Available online at www.sciencedirect.com

ScienceDirect

Chemometrics and Intelligent Laboratory Systems 91 (2008) 54 – 66

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Comparison of some chemometric tools for metabonomics biomarker identification

Réjane Rousseau [a,*], Bernadette Govaerts [a], Michel Verleysen [a,b], Bruno Boulanger [c]

[a] *Université Catholique de Louvain, Institut de Statistique, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium*
[b] *Université Catholique de Louvain, Machine Learning Group - DICE, Belgium*
[c] *Eli Lilly, European Early Phase Statistics, Belgium*

Received 29 December 2006; received in revised form 15 June 2007; accepted 22 June 2007
Available online 29 June 2007

Abstract

NMR-based metabonomics discovery approaches require statistical methods to extract, from large and complex spectral databases, biomarkers or biologically significant variables that best represent defined biological conditions. This paper explores the respective effectiveness of six multivariate methods: multiple hypotheses testing, supervised extensions of principal (PCA) and independent components analysis (ICA), discriminant partial least squares, linear logistic regression and classification trees. Each method has been adapted in order to provide a biomarker score for each zone of the spectrum. These scores aim at giving to the biologist indications on which metabolites of the analyzed biofluid are potentially affected by a stressor factor of interest (e.g. toxicity of a drug, presence of a given disease or therapeutic effect of a drug). The applications of the six methods to samples of 60 and 200 spectra issued from a semi-artificial database allowed to evaluate their respective properties. In particular, their sensitivities and false discovery rates (FDR) are illustrated through receiver operating characteristics curves (ROC) and the resulting identifications are used to show their specificities and relative advantages.The paper recommends to discard two methods for biomarker identification: the PCA showing a general low efficiency and the CART which is very sensitive to noise. The other 4 methods give promising results, each having its own specificities.

Keywords: Metabonomics; Multivariate statistics; Variable selection; Biomarker identification; ^{1}H NMR spectroscopy

1. Introduction

The recent biological ‘Omics’ domain is formed by several technological platforms (Genomics, Proteomics, Metabonomics) using multiparametric biochemical information derived from the different levels of biomolecular organization (respectively the genes, proteins and metabolites) to study the living organisms. All of these Omics sciences rely on analytical chemistry methods, resulting in complex multivariate datasets which require a large variety of statistical chemometric and bioinformatic tools for interpretation. Due to the central place of metabolites in organization of living systems, Metabonomics, the most recent technology in the world of “Omics”, is particularly indicated to extract biochemical information reflecting the actual biological events. Genomics and proteomics information describing transcriptional effects and protein synthesis do not provide a complete description of the perturbation cause by a disease or a xenobiotic on an organism. Alternatively, Metabonomics analyses the entire pool of endogenous metabolites in biofluids and creates a biological summary more complete and also closest to the phenotype. Consequently, the metabonomics approach is a promising framework to build detection tools of a response of an organism to a stressor.

Metabonomics is formally defined as “The quantitative measurement of the dynamic multi-parametric metabolic response of living systems to physiological stimuli or genetic modification” [1]. Metabonomics aims to approach the modifications of endogenous metabolites consecutive to a physiopathological stimulus or genetic modification by the combined use of an analytical technology and multivariate statistical methods. Proton nuclear magnetic resonance (^{1}H NMR) spectroscopy generates spectral profiles describing the concentration, size and structure of metabolites contained in collected biofluid samples. This analytical technology remains the more efficient in metabonomics as the analysis is non-destructive, non-selective, cost-effective and

[*] Corresponding author. Tel.: +32 10 47 30 49; fax: +32 10 47 30 32.
E-mail address: Rousseau@stat.ucl.ac.be (R. Rousseau).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.06.008

typically takes only a few minutes per sample requiring little or no sample pretreatment or reagent. Each resulting spectrum offers an overview of the metabolic state of the organism at the moment of the biofluid sampling. However, stressors from different categories affecting the organism will alter the concentration of the metabolites and consequently modify the spectral profile. On this basis, comparison of spectra in various specific states allows to detect alterations corresponding to biochemical changes inherent to the presence of a stressor. These resulting changes can be mapped by biologists to known pathways and to quickly build biochemical hypotheses. Anyway, the principal opportunity provided by the spectra of biofluids is the development of detection tools for the biological response to a stressor: viewing the recordered spectral changes as fingerprints of the reaction, the concerned regions of the spectrum can be employed on a new spectral profile to declare if this new observation develops the reaction.

However a typical ^{1}H NMR metabonomics study generates numerous biofluid samples and related complex ^{1}H NMR spectra, making impossible, even for a trained NMR-spectroscopist, to reveal all the changes by a visual inspection. Moreover, systematic differences between spectra are often hidden in biological noise. Adequate data pre-treatment and reduction tools and chemometric methodologies are then required to extract typical differences between spectra obtained in various states. In this aim, each spectrum domain is first transformed in a set of regions called descriptors corresponding each to the summed intensity below the spectrum in its region. The observed values of all spectra give rise to a multivariate ^{1}H NMR database, typically characterized by a large number of variables (the descriptors). Multivariate statistical methodologies are then applied to mine typical differences between spectral data in different conditions.

Although the application of metabonomics as detection tools was first developed in the pharmaceutical industry field for toxicity predictions and screening, recently applications have expanded the use of metabonomics to a large variety of domains of lifes sciences. Metabonomics is becoming a promising tools for clinical diagnosis, as well as environmental security applications [2].

Biomarkers and predictive models are the two different detection tools issued from the use of statistical methodologies on ^{1}H NMR metabonomics data. A ^{1}H NMR metabonomics biomarker is defined as a stable change in a ^{1}H NMR spectral region associated to the alteration of an endogenous metabolite in reaction to the contact with the considered stressor. Alteration of these regions or descriptor(s) serves in research as an indicator of the development of a response of the organism to the stressor. On the other hand, predictive model are useful in research to provide the probability of the development of this response. Predictive model quantitatively characterizes for each spectral region its pertinence to contribute to an adequate description of the presence of a response of the organism. Biologists then use this quantitative information in order to build a model aimed at validating the presence of reaction of the organism to a new stressor.

