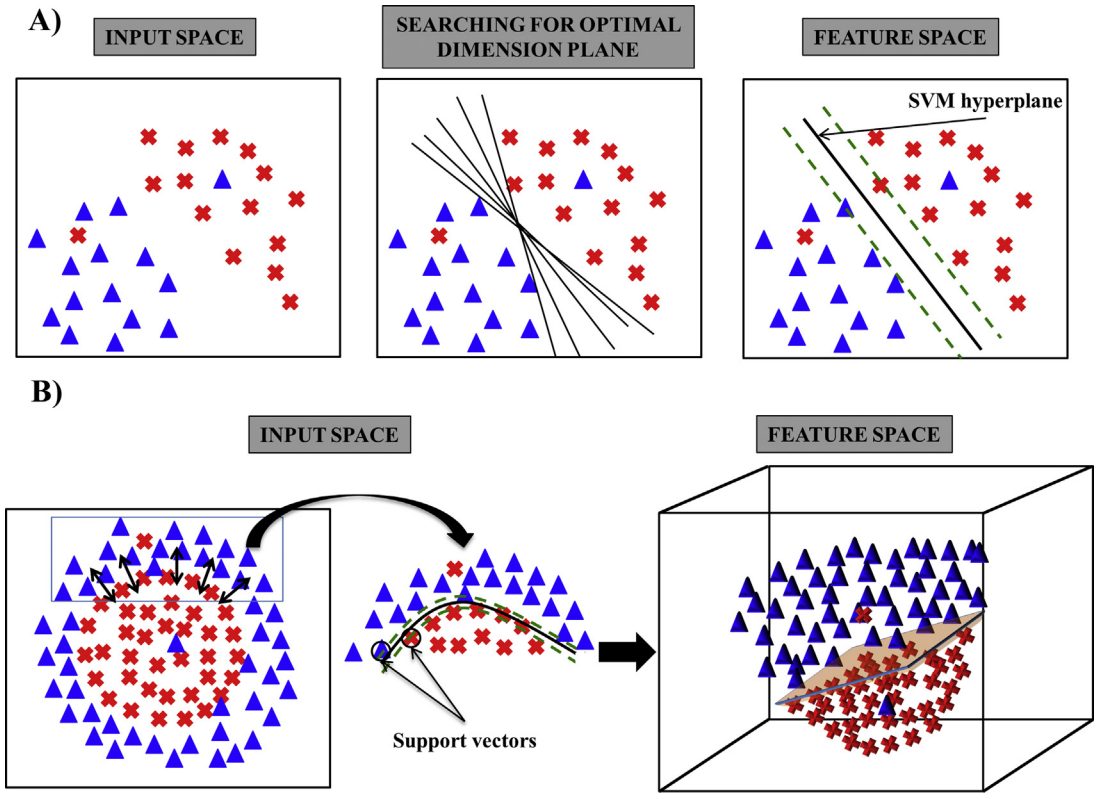


Fig. 5. SVM outline. (A) Application of SVM for linear cases, where the algorithm searches in the input space (left) for an optimal dimension plane (right) that will allow the maximization of the margins between two different groups. The closest points located on the both side of the margins here green line (right) are being defined as support vectors. Therefore, a new sample is classified based on which side of the margin it falls. (B) Application of SVM for non-linear cases (here "polynomial" kernel). In this case SVMs project a data space (left) that cannot be linear separable (middle) into a higher-dimensional feature space, where the optimal hyper-plane that separates two groups is identified. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

proper validation and the influence of parameter selection and, they recommended taking great care when optimizing both models. PC-DFA has been successfully implemented in several metabolomics studies which include: the analysis of a closely related groups of bacteria belonging to the genus *Bacillus* where PC-DFA was used to group these bacteria based on their SERS fingerprints [108]; in another study PC-DFA was used to demonstrate a linear trend with respect to recombinant protein production in cultures of Chinese hamster ovary (CHO) and murine myeloma non-secreting 0 (NS0) cell lines [109]; or more recently in the analysis of uropathogenic *E. coli* isolates [110].

3.3. Support vector machines (SVM)

SVM is an effective non-parametric machine learning technique suitable for both classification and regression problems [28]. This method is based on mapping data into a high-dimensional space that allows for the separation of two groups of samples into distinctive regions. This process is graphically explained in Fig. 5 where two simple examples of SVMs have been chosen: one for linear cases ("linear" kernel has been described) and one for non-linear problems ("polynomial" kernel has been explained). As shown in Fig. 5 separation is achieved by identifying only a small fraction of the samples, referred to as 'support vectors', between which the separating hyper-plane is identified. In order to maintain a good generalization performance when SVMs are applied to cases where perfect separation is not achievable (*i.e.*, to avoid over-fitting), a "slacking" variable is introduced so that a fraction of training samples are allowed to be misclassified before attempting to increase the complexity of the model. This is called soft margin in SVMs related literature [32,34,111]. The classification of the test set is determined by projecting each of the new points into this space and identifying to which side of the support vectors the new sample falls [32,34,111].

In comparison to PLS-DA, SVM is not influenced by the distribution of the different sample classes, on the contrary, it focuses on which side of the support vectors particular test samples fall [32]. The main advantage of this technique is its flexibility in the choice of the kernel function that allows the separation of two groups of samples, and this kernel can be chosen for either linear or non-linear problems. When one wants to deal with non-linear problems then the so-called 'kernel trick' is applied [34]. This allows for the transformation of the input space into a high-dimensional feature space where classes are linearly separable. The following kernels can be used: polynomial, radial basis function (RBF, also called Gaussian) or sigmoid functions [34,112]. When the 'kernel trick' is applied no assumptions about the functional form of the transformation, which makes data linearly separable, are necessary [34].

The main disadvantage of this technique is the lack of transparency of the results. There are no statistics such as scores and loadings available for easy visualization. The main caveat that needs to be considered when this technique is applied is that the algorithm was primarily developed to solve binary problems and so is ideal for case-control studies [28,29]. However, this problem may be overcome when three methods based on binary classifications are considered: 'one-against-all', 'one-against-one', and by application of directed acyclic graph (DAG) [111]. Finally, the identification of important variables is rather abstract, due to the lack of transparency. Yet again, this obstacle can be circumvented, either by the application of kernel-penalized SVM (KP-SVM), which allows the exclusion of the features that display low importance for

the classifier [113], or by selection of relevant features *via* application of SVM-based on recursive feature elimination (RFE) [114,115].

Numerous studies have compared SVM's performance against other commonly used models and SVM's robustness and excellent generalization performance have been demonstrated many times in various fields [35,116–119]. As an example in a study by Mahadevan *et al.* [120] two multivariate analyses, PLS-DA and SVM, were compared in terms of feature selection and classification. For this purpose data collected from healthy matched controls and patients with *Streptococcus pneumoniae* were used. In terms of feature selection variable importance on projection (VIP) for PLS-DA and RFE coupled to SVM were compared. Based on the reduced number of features the classification accuracies were measured, and it was shown that SVM outperformed PLS-DA, both in terms of feature selection as well as classification accuracy [120]. Other examples of successful application of SVM to metabolomics data include: application of SVM in understanding chemical or biological processes [121]; to classify lung cancer cases *versus* controls [122]; and in prediction of metabolic reactions based on atomic and molecular properties of small-molecules [123].

3.4. Random forests (RF)

RF belongs to the family of classification trees [30,33], whereby this approach generates many decision trees, each of which is constructed using different sets of randomly selected input (X) variables. This method as shown in the Fig. 6 begins with splitting the original data set using bootstrapping (with replacement). The sampling procedure consequently allows the construction of a series of training sets, which on average include 63.2% of all samples, and a series of test sets, which includes the remaining 36.8% samples [124,125]. The training data sets are used to construct the trees whereas the test sets are used to estimate classification accuracy [30,33].

The following advantages of random forests have been reported in the literature [33,126–136], including its ability to: deal with large datasets without variable deletion; provide a feature importance measure of the predictor variables (mean decrease in accuracy); a measure of the internal structure of the data (mean decrease in Gini index); as well as being able to handle missing values. Furthermore, this approach is robust to over-fitting and no scaling is required prior to the analysis. In summary, RF is a highly accurate classifier which produces an internal unbiased estimation of the generalization error, and finally the model is robust to outliers.

However, random forests can tend to overfit under some data distributions where maximal unpruned trees are used, consequently reducing performance of the model which is something that needs to be considered when applying this approach [137]. Moreover, the feature selections implemented in the model are not reliable in situations where variables differ in their scale of measurement or their number of categories [132–134].

Menze *et al.* compared the random forests classifier, with its associated Gini feature importance to PLS-DA and its associated variable importance on projection (VIP) coefficient [138]. Here the

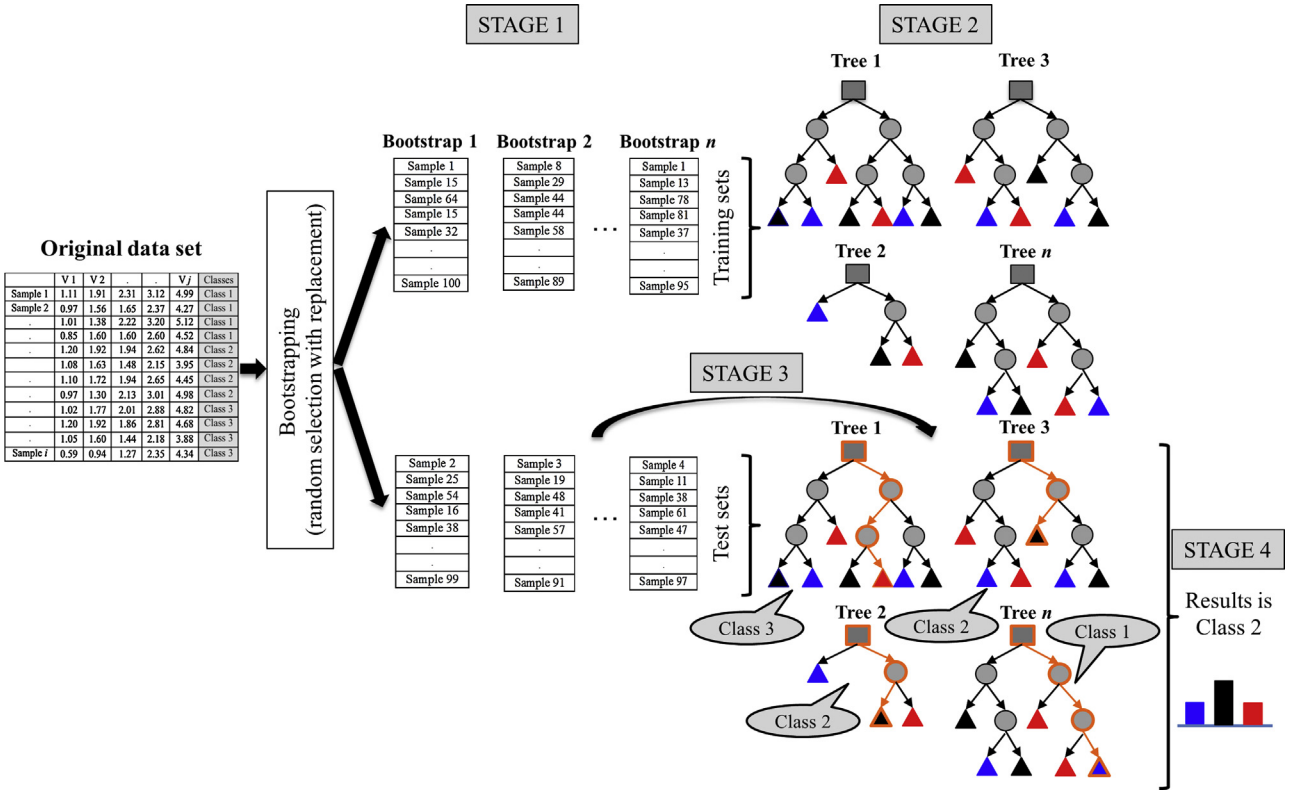


Fig. 6. Random forests outline. The process starts with the original data as the input matrix. In the first stage (Stage 1) the data are divided into training and test sets using bootstrapping (with replacement). This method allows for the generation of large collection of data sets encompassing different variations: on average 63.2% of all samples from the original data set are placed into the training sets and the remaining 36.8% used as the test set. In Stage 2 the training sets are used to build large collection of de-correlated decision trees that later are used in Stage 3 for classification. In Stage 2 each tree is grown using bootstrapped samples of training data set and starts with a root node (represented here as squares) where small subsets of input variables are selected randomly (usually according to the square root of the number of variables). This process allows optimizing the splits in the trees within the forest. Circles correspond to internal nodes where the samples are split based on the values of different variables. Finally, triangles correspond to leaf nodes where the classes are recognized. In Stage 3 each tree within the forest is challenged with the test sets: here each of the bootstrap partitions are run through each of the trees in the forest (as highlighted in orange) and the votes are counted. In Stage 4 when all the bootstrap samples from the test set are run through decision trees the numbers of votes are aggregated to get multi-class classification results. Finally, the algorithm generates a probability distribution of random forests. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of the four common chemometrics algorithms used in the analysis of metabolomics data. These comparisons include generally recognised advantages (pros) along with specific caveats for each of the methods.