The metabonomics biomarkers, used in research by biologists, are developped or identified beforehand in an experimental ^{1}H NMR database with chemometric analysis. The goal of the methods is to find, in the range covered by the ^{1}H NMR spectrum, the area(s) which is (are) consistently altered by the given factor of interest (disease, toxicity). Several methods can be considered in order to identify in multivariate data the more altered variables in presence of a chosen characteristic [3]. Usually, unsupervised methods (Principal component analysis, Hierarchical cluster analysis, Nonlinear mapping) constitute a first step in metabonomics data analysis. Without assuming any previous knowledge of sample class, these methods enable the visualization of the data in a reduced dimensional space built on the dissimilarities between samples with respect to their biochemical composition. In this step, biomarkers are identified in a pertinent space of reduced dimension. For this purpose, principal component analysis (PCA) has been extensively used in metabonomics litterature. Despite apparent satisfying published results, the known large sensitivity of PCA to noise can suggest that improvements are expected with more robust methods to identify biomarkers in noisy data. Moreover, the traditional use of PCA remains highly questionable: biomarkers are identified from the loadings of the two first principal components, while the two first components do not necessarily contain the most relevant variations between altered and normal spectra. Sometimes, the results of the initial unsupervised analysis are confirmed by a second supervised analysis. This one employs classification methods as Partial Least Squares (PLS), SIMCA and neural networks, allowing first to separate normal and altered spectra, and secondly to identify more robust biomarkers [2].

This paper aims in investigating the relatives properties of advanced statistical methods for the identification of biomarkers from ^{1}H NMR data. As the performances of the PCA usual tool are questionable, the choice of an appropriate method stays a open domain. All methods covered in this paper are published tools selected for their frequent uses in statistics or chemometrics. Nevertheless, some of them are extended in order to fit with the biomarker identification goal. Some of the methods are used solely for identification, others may be used as predictive models too. All are compared based on both qualitative and quantitative considerations.

This paper is organized as follows. Section 2 presents six possible methods to identify biomarkers. The first one (MHT) is based on traditional descriptor-wise multiple hypothesis testing. The next two (s-PCA and s-ICA) are extensions of corresponding traditional multivariate data reduction tools. The final three methods (PLS, linear logistic regression and CART classification trees) provide predictive models from which biomarkers can be extracted. The next sections are devoted to the illustration and comparison of these methods. Section 3 presents a semi-artificial metabonomic data base built for this purpose. Section 4 illustrates the methods on one data set and emphasizes their particular characteristics. Finally, Section 5 tests systematically the six methods on several data samples and compare their performances in terms of various criteria as sensitivity, specificity and stability. This comparison will show that all methods, except s-PCA and CART, give promising results. Each of these has its own advantages and provides specific information.

2. Methods

In this section, six methods envisaged for biomarker identification and/or prediction are described, after the presentation of unified notations. All methods are described both in an intuitive way and an algorithmic form.

Let X be an $n \times m$ matrix of spectral data containing n spectra, each of them being described by m descriptors. A binary vector Y of size n identifies the class of each of the n spectra; n_0 are normal spectra ($y=0$) and n_1 ($y=1$) are altered ones. Section 3 will detail the difference between normal and altered spectra.

The goal here is to find among the m descriptors of the spectrum those which are associated to the concept of class membership, i.e. those which show systematic differences between the normal and altered classes. Each method provides so-called "biomarker scores" for all spectral descriptors in a vector b of size $m \times 1$. The descriptors with the highest scores will be considered as potential biomarkers. The number m_b of potential biomarkers of interest is chosen either by the biologist or recommended based on statistical criterion included in the method.

2.1. Multiple hypothesis testing (MHT)

2.1.1. Presentation

In the emerging-omics techniques, multiple hypothesis testing (MHT) has become a very popular technique to determine simultaneously if some descriptors, of a large set of possible ones, are altered by a (biological) factor of interest. Micro-array data analysis has been the privileged area of application of such methods. MHT consists in calculating, for each descriptor, a test statistic measuring the effect of the factor of interest; in our case, this factor is Y, the altered and normal spectral class identifier. The multiple test procedure aims then at giving a rule to decide which of the calculated statistics are statistically significant in controlling the total error rate of the test. When the number of tests is very high, the false discovery rate (FDR), i.e. the number of false discoveries over the number of discoveries, is usually taken as the error factor to be controlled. Several methods have been developed in a attempt to control the FDR under independent or dependent hypotheses [4–7]. In this paper, MHT is applied in two steps. First, for each descriptor, a classical t statistic is calculated to compare the mean spectral intensities of the two groups. The p-values attached to each statistic are then calculated and transformed to build biomarker scores. The Benjamini and Yekutieli [4] rule for multiple testing under dependency is then applied to set up a cut point between significant descriptors and non-significant ones.

2.1.2. Algorithm

- From the spectral matrix X, calculate, for each descriptor $j=1,\ldots, m$, the following t statistic:

$$t_j = \frac{\overline{x}_{1_j} - \overline{x}_{0_j}}{\sqrt{\frac{S^2_{1_j}}{n_1} + \frac{S^2_{0_j}}{n_0}}} \text{ where } \overline{x}_{kj} = \frac{1}{n_k}\sum_{\{i:Y_i=k\}} X_{ij} \text{ and } \tag{1}$$
$$S^2_{kj} = \frac{1}{n_k}\sum_{\{i:Y_i=k\}} (X_{ij} - \overline{x}_{kj})^2$$

- Calculate the individual p-values attached to these t statistics as

$$p_j = 2P(t_{vj} > |t_j|)$$

where t_{v_j} is a Student t random variable with v_j degrees of freedom. v_j is defined by the Welch formula as the closest integer to $\frac{\left(\frac{S_1^2}{n_1}+\frac{S_0^2}{n_0}\right)^2}{\left(\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_0^4}{n_0^2(n_0-1)}\right)}$

- Define the vector of signed biomarquer scores $b=(b_1, b_2,\ldots, b_m)$ from the following transformation of the p_j's:

$$b_j = \text{sign}(t_j)(\log(1/p_j))$$

- Choice as potential biomarkers the m_b descriptors with the highest b_j's. m_b can be chosen by the analyst or by the Benjamini–Yekutieli FDR controlling rule [4]: $m_b = \max\left\{i : p_i \leq \frac{i\alpha}{m\sum_{i=1}^{n}\frac{1}{i}}\right\}$ where α is the maximum expected FDR desired for the multiple testing procedure.