	PLS-DA	PC-DFA	SVM	RF
PROS	• can be used to predict either continuous (PLS1) or categorical (PLS2) variables • availability of packages • ranks variables • reduces dimensionality • resistant to multi-collinearity and a much larger number of variables than observations • handles noisy data • ability to draw confidence intervals around different groups • robust classifier that can separate multiple groups	• simple and robust classifier • handles noisy data • identifies a linear separation between groups • ability to separate multiple groups • reduces dimensionality • ability to draw confidence intervals around different groups	• flexibility by having many potential kernel functions • not influenced by the distribution of samples • provides a single solution • targets both linear and non-linear data (dependent on kernel) • less likely to overfit and robust to noise • no local minima • robustness to outliers	• handles thousands of input variables without variable deletion • produce estimates of what variables are important in the classification • ability to handle missing values • robust to overfitting and outliers • handles data without pre-processing • handles noisy data
CAVEATS	• model validation is essential and often overlooked • scores plot may present an overoptimistic view of the separation between the classes • tendency to overfitting • limitations of R^2 and Q^2 in terms of description of the predictive ability of categorical models	• may overfit data if not properly validated • visualization of output in binary classification can be somewhat abstract as only single latent variable can be plotted • the relationship between variables are assumed to be linear in all groups • DFA without prior PCA is sensitive to multi-collinearity	• visualization for some kernels might be abstract (*i.e.,* polynomial, sigmoid or radial) • solves only two class problems (binary classifier); for multi-class problems, pair-wise estimation can be used (one against all) • computationally expensive with large number of classes • selection of parameters • variable selection is rather abstract	• complex visualization of output decision trees due to large number of trees • selection of parameters however in many cases default values are essential • higher correlation between any two trees in the forest increases the forest error rate, therefore, sensitive to the number of variables selected at each node

authors emphasize that RF may not always be adapted to spectral data due to 'the topology of its constituent classification trees' [138]; however, they appreciate the capabilities of RF in terms of feature selection. The authors showed that RF outperformed PLS-DA in terms of variable selection, whilst PLS-DA outperformed RF in terms of classification. As a result the authors recommended exploring the advantages of both approaches by connection of Gini index for feature selection and PLS-DA for classification. Further successful applications of RF classifier were demonstrated by Patterson *et al.* in the classification of type 2 diabetes mellitus (T2DM) *versus* controls [139] and by Fan *et al.* in the identification of serum protein biomarkers panels of prostate cancer [140].

3.5. Parameters selection

One of the most important considerations when using any of the methods described above is parameter selection: Table 2 summarizes the main advantages of all statistical methods presented in this review, along with important caveats that must be considered during calibration or when used for predictions. For PLS-DA and PC-DFA it will be the number of components (latent variables) to be used for modelling and these need to be judiciously chosen. However, this obstacle can be easily circumvented by employing a validation technique such as bootstrapping [124,125]. For SVM the parameters are not pre-defined and therefore careful parameter selection is required [141]. However, this can be easily optimized by the application of grid search, which identifies the correct values for each of the parameters based on misclassification error [34,112]. Finally, RF is governed by parameters, such as the number of samples to be selected at each node and the number of trees to be grown. Again these parameters can be selected sensibly by applying a grid search where different parameters are tested against each other [33,135,136].

In summary, after introducing PLS-DA we have presented three different approaches as alternatives to PLS-DA. This is of course not an exhaustive list and new multivariate approaches are continually being developed, as well as researchers using other established chemometrics approaches. The reader is directed to more detailed examples of these specific techniques which are provided in Table 3.

Table 3
Glossary of chemometric models applied in the analysis of metabolomics data.

Abbreviation	Term	Example of use
ANN	Artificial neural network [142]	Identification of 1,3-dicyclohexylurea in human serum [143]
BN	Bayesian network [144]	Metabolomic analysis of *Shewanella oneidensis* [145]
GA-BN	Genetic algorithm-Bayesian network [144,146]	Rapid identification of *Bacillus* spores and classification of *Bacillus* species [147]
GP	Genetic programming [148]	Pattern identification of metabolites that distinguish plasma from case and control [149]
KPLS	Kernel partial least squares [150]	Detection of the metabolite dipicolinic acid from *Bacillus* spores using surface enhanced Raman scattering [151]
MB-PCA	Multiblock principal component analysis [152]	Analysis of metabolomics data with two influential factors [152]
MB-PLS	Multiblock partial least squares [153]	Application to metabolic profiling of meat spoilage detection [154] or in the analysis of HIV protease inhibitors on expressing cervical carcinoma cells [155]
ML-PLSDA OPLS-DA	Multilevel partial least squares discriminant analysis [156] and orthogonal partial least squares discriminant analysis [157]	Metabolomics data from human nutritional intervention studies [39]
PARAFAC	Parallel factor analysis [158]	In the analysis of metabolites in regulatory mutants of yeast [159]
PLS-R	Partial least squares regression [1]	Metabolite profiling on serum and plasma sample [160]

Table 4
Comparison of four classifiers based on common characteristics.

Characteristic	PC-DFA	PLS-DA	RF	SVM
Handling of missing values	Poor	Poor	Good	Poor
Robustness to outliers	Fair	Fair	Good	Good
Predictive power	Good	Fair	Good	Fair/Good[a]
Ability to rank variables	Fair	Good	Good	Poor
Interpretability and visualization	Good	Fair	Poor	Poor
Resistance to overtraining	Fair	Fair	Good	Fair
Dimensionality reduction	Good	Good	Fair	Poor
Resistance to overfitting	Fair	Fair	Good	Good
Selection of parameters	Yes	Yes	Yes	Depends[a]
Data pre-processing	Yes	Yes	No	Yes

[a] Dependent on the applied kernel function; yes for linear classifiers, but no when non-linear kernels are used (*viz*., polynomial, RBFs or sigmoid functions).

4. Conclusions

Metabolomics is a relatively young, vigorous, and continuously expanding field, which has been said to offer some unique advantages in comparison to the other 'omics' [161]. Metabolomics is providing novel insights and exciting opportunities for a wide-range of disciplines such as the clinical/medical and environmental sciences, agri-food, systems and synthetic biology and industrial biotechnology amongst others. Therefore continual developments in this area require specialized chemometrics and (bio) informatics tools to meet the increasing challenges (and ever larger and more complex datasets) within the field. That being the case, despite the wide availability of chemometrics techniques many of them are not commonly used. Here, we have hopefully illustrated that there is no universal choice of method which is superior in all cases, but understanding how the method actually works is absolutely crucial. These have been summarized in Table 4 where the main characteristics of each algorithm are highlighted, many of which have been confirmed by our latest study [162]. We believe that PLS-DA is an excellent tool for the analysis of metabolomics data, despite its limitations, but only if well understood by those choosing to apply it. Though, it must be said, in many cases it can potentially lead to incorrect interpretations when used by inexperienced researchers.

In summary, we have shown that in many cases PLS-DA is outperformed in terms of variable selection and classification by the other approaches described above such as PC-DFA, SVM and RF. Of course, the selection of the classification models presented here are not exhaustive, but should shed at least some light on alternative approaches to PLS-DA. One point being that there are always alternatives available and the researcher need not feel obliged to use PLS-DA to analyze metabolomics datasets without further exploration, or at least awareness, of other options, several of which we have mentioned in this article. We would heartily encourage this exploration, both experimentally and through the literature, as we believe that further progress within metabolomics strongly depends on the understanding and application of a variety of statistical approaches, as one model is not always suitable for all data. Metabolomics may go hand in hand with PLS-DA at the moment, and this may be seen as a marriage of convenience and perhaps researchers should look outside this default algorithm and explore the 'chemometrics zoo' that exists for multivariate analysis [21].

Acknowledgments

The authors would like to thank to PhastID (grand agreement no: 258238) which is a European project supported within the Seventh Framework Programme for research and Technological Development and funding the studentship for PSG. In addition, the authors would like to thank the referees and editor for their useful comments and suggestions, which have helped us improve our manuscript.

References

[1] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. 58 (2001) 109–130.
[2] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213–225.
[3] E. Szymanska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8 (2012) S3–S16.
[4] C. Christin, H.C.J. Hoefsloot, A.K. Smilde, B. Hoekman, F. Suits, R. Bischoff, P. Horvatovich, A critical assessment of feature selection methods for biomarker discovery in clinical proteomics, Mol. Cell. Proteomics 12 (2013) 263–276.
[5] A.-L. Boulesteix, K. Strimmer, Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, Brief. Bioinf. 8 (2007) 32–44.
[6] K.M. Oksman-Caldentey, D. Inze, Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites, Trends Plant Sci. 9 (2004) 433–440.
[7] G. Blekherman, R. Laubenbacher, D.F. Cortes, P. Mendes, F.M. Torti, S. Akman, S.V. Torti, V. Shulaev, Bioinformatics tools for cancer metabolomics, Metabolomics 7 (2011) 329–343.
[8] J.L. Izquierdo-Garcia, I. Rodriguez, A. Kyriazis, P. Villa, P. Barreiro, M. Desco, J. Ruiz-Cabello, A novel R-package graphic user interface for the analysis of metabonomic profiles, BMC Bioinformatics 10 (2009) 363.
[9] K.-A. Le Cao, I. Gonzalez, S. Dejean, integrOmics: an R package to unravel relationships between two omics datasets, Bioinformatics 25 (2009) 2855–2856.
[10] T. Wang, K. Shao, Q. Chu, Y. Ren, Y. Mu, L. Qu, J. He, C. Jin, B. Xia, Autonics: an integrated platform for NMR-based metabonomics spectral processing and data analysis, BMC Bioinformatics 10 (2009) 83.
[11] E. Want, P. Masson, Processing and analysis of GC/LC–MS-based metabolomics data, Methods Mol. Biol. 708 (2011) 277–298.
[12] J. Xia, N. Psychogios, N. Young, D.S. Wishart, MetaboAnalyst: a web server for metabolomic data analysis and interpretation, Nucleic Acids Res. 37 (2009) W652–W660.
[13] G. Quintas, N. Portillo, J. Carlos Garcia-Canaveras, J. Vicente Castell, A. Ferrer, A. Lahoz, Chemometric approaches to improve PLSDA model outcome for predicting human non-alcoholic fatty liver disease using UPLC-MS as a metabolic profiling tool, Metabolomics 8 (2012) 86–98.
[14] D.I. Broadhurst, D.B. Kell, Statistical strategies for avoiding false discoveries in metabolomics and related experiments, Metabolomics 2 (2006) 171–196.
[15] O. Fiehn, D. Robertson, J. Griffin, M. van der Werf, B. Nikolau, N. Morrison, L.W. Sumner, R. Goodacre, N.W. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. van Ommen, J.C. Lindon, S.-A. Sansone, The metabolomics standards initiative (MSI), Metabolomics 3 (2007) 175–178.
[16] N.W. Hardy, C.F. Taylor, A roadmap for the establishment of standard data exchange structures for metabolomics, Metabolomics 3 (2007) 243–248.
[17] S.-A. Sansone, D. Schober, H.J. Atherton, O. Fiehn, H. Jenkins, P. Rocca-Serra, D. V. Rubtsov, I. Spasic, L. Soldatova, C. Taylor, A. Tseng, M.R. Viant, M. Ontology, Working Grp Metabolomics standards initiative: ontology working group work in progress, Metabolomics 3 (2007) 249–256.
[18] Bioinformatics Market Analysis And Segment Forecasts To 2020, Grand View Research, Inc., 2014. Available from: http://www.grandviewresearch.com/industry-analysis/bioinformatics-industry (27.04.2014).
[19] M. Sugimoto, M. Kawakami, M. Robert, T. Soga, M. Tomita, Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis, Curr. Bioinf. 7 (2012) 96–108.
[20] M. Brown, W.B. Dunn, D.I. Ellis, R. Goodacre, J. Handl, J.D. Knowles, S. O'Hagan, I. Spasic, D.B. Kell, A metabolome pipeline: from concept to data to knowledge, Metabolomics 1 (2005) 39–51.
[21] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Metabolomics by numbers: acquiring and understanding global metabolite data, Trends Biotechnol. 22 (2004) 245–252.