2.2. Supervised principal component analysis (s-PCA)

2.2.1. Presentation

In metabonomics and ^{1}H NMR data exploration, the Principal Component Analysis (PCA) [8] is the most commonly used method by practitioners. However, as discussed above, the traditional use of PCA in metabonomics remains questionnable, notably due to the use of the two first principal components. As PCA is here presented as a reference tool for biomarker identification, some improvements are suggested through a method called s-PCA.

s-PCA is performed as follows. A PCA is first applied to the matrix of centered by columns spectra X^c. The normalized score matrix is then used to find the two components which discriminate best between the two groups. This is an unusual, but effective way of using PCA. Indeed, PCA is an unsupervised method: the principal components are computed without taking the class information into account. The two first directions, which are usually selected, are therefore not necessarily those that maximize the discrimination. Here all directions are first computed, and only the two ones that discriminate the most the classes are kept. Then, in the plane defined by these two principal components, the direction that maximizes the discrimination is calculated and the corresponding loadings are chosen as biomarker scores.

2.2.2. Algorithm

- Center X by columns: $X^c = X - 1_n \cdot \overline{\mathrm{X}}^T$ where $\overline{\mathrm{X}}$ is the $m \times 1$ vector of column means and 1_n a $n \times 1$ vector of ones.
- Perform a PCA on X^c:$X^c = TP^T$ where T is the $n \times k$ matrix of scores and P is the $m \times k$ matrix of loadings ($k=\min(n, m)$).
- Normalise the score matrix as $C = T\Gamma^{-1/2}$ where Γ is the diagonal matrix with k eigenvalues.

- By applying formulae (1) to C (instead of X), calculate t statistics to compare, for each principal component, the normalized scores of both groups.
- Search the two components j_1 and j_2 that maximize $|t_j|$, i.e. which discriminate the best between the two groups.
- In the space of the j_1 and j_2 components, find the direction which maximizes the distance between the two groups centroïds and evaluate the contribution of loadings to this direction: $p^* = p_{j_1}\overline{c}_{1_{j_1}} - p_{j_0}\overline{c}_{0_{j_2}}$ where $\overline{c}_{1_{j_1}}$ is the coordinate on j_1 of the mean $\overline{c}_1$ of spectra scores from class 1 and $\overline{c}_{0_{j_1}}$ is defined in the same way.
- Define the biomarkers scores as $b=p^*$
- Choose a predefined number of descriptors with highest (absolute) scores as candidate biomarkers.

2.3. Supervised independent component analysis (s-ICA)

2.3.1. Presentation

Independent component analysis (ICA) [9] is methods that originally aimed in recovering unobserved signals or sources from linear mixtures of them. In the context of metabonomics ^{1}H NMR data, the media analyzed (e.g. plasma, urine) can be seen as a mixture of individual metabolites and NMR spectra may then be interpreted as weighted sums of NMR spectra of these single metabolites. If the matrix X of ^{1}H NMR spectra is rich enough, the application of ICA to ^{1}H NMR data should then ideally recover source products included in the analyzed media, in particular those that are biomarkers of the causal factor of interest in the study.

ICA is applied in this context as follows. First ICA is applied to the matrix of spectra. t statistics are then calculated from the mixing coefficients and used to identify sources that are able to discriminate the two groups of interest. Identified sources can ideally be interpreted as spectra of pure or complex metabolites whose quantities have been altered by the factor of interest. Mean NMR spectra for both altered and normal group are then reconstructed from the identified sources and the difference between these mean spectra are used as biomarker scores.

2.3.2. Algorithm

- Center X by lines and transpose it: $X^{Tc}=X^T-1_{\mathrm{m}}\cdot\tilde{\mathrm{X}}^T$ where $\tilde{\mathrm{X}}$ is the vector of lines (spectra) means.
- Apply ICA to X^{Tc}. e.g. the fastICA algorithm with parallel extraction of components proceeds in three steps:
  - Reduce by PCA the $m\times n$ matrix X^{Tc} to a $m\times k$ matrix of scores T ($k\leq\min(n, m)$): $X^{T\ c}=TP^T+E$ where E is the error.
  - Apply ICA to T: $T=WS$ where S is the $m\times k$ matrix of sources and W is the $k\times k$ unmixing matrix.
  - Derive the mixing matrix A such that $X^{T\ c}=SA$.
- Search for the sources that discriminate the most normal and altered spectra.
  - Calculate t statistics to compare, for each source, the mixing coefficients in both groups. The t statistics are derived by applying formulae (1) to A^T, the transposed mixing matrix.
  - Choose the k* sources with the highest t_j as those that discriminate the most the two groups. k* can be chosen either visually or with a FDR based method applied on the t_j's.
  - Build S* and A* the subset matrices of S and A corresponding to these k* sources.
- Calculate the biomarker scores as $b=S^*\ A^*\ Z$ where Z is a $(n\times 1)$ vector with $Z_i=-1/n_0$ if $Y_i=0$ and $1/n_1$ otherwise.
- Choose a predefined number of descriptors with highest (absolute) scores as candidate biomarkers.

2.4. Discriminant Partial Least Squares (PLS-DA)

2.4.1. Presentation

Partial least squares discriminant analysis (PLS-DA) [10,11] is a partial least squares regression aimed at predicting one (or several) binary responses(s) Y from a set X of descriptors. PLS-DA implements a compromise between the usual discriminant analysis and a discriminant analysis on the significant principal components of the descriptor variables. It is specifically suited to deal with problems where the number of predictors is large (compared to the number of observations) and collinear, two major challenges encountered with ^{1}H NMR data.

This paper suggests to apply PLS-DA for biomarker identification as follows. First, a PLS-DA prediction model is estimated [12]; the number of significant components is then chosen according to a cross-validation based criterion. The model provides regression parameters that can be used as biomarker scores b. The descriptors with the highest (absolute) coefficients are candidate biomarkers.

2.4.2. Algorithm

- Center X by columns : $X^c=X-1_n\cdot\overline{\mathrm{X}}^T$.
- Choose the optimal number of components of the PLS model using an adequate validation technique and criterion. The RMSEP (Root mean square error of prediction) [13] is a traditional criterion used in this context. It can be calculated for each size of model on an external validation set or, when no validation set is available, by k-fold cross-validation.
- Build the PLS regression prediction equation $\hat{Y}=X^cb$ using the previously chosen number of components.
- Define b as the vector of biomarker scores.
- Choose a predefined number of descriptors with highest (absolute) scores as candidate biomarkers.

2.5. Linear logistic regression (LLR)

2.5.1. Presentation

Linear logistic regression (LLR) [14] generalises classical multiple regression to binary responses. It aims at predicting the probability of class membership $\pi=P\ (Y=1)$ as a function of a set of exploratory variables $x=(x_1, x_2,\ldots, x_k)'$. In order to get a model response in the [0, 1] interval, the π is transformed with the logistic transformation: $\eta=\log\left(\frac{\pi}{1-\pi}\right)$ and η is expressed as a linear combination of x as $\eta=\alpha+\delta'x$. The parameters are estimated by maximum likelihood to take into account the Bernouilli distribution of the response Y.

Several points must be discussed when applying LLR to biomarker identification. As the number of potential regressors (descriptors) m is high and, in most cases, larger than the number of spectra ($n \ll m$), a dimension reduction or variable selection technique must be first applied to allow model estimation. Variable selection is privileged in this paper because the variables (descriptors) selected in the model can directly be seen as potential biomarkers. Forward selection (a technique that adds descriptors and never deletes them) and stepwise-forward (which starts with an empty set and adds or removes a single predictor variable at each step of the procedure) have been tested. Forward selection has demonstrated to be adequate in this context. The Akaike AIC criterion [15] is commonly used to select the variables to be entered into the model. The AIC criterion is defined as $\mathrm{AIC}=-2\log(L)+2(k+1)$ where L is the likelihood of the estimated model, and k is the number of variable included. When a model has been set up, biomarkers scores may be derived from the p-values of the regression coefficients.

2.5.2. Algorithm

- Estimate a model with the constant term only and calculate the corresponding AIC.
- Repeat for $k=1,\dots, m_b$:
  – Try to enter each descriptor x_j ($j=1,\dots, m$) as a supplementary variable in the model and calculate the corresponding AIC_j;
  – enter in the model variable x_j such that AIC_j is minimum.
- Stop either when a predefined maximum number mb of descriptor is reached or when the AIC criterion can not be decreased any more.
- Take as biomarker scores $b_j=0$ if descriptor j is not chosen in the model and $b_j=\mathrm{sign}(\delta_j)(1/p_j)$ for the other descriptors. p_j is the p-value of the Wald test on the regression parameter δ_j.

2.6. Classification and regression trees (CART)

2.6.1. Presentation

The CART tree classifier [16] implements a strategy where a complex problem is divided into simplest sub-problems, with the advantage that it becomes possible to follow the classification process through each node of the decision tree. In the context of this paper, CART is proposed to realise recursive and iterative binary segmentations of the descriptor space in order to direct spectra to smaller and smaller groups that are more and more homogeneous with regards to the class. When looking for biomarkers, the tree is not developed for its capacity to predict the class membership but for its stepwise selection of a subset of features relevant for class discrimination: the construction of the tree highlights in segmentation rules the descriptors with a good discriminant power between the two classes of spectra.

2.6.2. Algorithm

- Build the maximal tree model $T_{\max}$ by repeating segmentation until the number of spectra in each subgroup is less or equal than 5 as suggested by Breinman ([16] p82):
  – Define a binary segmentation rule by a descriptor x_j and its threshold $x_j^{(t)}$ chosen as to maximise the decrease of the Gini impurety criterion ([16] p38);
  – based on the value of x_{ij}, direct each spectrum i of the node to the left or the right child-node according to the chosen segmentation rule ($x_{ij} \leq \text{or} > x_j^{(t)}$).
- If a fixed m_b is required, take as biomarkers the descriptors x_j's corresponding to the m_b segmentation rules with the highest number of spectra in the branch under the corresponding node. The biomarker score b_j is this number when x_j is in this biomarker list and 0 otherwise. Note that the number of segmentation rules in the tree may be smaller than m_b.
- If one want to choose automatically the number of biomarkers m_b, the tree $T_{\max}$ may be reduced to a smaller (and optimal) tree $T_{\mathrm{op}}t$ by cost-complexity pruning [17]. This method cuts stepwise the branches of the initial tree which minimise the increase of error rate. In this sequence of nested trees, the optimal one is chosen with respect to its predictive accuracy (measured by a deviance) on an external data set or k-fold cross-validation ones.

2.7. Implementation

All algorithms have been implemented in the *R* language. Links to the used libraries are available on www.cran.r-project.org/src/contrib/Descriptions/. The following libraries were used: for PCA, the *pcurve* library, for ICA: the *fastICA* library, for PLS-DA: the *pls-pcr* library, for LLR: the *Design* library, for CART: the *tree* library. The MHT method has been implemented specifically for this study.

3. Description of the data

3.1. Typical metabonomics data

A typical experimental database is formed by three sets of data: a design, a set of ^{1}H NMR spectra and biological and/or hysto-pathological data. The design describes the experimental conditions underlying each available spectrum. Typical design factors are: subject (animal or human) ID and characteristics, treatment, dose, time of prelevement. The ^{1}H NMR dataset contains the spectral evaluations of biofluid samples collected according to the design. After spectra are accumulated, a primary data reduction ("binning") is carried out by digitizing the one-dimensional spectrum into a series of 250 to 3000 integrated regions or *descriptor* variables. However, a typical metabonomics study involves about 30 to 200 spectra or sample measurements. The resulting dataset is thus typically characterized by a larger number m of variables than the number n of observations. Another important characteristic of ^{1}H NMR data is the strong association (dependency) existing between some descriptors, due to the fact each molecule can have more than one spectral peak and hence contribute to a lot more than one descriptor. As a large variety of dynamic biological systems and processes are reflected in spectra, a range of physiological conditions, as for example the nutritional status, can also modify spectra. Noise or biological fluctuation are thus natural in the spectral data. Each spectrum of the ^{1}H NMR dataset is also usually linked to one or more variable(s) aimed at confirming by an independent measure the presence of a response of the organism towards the stressor. This confirmation is obtained by means of the current gold-standard examinations (biological measures or hysto-pathological ones) generated for the subject from which the spectrum is measured.