[22] M.M.W.B. Hendriks, F.A. van Eeuwijk, R.H. Jellema, J.A. Westerhuis, T.H. Reijmers, H.C.J. Hoefsloot, A.K. Smilde, Data-processing strategies for metabolomics studies, Trends Anal. Chem. 30 (2011) 1685–1698.
[23] K.H. Liland, Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis, Trends Anal. Chem. 30 (2011) 827–841.
[24] M. Eliasson, S. Rannar, J. Trygg, From data processing to multivariate validation – essential steps in extracting interpretable information from metabolomics data, Curr. Pharm. Biotechnol. 12 (2011) 996–1004.
[25] S.P. Putri, Y. Nakayama, F. Matsuda, T. Uchikata, S. Kobayashi, A. Matsubara, E. Fukusaki, Current metabolomics: Practical applications, J. Biosci. Bioeng. 115 (2013) 579–589.
[26] A. Smolinska, L. Blanchet, L.M.C. Buydens, S.S. Wijmenga, NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: a review, Anal. Chim. Acta 750 (2012) 82–97.
[27] B.F.J. Manly, Multivariate Statistical Methods: A Primer, Chapman and Hall, Boca Raton, 1986.
[28] V.N. Vapnik, Statistical Learning Theory, John Willey & Sons, New York, 1998.
[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[30] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[31] J.W. Allwood, D.I. Ellis, J.K. Heald, R. Goodacre, L.A.J. Mur, Metabolomic approaches reveal that phosphatidic and phosphatidyl glycerol phospholipids are major discriminatory non-polar metabolites in responses by *Brachypodium distachyon* to challenge by *Magnaporthe grisea*, Plant J. 46 (2006) 351–368.
[32] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167.
[33] D.R. Cutler, T.C. Edwards Jr., K.H. Beard, A. Cutler, K.T. Hess, Random forests for classification in ecology, Ecology 88 (2007) 2783–2792.
[34] Y. Xu, S. Zomer, R.G. Brereton, Support vector machines: a recent method for classification in chemometrics, Crit. Rev. Anal. Chem. 36 (2006) 177–188.
[35] R.M. Balabin, E.I. Lomakina, Support vector machine regression (SVR/LS-SVM)-an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data, Analyst 136 (2011) 1703–1712.
[36] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) .
[37] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173.
[38] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (2008) 81–89.
[39] J.A. Westerhuis, E.J.J. van Velzen, H.C.J. Hoefsloot, A.K. Smilde, Multivariate paired data analysis: multilevel PLSDA versus OPLSDA, Metabolomics 6 (2010) 119–128.
[40] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recogn. Lett. 31 (2010) 2225–2236.
[41] J.H. Moore, F.W. Asselbergs, S.M. Williams, Bioinformatics challenges for genome-wide association studies, Bioinformatics 26 (2010) 445–455.
[42] S.G. Oliver, M.K. Winson, D.B. Kell, F. Baganz, Systematic functional analysis of the yeast genome, Trends Biotechnol. 16 (1998) 373–378.
[43] O. Fiehn, Metabolomics – the link between genotypes and phenotypes, Plant Mol. Biol. 48 (2002) 155–171.
[44] D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D.D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G.E. Duggan, G.D. MacInnis, A.M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B.D. Sykes, H.J. Vogel, L. Querengesser, HMDB: the human metabolome database, Nucleic Acids Res. 35 (2007) D521–D526.
[45] D.B. Kell, Metabolomic biomarkers: search, discovery and validation, Expert Rev. Mol. Diagn. 7 (2007) 329–333.
[46] W.B. Dunn, D.I. Ellis, Metabolomics: current analytical platforms and methodologies, Trends Anal. Chem. 24 (2005) 285–294.
[47] W.B. Dunn, D.I. Broadhurst, H.J. Atherton, R. Goodacre, J.L. Griffin, Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy, Chem. Soc. Rev. 40 (2011) 387–426.
[48] V. Shulaev, Metabolomics technology and bioinformatics, Brief. Bioinform. 7 (2006) 128–139.
[49] A. Zhang, H. Sun, P. Wang, Y. Han, X. Wang, Modern analytical techniques in metabolomics analysis, Analyst 137 (2012) 293–300.
[50] J.L. Griffin, J.P. Shockcor, Metabolic profiles of cancer cells, Nat. Rev. Cancer 4 (2004) 551–561.
[51] J.K. Nicholson, I.D. Wilson, Understanding 'global' systems biology: metabonomics and the continuum of metabolism, Nat. Rev. Drug Discov. 2 (2003) 668–676.
[52] D.I. Ellis, V.L. Brewster, W.B. Dunn, J.W. Allwood, A.P. Golovanov, R. Goodacre, Fingerprinting food: current technologies for the detection of food adulteration and contamination, Chem. Soc. Rev. 41 (2012) 5706–5727.
[53] K.A. Hollywood, M. Maatje, I.T. Shadi, A. Henderson, D.A. McGrouther, R. Goodacre, A. Bayat, Phenotypic profiling of keloid scars using FT-IR microspectroscopy reveals a unique spectral signature, Arch. Dermatol. Res. 302 (2010) 705–715.