3.2. Construction of a semi-artificial database

To explore the capabilities of multivariate statistical methods to identify biomarkers, a semi-artificial database was built in which the descriptors to be identified by the methods are controlled. Knowing the biomarkers to be found offers the advantage to evaluate important characteristics of a method as the sensitivity and the specificity of a method. The principles of this construction lays on the addition of random artificial alterations to normal or *placebo* real rat urine spectra. By convention, this paper uses the terms “biomarkers” and “identifications” to make a distinction between respectively the “real alterations that we want to detect” and “the results or selection of descriptors indicated by the method as biomarker”.

3.2.1. Placebo data

The placebo data are composed of more than 800 spectra supposed to reflect a situation of physiological stability in rat urine. Each of them is issued from the COMET [18] database and corresponds to the spectral profile obtained from a “control” rat (which did not receive any treatment). All spectra were measured at a ^{1}H NMR frequency of 600 MHZ at the Imperial College London, using a flow injection process.

After acquisition, spectral FID signals were automatically treated and converted to variables using a Matlab library (BUBBLE) developed at Eli Lilly [19]. Bubble automatically performed, in sequence, suppression of the water resonance, an apodisation, a baseline correction, a warping to align shifted peaks. To decrease the inter-sample variability, a normalization is also realized, dividing each spectra by the median of a well chosen part of spectral descriptors. The last step reduces, by simple integration, the part of the spectrum between 0.2 and 10 ppm to 500 descriptors. Finally, some statistical tools (euclidian and Mahalanobis distances and PCA) were used to find outliers in the set of spectra and some of them were removed (less than 20). A typical urine spectra coming out of this process is given in Fig. 1a.

3.2.2. Simulation of alterations

Among the 500 descriptors, 46 were chosen to become biomarkers. These 46 descriptors were, for half of the spectra in the database, altered according to the description below, in order to simulate the response of an organism to a stressor or treatment. The 46 descriptors are chosen in ten consecutive regions of the spectra as shown in Fig. 1b. Half of these 46 descriptors and five regions are localized in a first part (index from 140 to 180) of the spectra contain a low level of noise; the other descriptors and regions are localized in a second part (index from 340 to 380) of the spectra, where the level of noise is higher. The alterations consist in adding to the placebo spectra random draws of Gamma distributions, whose means take the form of 10 peaks with different widths localized in the ten regions. Fig. 1b shows the mean height of these peaks. The signal of these alterations represents in average 30 percent of the noise of the spectra. Note also that four peaks have a width of 7 descriptors and the 6 other peaks have a smaller width of 3 descriptors, for a total of 46 descriptors or biomarkers. Moreover, some pairs (see arrows in Fig. 1b) of alterations were designed to be correlated, by using the same Gamma distribution to generate peaks, so that only six independent Gamma distributions have been used instead of 10. This last feature of the database makes it possible to test if the biomarker identification methods are able or not to discover correlated biomarkers. The 6 sets of 3 to 14 single descriptors or biomarkers generated independently will be called below the “independent” biomarkers. Note finally that each placebo spectrum was only altered with two (randomly chosen) from the six possible independent biomarkers. This simulates the fact that each organism doesn’t necessarily respond the same way to a stimulus. A dataset of 400 altered spectra (randomly chosen from the 800 placebos) was built with this methodology.

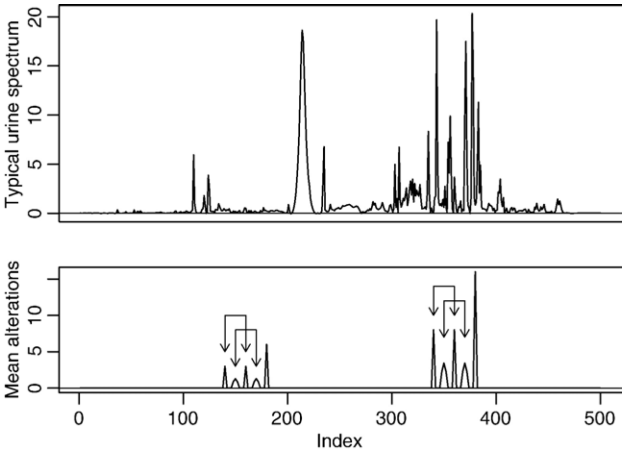

Fig. 1. (a) Typical rat urine NMR spectrum, (b) positions and mean amplitude of alterations added to urine spectra.

3.3. Construction of the datasets for method testing and comparison

The dataset available consists therefore in 400 placebo spectra and 400 “altered” spectra. The goal of the next sections is to show that the methods are able to identify the altered descriptors. For this purpose, subsamples of size 60 (2 × 30) and 200 (2 × 100) have been extracted out this dataset to simulate typical metabonomics sample sizes. In Section 4, one sample of 200 spectra was drawn randomly to illustrate the methods. In

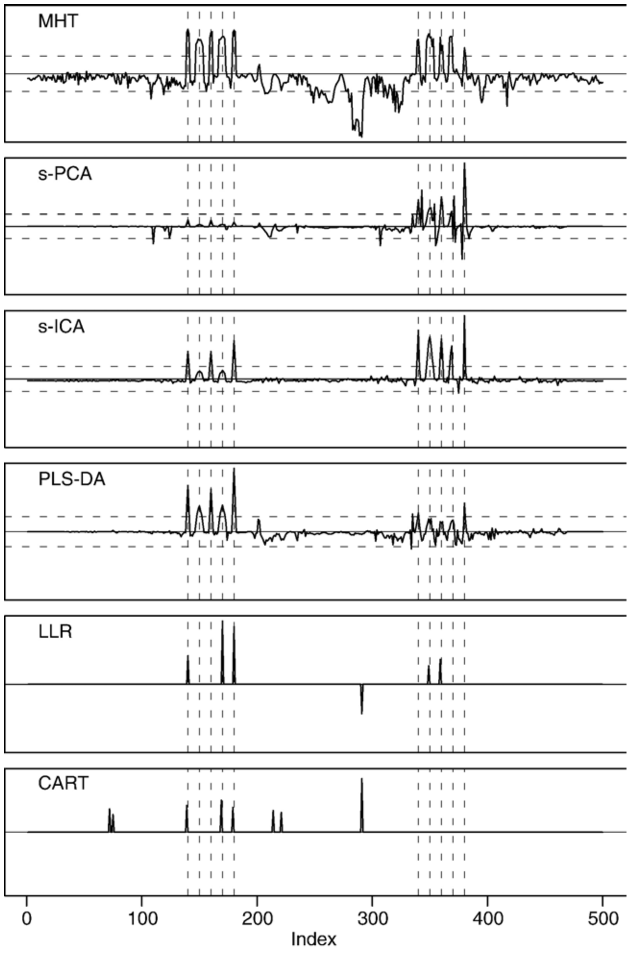

Fig. 2. Biomarker scores for all tested methods.

Section 5, twenty samples for each size were drawn to compare the performances of the methods. Note that validation sets were not used when needed in the methods (PSL and CART); k-fold cross validation was preferred instead.

4. Illustration of the methods

The purpose of this section is to illustrate the six methods on a subsample of 200 (100 normal and 100 altered) spectra extracted from the semi-artificial database described in Section 3. The main results available from the six methods are reported graphically in Figs. 2–5. Fig. 2 provides the biomarker scores calculated for each one. Figs. 3, 4 and 5 present intermediate outputs for s-PCA, s-ICA and CART that offer additional support to visualize the identified descriptors and/or some specific features of these procedures. All these graphics serve below to describe qualitative behaviour resulting from the design of the methods while systematic performance method evaluation is kept for Section 5. Let's first interpret the score plot figure.

- Four methods (MHT, s-PCA, s-ICA and PLS-DA) provide non null scores for all descriptors leading to possible complex score profiles. Score profiles for LLR and CART methods are simpler since they come from variable selection procedures and non null score values only exist when a descriptor is selected.
- The vertical lines on the graphic show the locations of the 46 "real" biomarkers. High positive scores in these regions denote that the method was able to identify them. High (positive or negative) scores elsewhere must be interpreted as false discoveries.
- The sign of each score has also a meaning: a positive score indicates that the corresponding descriptor has potentially a positive effect on the intensity of the altered spectra. A negative score leads to a decrease of intensity from normal to altered spectra.
- Three methods (MHT, LLR, CART) provide criteria to select automatically "significant" scores. These have been applied here. The threshold of selection, drawn by an horizontal line in the MHT graph, is calculated from the Benjamini Yekutieli (B–Y) FDR rule with $\alpha = 0.05$. The descriptors in LLR are those selected by minimizing AIC and in CART by applying the pruning algorithm. This example shows that LLR and CART seem to be very selective methods, but in the contrary, the B–Y does not seem to really control FDR since many false discoveries appear (much more than 5%). The horizontal lines drawn for the three other methods (s-PCA, s-ICA and PLS-DA) have no statistical interpretation and are there only for visual purpose. They have been drawn such that the number of scores appearing out of the interval is 23, the half of the 46 "real" biomarkers.
- The MHT method identifies well all biomarkers but is also the method that generates the most noisy score vector leading to many false identifications. It may be shown that this behavior increases when the sample size decreases.
- s-PCA, s-ICA and PLS-DA methods have quite similar score profiles: they are all able to identify some (s-PCA and PLS-DA) or all (s-ICA) biomarkers and display reasonable noise in the other regions. s-ICA is specially able to extract signal from noise. PLS-DA seems, as it is a predictive tool,

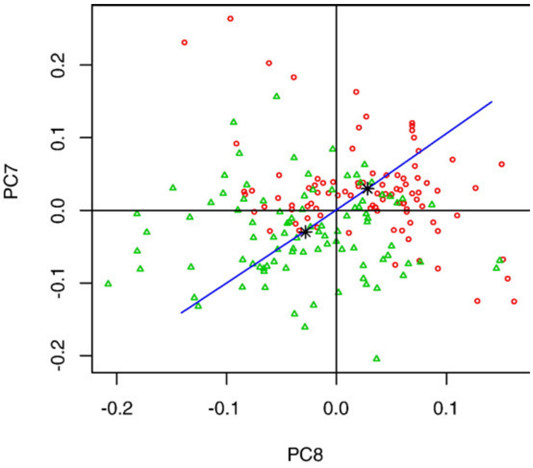

Fig. 3. Projection of the spectra on the principal components which best discriminate between normal and altered spectra. A △ symbol represents a projection of a normal spectrum and a ◦ of a altered one. The two * indicate the centroids of the clouds respectively formed by each kind of spectra.

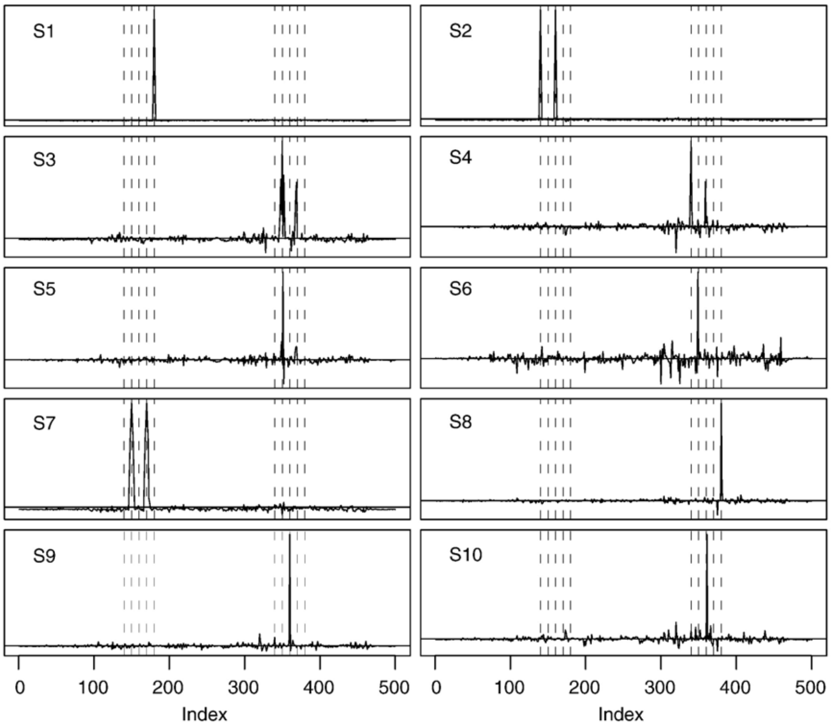

Fig. 4. The 10 ICA sources which best discriminate between normal and altered spectra.

to privilege biomarkers from the less noisy spectral region. More curiously, s-PCA performs better in the noisy biomarker region. This may be due to the fact that the t statistic privileges high signal even in a noisy area of the spectrum.