[54] A.J. Lloyd, J.W. Allwood, C.L. Winder, W.B. Dunn, J.K. Heald, S.M. Cristescu, A. Sivakumaran, F.J.M. Harren, J. Mulema, K. Denby, R. Goodacre, A.R. Smith, L.A. J. Mur, Metabolomic approaches reveal that cell wall modifications play a major role in ethylene-mediated resistance against Botrytis cinerea, Plant J. 67 (2011) 852–868.
[55] C.L. Winder, R. Cornmell, S. Schuler, R.M. Jarvis, G.M. Stephens, R. Goodacre, Metabolic fingerprinting as a tool to monitor whole-cell biotransformations, Anal. Bioanal. Chem. 399 (2011) 387–401.
[56] D.I. Ellis, R. Goodacre, Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy, Analyst 131 (2006) 875–885.
[57] W. Petrich, B. Dolenko, J. Fruh, M. Ganz, H. Greger, S. Jacob, F. Keller, A.E. Nikulin, M. Otto, O. Quarder, R.L. Somorjai, A. Staib, G. Warner, H. Wielinger, Disease pattern recognition in infrared spectra of human sera with diabetes mellitus as an example, Appl. Opt. 39 (2000) 3372–3379.
[58] A. Boskey, N.P. Camacho, FT-IR imaging of native and tissue-engineered bone and cartilage, Biomaterials 28 (2007) 2465–2478.
[59] P. Lasch, W. Haensch, D. Naumann, M. Diem, Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis, Biochim. Biophys. Acta-Mol. Basis Dis. 1688 (2004) 176–186.
[60] D.I. Ellis, D.P. Cowcher, L. Ashton, S. O'Hagan, R. Goodacre, Illuminating disease and enlightening biomedicine: Raman spectroscopy as a diagnostic tool, Analyst 138 (2013) 3871–3884.
[61] R. Salzer, H.W. Siesler, Infrared and Raman Spectroscopic Imaging, first ed., Wiley, Weinheim, 2009.
[62] J.W. Allwood, R. Goodacre, An introduction to liquid chromatography–mass spectrometry instrumentation applied in plant metabolomic analyses, Phytochem. Anal. 21 (2010) 33–47.
[63] D.I. Ellis, R. Goodacre, Metabolomics-assisted synthetic biology, Curr. Opin. Biotechnol. 23 (2012) 22–28.
[64] H.K. Kim, Y.H. Choi, R. Verpoorte, NMR-based plant metabolomics: where do we stand, where do we go? Trends Biotechnol. 29 (2011) 267–275.
[65] Z. Lei, D.V. Huhman, L.W. Sumner, Mass spectrometry strategies in metabolomics, J. Biol. Chem. 286 (2011) 25435–25442.
[66] N. Psychogios, D.D. Hau, J. Peng, A.C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T.L. Pedersen, S.R. Smith, F. Bamforth, R. Greiner, B. McManus, J.W. Newman, T. Goodfriend, D.S. Wishart, The human serum metabolome, PLoS One 6 (2011) .
[67] W.B. Dunn, W. Lin, D. Broadhurst, P. Begley, M. Brown, E. Zelena, A.A. Vaughan, A. Halsall, N. Harding, J.D. Knowles, T.S.A. Francis-McIntyre, D.I. Ellis, S. O'Hagan, G. Aarons, B. Benjamin, S. Chew-Graham, C. Moseley, P. Potter, C.L. Winder, C. Potts, P. Thornton, C. McWhirter, M. Zubair, M. Pan, A. Burns, J.K. Cruickshank, G.C. Jayson, N. Purandare, F.C.W. Wu, J.D. Finn, J.N. Haselden, A.W. Nicholls, I.D. Wilson, R. Goodacre, D.B. Kell, Molecular phenotyping of a UK population: defining the human serum metabolome, Metabolomics 11 (2014) 9–26.
[68] R. Goodacre, D. Broadhurst, A.K. Smilde, B.S. Kristal, J.D. Baker, R. Beger, C. Bessant, S. Connor, G. Calmani, A. Craig, T. Ebbels, D.B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjostrom, J. Trygg, F. Wulfert, Proposed minimum reporting standards for data analysis in metabolomics, Metabolomics 3 (2007) 231–241.
[69] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J. J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis, Metabolomics 3 (2007) 211–221.
[70] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genom. 7 (2006) .
[71] M. Brown, D.C. Wedge, R. Goodacre, D.B. Kell, P.N. Baker, L.C. Kenny, M.A. Mamas, L. Neyses, W.B. Dunn, Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets, Bioinformatics 27 (2011) 1108–1112.
[72] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat. Protoc. 6 (2011) 1060–1083.
[73] R.A. Scheltema, A. Jankevics, R.C. Jansen, M.A. Swertz, R. Breitling, PeakML/mzmatch: a file format, java library, R library, and tool-chain for mass spectrometry data analysis, Anal. Chem. 83 (2011) 2786–2793.
[74] J.P.A. Ioannidis, M.J. Khoury, Improving validation practices in omics research, Science 334 (2011) 1230–1232.
[75] X. Duportet, R.B.M. Aggio, S. Carneiro, S.G. Villas-Boas, The biological interpretation of metabolomic data can be misled by the extraction method used, Metabolomics 8 (2012) 410–421.
[76] P.S. Gromski, Y. Xu, H.L. Kotze, E. Correa, D.I. Ellis, E.G. Armitage, M.L. Turner, R. Goodacre, Influence of missing values substitutes on multivariate analysis of metabolomics data, Metabolites 4 (2014) 433–452.
[77] R.G. Brereton, Consequences of sample size, variable selection, and model validation and optimisation for predicting classification ability from analytical data, Trends Anal. Chem. 25 (2006) 1103–1111.
[78] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, L. Snipen, A partial least squares based algorithm for parsimonious variable selection, Algorithms Mol. Biol. 6 (2012) 27.