- For LLR and CART methods, one can observe that true discoveries are all coming from independent biomarkers. When one descriptor from an independent biomarker is selected by the procedure, all others are discarded because they constitute redundant (and correlated) information. This behavior is typical in a forward regression selection technique and in decision trees. Fig. 2 shows that LLR identifies 5 (of the 6) independent biomarkers with a little sensitivity to noise while CART identifies only the three in the first part of the spectra. The CART method presents indeed a high sensitivity to noise, as illustrated by the lack of identification of biomarkers in the second (noisy) part of the spectra. Many

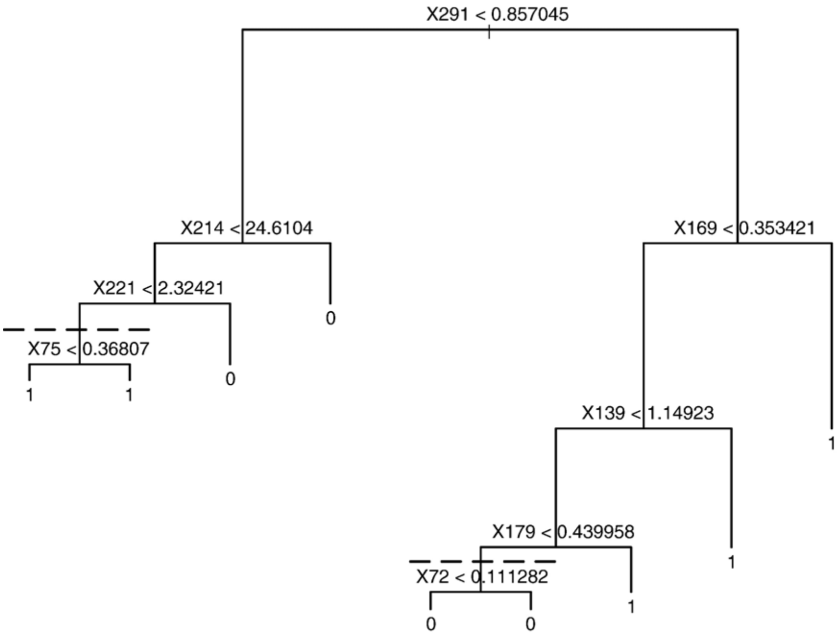

Fig. 5. Classification tree before and after pruning. Horizontal bars indicate where the tree is pruned.

other simulations have confirmed that basic CART can be efficient in situations without or with low noise but is not able to find signal in presence of higher noise.

The following comments can be made from Figs. 3, 4, and 5:

- Fig. 3 presents, for s-PCA, the projection of the 200 spectra in the space of the two principal components which discriminate best normal and altered spectra. This graphic is certainly helpful for most biologists very used to PCA methods. It shows how well spectra are separated and can detect outliers. Note that, in this example, the two best components are the 8th and the 7th. Biologists used to work with first components should then figure out that high variance explained does not mean high discrimination as the classification factor is not taken into account in a PCA. More precisely, the space of the 7th and 8th components explains an amount of variance smaller than the space formed by the PC1 and PC2 (7.9% with respect to 14,8%). However, components 7 and 8 contain the part of the variance that is informative for the identification of biomarkers, as illustrated by the distinction between the two kinds of spectra in Fig. 3.
- Among the 100 estimated sources in s-ICA, 10 independent sources were selected (by a FDR based procedure applied to the *t* statistics) to be significantly discriminant. Each of these independent sources is illustrated in Fig. 4. The graphic is impressive: sources S1, S2, S3, S4, S7 and S8 correspond nearly perfectly to the 6 independent biomarkers added to