[79] T. Mehmood, K.H. Liland, L. Snipen, S. Saebo, A review of variable selection methods in partial least squares regression, Chemom. Intell. Lab. 118 (2012) 62–69.
[80] A. Krishnan, L.J. Williams, A.R. McIntosh, H. Abdi, Partial least squares (PLS) methods for neuroimaging: a tutorial and review, Neuroimage 56 (2011) 455–475.
[81] B. Worley, S. Halouska, R. Powers, Utilities for quantifying separation in PCA/PLS-DA scores plots, Anal. Biochem. 433 (2013) 102–104.
[82] K. Hasegawa, K. Funatsu, Evolution of PLS for modeling SAR and omics data, Mol. Inform. 31 (2012) 766–775.
[83] C.M. Rubingh, S. Bijlsma, E.P.P.A. Derks, I. Bobeldijk, E.R. Verheij, S. Kochhar, A. K. Smilde, Assessing the performance of statistical validation tools for megavariate metabolomics data, Metabolomics 2 (2006) 53–61.
[84] J.A. Westerhuis, E.J.J. van Velzen, H.C.J. Hoefsloot, A.K. Smilde, Discriminant Q (2) (DQ(2)) for improved discrimination in PLSDA models, Metabolomics 4 (2008) 293–296.
[85] A. Golbraikh, A. Tropsha, Beware of q(2)!, J. Mol. Graph. Model. 20 (2002) 269–276.
[86] R. Nuzzo, Statistical errors, Nature 506 (2014) 150–152.
[87] P.S. Gromski, Y. Xu, E. Correa, D.I. Ellis, M.L. Turner, R. Goodacre, A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data, Anal. Chim. Acta. 829 (2014) 1–8.
[88] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis: Principles and Applications, Umetrics Academy, Umeå, 2001.
[89] A.P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159.
[90] J. Carrola, C.M. Rocha, A.S. Barros, A.M. Gil, B.J. Goodfellow, I.M. Carreira, J. Bernardo, A. Gomes, V. Sousa, L. Carvalho, I.F. Duarte, Metabolic Signatures of Lung Cancer in Biofluids: NMR-Based Metabonomics of Urine, J. Proteome Res. 10 (2011) 221–230.
[91] H.-J. Kim, J.H. Kim, S. Noh, H.J. Hur, M.J. Sung, J.-T. Hwang, J.H. Park, H.J. Yang, M.-S. Kim, D.Y. Kwon, S.H. Yoon, Metabolomic analysis of livers and serum from high-fat diet induced obese mice, J. Proteome Res. 10 (2011) 722–731.
[92] X. Wang, B. Yang, H. Sun, A. Zhang, Pattern recognition approaches and computational systems tools for ultra performance liquid chromatography–mass-spectrometry-based comprehensive metabolomic profiling and pathways analysis of biological data sets, Anal. Chem. 84 (2012) 428–439.
[93] H.J.H. MacFie, C.S. Gutteridge, J.R. Norris, Use of canonical variates analysis in differentiation of bacteria by pyrolysis gas–liquid chromatography, J. Gen. Microbiol. 104 (1978) 67–74.
[94] W. Windig, J. Haverkamp, P.G. Kistemaker, Interpretation of sets of pyrolysis mass spectra by discriminant analysis and graphical rotation, Anal. Chem. 55 (1983) 81–88.
[95] R. Hoogerbrugge, S.J. Willig, P.G. Kistemaker, Discriminant analysis by double stage principal component analysis, Anal. Chem. 55 (1983) 1710–1712.
[96] R. Goodacre, E.M. Timmins, R. Burton, N. Kaderbhai, A.M. Woodward, D.B. Kell, P.J. Rooney, Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks, Microbiology 144 (1998) 1157–1170.
[97] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.
[98] L. Clemmensen, T. Hastie, D. Witten, B. Ersboll, Sparse discriminant analysis, Technometrics 53 (2011) 406–413.
[99] T. Hastie, A. Buja, R. Tibshirani, Penalized discriminant analysis, Ann. Stat. 23 (1995) 73–102.
[100] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, second ed., Springer, New York, 2009.
[101] W.J. Krzanowski, Orthogonal canonical variates for discrimination and classification, J. Chemom. 9 (1995) 509–520.
[102] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, M.R. Thomas, Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data, J. R. Stat. Soc. Ser. C Appl. Stat. 44 (1995) 101–115.
[103] L. Norgaard, R. Bro, F. Westad, S.B. Engelsen, A modification of canonical variates analysis to handle highly collinear multivariate data, J. Chemom. 20 (2006) 425–435.
[104] D.M. Witten, R. Tibshirani, Penalized classification using Fisher's linear discriminant, J. R. Stat. Soc. Series B-Statist. Methodol. 73 (2011) 753–772.
[105] D.M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (2009) 515–534.
[106] P. Xu, G.N. Brock, R.S. Parrish, Modified linear discriminant analysis approaches for classification of high-dimensional microarray data, Comput. Stat. Data Anal. 53 (2009) 1674–1687.
[107] P.S. Gromski, Y. Xu, K.A. Hollywood, M.L. Turner, R. Goodacre, The influence of scaling metabolomics data on model classification accuracy, Metabolomics (2014) , doi:http://dx.doi.org/10.1007/s11306-014-0738-7.
[108] R.M. Jarvis, A. Brooker, R. Goodacre, Surface-enhanced Raman scattering for the rapid discrimination of bacteria, Faraday Discuss. 132 (2006) 281–292.
[109] C.A. Sellick, R. Hansen, R.M. Jarvis, A.R. Maqsood, G.M. Stephens, A.J. Dickson, R. Goodacre, Rapid monitoring of recombinant antibody production by mammalian cell cultures using Fourier transform infrared spectroscopy and chemometrics, Biotechnol. Bioeng. 106 (2010) 432–442.
[110] H. AlRabiah, E. Correa, M. Upton, R. Goodacre, High-throughput phenotyping of uropathogenic *E. coli* isolates with Fourier transform infrared spectroscopy, Analyst 138 (2013) 1363–1369.
[111] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2002) 415–425.
[112] V. Cherkassky, Y.Q. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Netw. 17 (2004) 113–126.
[113] S. Maldonado, R. Weber, J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, Inf. Sci. 181 (2011) 115–128.
[114] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[115] K.B. Duan, J.C. Rajapakse, H.Y. Wang, F. Azuaje, Multiple SVM-RFE for gene selection in cancer classification with expression data, IEEE Trans. Nanobiosci. 4 (2005) 228–234.
[116] J.M. Kriegl, T. Arnhold, B. Beck, T. Fox, A support vector machine approach to classify human cytochrome P450 3A4 inhibitors, J. Comput.-Aided Mol. Des. 19 (2005) 189–201.
[117] F. Liu, Y. He, L. Wang, Determination of effective wavelengths for discrimination of fruit vinegars using near infrared spectroscopy and multivariate analysis, Anal. Chim. Acta 615 (2008) 10–17.
[118] M. Sattlecker, C. Bessant, J. Smith, N. Stone, Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics, Analyst 135 (2010) 895–901.
[119] U. Schmid, P. Roesch, M. Krause, M. Harz, J. Popp, K. Baumann, Gaussian mixture discriminant analysis for the single-cell differentiation of bacteria using micro-Raman spectroscopy, Chemom. Intell. Lab. Syst. 96 (2009) 159–171.
[120] S. Mahadevan, S.L. Shah, T.J. Marrie, C.M. Slupsky, Analysis of metabolomic data using support vector machines, Anal. Chem. 80 (2008) 7562–7570.
[121] P.W.T. Krooshof, B. Ustun, G.J. Postma, L.M.C. Buydens, Visualization and recovery of the (bio) chemical interesting variables in data analysis with support vector machine classification, Anal. Chem. 82 (2010) 7000–7007.
[122] P.G. Lokhov, O.N. Kharybin, A.I. Archakov, Diagnosis of lung cancer based on direct-infusion electrospray mass spectrometry of blood plasma metabolites, Int. J. Mass Spectrom. 309 (2012) 200–205.
[123] F. Mu, C.J. Unkefer, P.J. Unkefer, W.S. Hlavacek, Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds, Bioinformatics 27 (2011) 1537–1545.
[124] B. Efron, 1977 Rietz lecture. Bootstrap methods: another look at the jackknife, Ann. Stat. 7 (1979) 1–26.
[125] B. Efron, G. Gong, A leisurely look at the bootstrap the jackknife, and cross-validation, Am. Stat. 37 (1983) 36–48.
[126] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[127] G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers, J. Mach. Learn. Res. 9 (2008) 2015–2033.
[128] A.-L. Boulesteix, A. Bender, J.L. Bermejo, C. Strobl, Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations, Brief. Bioinform. 13 (2012) 292–304.
[129] A. Hapfelmeier, T. Hothorn, K. Ulm, C. Strobl, A new variable importance measure for random forests with missing data, Stat. Comput. 24 (2014) 21–34.
[130] A.M. Prasad, L.R. Iverson, A. Liaw, Newer classification and regression tree techniques: Bagging and random forests for ecological prediction, Ecosystems 9 (2006) 181–199.
[131] T. Shi, D. Seligson, A.S. Belldegrun, A. Palotie, S. Horvath, Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma, Modern Pathol. 18 (2005) 547–557.
[132] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC Bioinformatics 9 (2008) .
[133] C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinformatics 8 (2007) 25.
[134] C. Strobl, J. Malley, G. Tutz, An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests, Psychol. Methods 14 (2009) 323–348.
[135] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832–844.
[136] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (2002) 18–22.
[137] A. Statnikov, L. Wang, C.F. Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, BMC Bioinformatics 9 (2008) 319.
[138] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F.A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinformatics 10 (2009) 213.
[139] A.D. Patterson, J.A. Bonzo, F. Li, K.W. Krausz, G.S. Eichler, S. Aslam, X. Tigno, J. N. Weinstein, B.C. Hansen, J.R. Idle, F.J. Gonzalez, Metabolomics reveals attenuation of the SLC6A20 kidney transporter in nonhuman primate and mouse models of type 2 diabetes mellitus, J. Biol. Chem. 286 (2011) 19511–19522.
[140] Y. Fan, T.B. Murphy, J.C. Byrne, L. Brennan, J.M. Fitzpatrick, R.W.G. Watson, Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer, J. Proteome Res. 10 (2011) 1361–1373.

[141] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Mach. Learn. 46 (2002) 131–159.
[142] C.M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.
[143] L.M. Hall, L.H. Hall, T.M. Kertesz, D.W. Hill, T.R. Sharp, E.Z. Oblak, Y.W. Dong, D. S. Wishart, M.-H. Chen, D.F. Grant, Development of ecom(50) and retention index models for nontargeted metabolomics: identification of 1,3-dicyclohexylurea in human serum by HPLC/mass spectrometry, J. Chem. Inf. Model. 52 (2012) 1222–1237.
[144] N. Friedman, M. Linial, I. Nachman, D. Pe'er, Using Bayesian networks to analyze expression data, J. Comput. Biol. 7 (2000) 601–620.
[145] H. Wang, E. Correa, W.B. Dunn, C.L. Winder, R. Goodacre, J.R. Lloyd, Metabolomic analyses show that electron donor and acceptor ratios control anaerobic electron transfer pathways in Shewanella oneidensis, Metabolomics 9 (2013) 642–656.
[146] J.H. Holland, Genetic algorithms, SciAm 267 (1992) 66–72.
[147] E. Correa, R. Goodacre, A genetic algorithm-Bayesian network approach for the analysis of metabolomics and spectroscopic data: application to the rapid identification of Bacillus spores and classification of Bacillus species, BMC Bioinformatics 12 (2011) 33.
[148] J.R. Koza, Genetic Programming: on the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, 1992.
[149] L.C. Kenny, W.B. Dunn, D.I. Ellis, J. Myers, P.N. Baker, D.B. Kell, Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning, Metabolomics 1 (2005) 227–234.
[150] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for PLS, J. Chemom. 7 (1993) 45–59.
[151] D.P. Cowcher, Y. Xu, R. Goodacre, Portable, quantitative detection of bacillus bacterial spores using surface-enhanced raman scattering, Anal. Chem. 85 (2013) 3297–3302.
[152] Y. Xu, R. Goodacre, Multiblock principal component analysis: an efficient tool for analyzing metabolomics data which contain two influential factors, Metabolomics 8 (2012) S37–S51.
[153] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemom. 12 (1998) 301–321.
[154] Y. Xu, E. Correa, R. Goodacre, Integrating multiple analytical platforms and chemometrics for comprehensive metabolic profiling: application to meat spoilage detection, Anal. Bioanal. Chem. 405 (2013) 5063–5074.
[155] D.-H. Kim, J.W. Allwood, R.E. Moore, E. Marsden-Edwards, W.B. Dunn, Y. Xu, L. Hampson, I.N. Hampson, R. Goodacre, A metabolomics investigation into the effects of HIV protease inhibitors on HPV16 E6 expressing cervical carcinoma cells, Mol. Biosyst. 10 (2014) 398–411.
[156] E.J.J. van Velzen, J.A. Westerhuis, J.P.M. van Duynhoven, F.A. van Dorsten, H.C. J. Hoefsloot, D.M. Jacobs, S. Smit, R. Draijer, C.I. Kroner, A.K. Smilde, Multilevel data analysis of a crossover designed human nutritional intervention study, J. Proteome Res. 7 (2008) 4483–4491.
[157] M. Bylesjo, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, J. Chemom. 20 (2006) 341–351.
[158] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171.
[159] E.M. Humston, K.M. Dombek, B.P. Tu, E.T. Young, R.E. Synovec, Toward a global analysis of metabolites in regulatory mutants of yeast, Anal. Bioanal. Chem. 401 (2011) 2387–2402.
[160] A.A. Vaughan, W.B. Dunn, J.W. Allwood, D.C. Wedge, F.H. Blackhall, A.D. Whetton, C. Dive, R. Goodacre, Liquid chromatography-mass spectrometry calibration transfer and metabolomics data fusion, Anal. Chem. 84 (2012) 9848–9857.
[161] D.I. Ellis, W.B. Dunn, J.L. Griffin, J.W. Allwood, R. Goodacre, Metabolic fingerprinting as a diagnostic tool, Pharmacogenomics 8 (2007) 1243–1266.
[162] P.S. Gromski, E. Correa, A.A. Vaughan, D.C. Wedge, M.L. Turner, R. Goodacre, A comparison of different chemometrics approaches for the robust classification of electronic nose data, Anal. Bioanal. Chem. 406 (2014) 7581–7590.