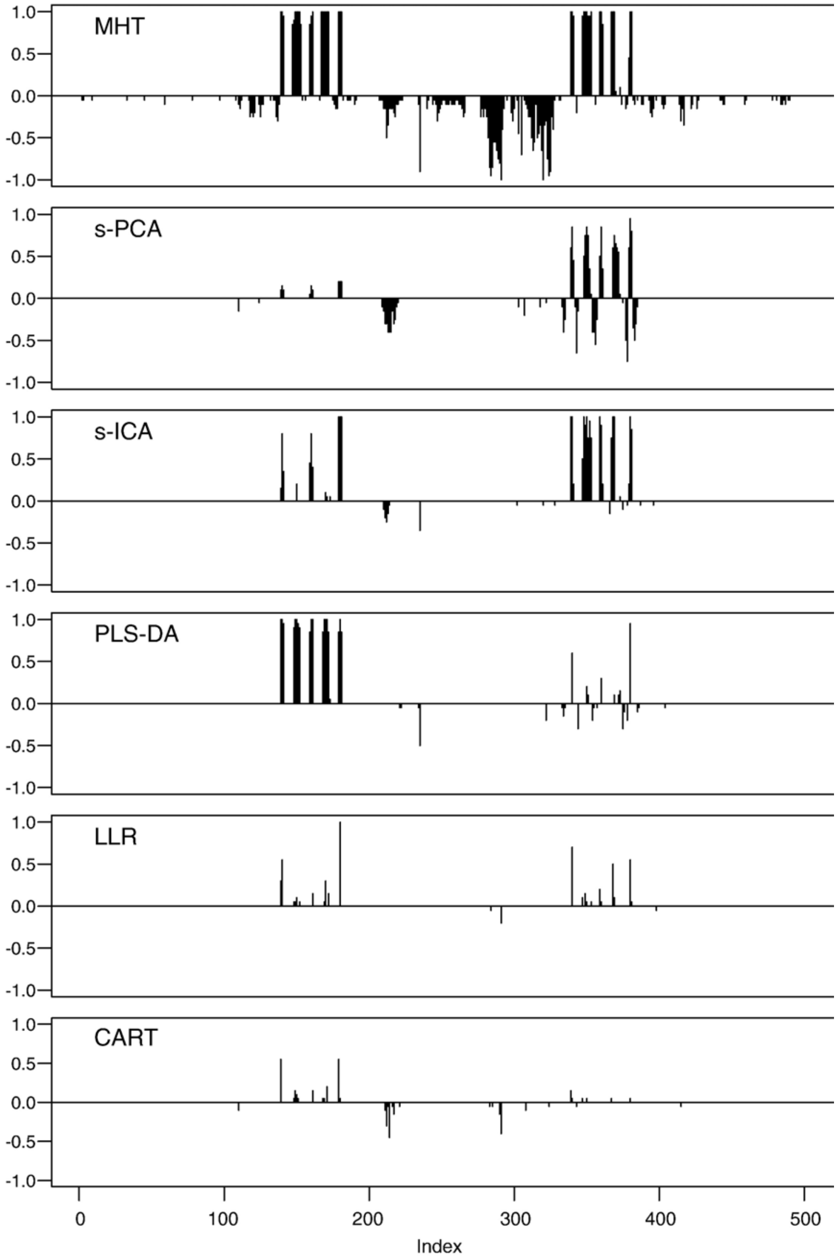

Fig. 6. Proportions of occurrences of true (positive bars) and false (negative bars) biomarker identification in simulations.

highly variable urine spectra in the artificial database. ICA is therefore able to extract these independent multi-descriptor biomarkers without prior information on their number and characteristics. The "purity" of the sources (especially S1, S2 and S7) must also be highlight: the signal can really be extracted from the noise. Sources S5, S6, S9 and S10 are unfortunately less useful: these represent correct biomarkers but are only parts of the multi-peak independent biomarkers. They are then redundant and it is difficult to explain why they appear as independent sources. Note that the authors realized similar graphics with the loadings of the principal components calculated from s-PCA or PLS-DA. They are not shown here because they do not reveal any useful information about independent biomarkers and are much more noisy.

- Cart tree representation shown in Fig. 5 presents the sequence of descriptors issued from the recursive segmentation, providing supplementary information on the order of descriptor selection and the exact segmentation rules. The horizontal line shows where the tree was cut by the pruning algorithm.

5. Method comparison

The six methods described in this paper have been illustrated using a single dataset in Section 4. The present section compares their performances on several datasets of different sample sizes. For this purpose, 20 samples of 200 spectra (100 altered and 100 placebo ones) and of 60 spectra (30 altered and 30 placebo ones) were drawn at random from the semi-artificial database described in Section 3. The 6 methods were then applied to each of these 40 datasets. Different samples sizes are used to test the robustness of the methods to small (but realistic in real-life situations) samples. In addition, generating 20 samples eliminates possible effects of a particular draw while the variability of the results can also be studied too.

5.1. Number of identifications

The first results concern the identifications obtained from each method. Fig. 6 provides for the 200 spectra and for each method, the proportion of simulations where each descriptor has been identified as a biomarker. The positive bars represent the correct identifications; the negative bars indicate false discoveries. Results for the 60 spectra databases are not given because they are very similar and only accentuate the observations coming out of Fig. 6.

For three methods (s-PCA, s-ICA and PLS-DA), the number of descriptors mb that the method identifies as biomarkers has to be fixed by the analyst. As a method can be supposed to have a limited number of correct detections, the number mb for these methods has been fixed here at 23, the half of the total 46 biomarkers randomly added to the altered spectra. For the three other algorithms (MHT, LLR and CART), the number of identifications mb is chosen automatically by a statistical criteria as detailed in Section 2.

These results will be interpreted together with the ROC curves after the next subsection.

5.2. ROC curves

As the performances of a method strongly depend on the total number of identifications m_b (both false identifications and biomarkers correctly identified), it is sometimes difficult to compare several methods which do not deliver the same number of identifications. The receiver operating characteristic curve (ROC [20]) provides a way to visualize the performances of a method for a whole range of possible m_bs. It must be noticed that in the presented ROC curves the performances evolve according an experimental condition (the value of m_b) and not according to a parameter of the method as in the traditional ROC curves. Consequently, the ROC curves shown in this paper can be non-monotonic. More precisely, it gives for each number m_b of identifications, in a chosen range, the method sensitivity and FDR (false discovery rate). The sensitivity is defined as the proportion of biomarkers correctly identified (among all biomarkers); the FDR is the percentage of false identifications (among all the m_b identifications).

As explained in Section 3, there are only 6 independent (multiple) biomarkers among the 46 ones. The sensitivity may thus be defined with respect to the proportion of correct identifications among the 46 biomarkers, or with respect to the proportion of correct identifications found among the 6 independent ones. These two definitions of sensitivity (therefore of ROC curves) give the four diagrams of Fig. 7, two for the 200 spectra case and two for the 60 spectra one. One curve represents the mean of the 20 ROC curves obtained from the application of one method to the 20 datasets. Each curve presents a method performance in a range of 1 to 46 number of identifications.

Good methods are those whose curves are mostly concentrated in or at least reach the upper left part of the ROC diagrams.

5.3. Discussion

The following comments can be made from Figs. 6 and 7.

The MHT method, with its FDR based threshold criterion, selects the higher number of descriptors as potential biomarkers (75 in average for the 200 spectra case). It is thus natural that this method has a high sensitivity, but this is at the price of a high FDR. This confirms the poor performance of the B–Y decision rule in this context. The MHT t scores are able to discover biomarkers in both low and high noise region of the spectra but loose clearly its performance in small sample. The MHT method has also the tendency to make false discoveries in the nosier part of the spectra. The main advantage of MHT is then certainly its simplicity coupled with overall acceptable performances especially in large samples. However, as other methods don't require it, the normality of the present H NMR data haven't be taken into account. A possible lack of normality can then have consequence on the MHT score and its performance.

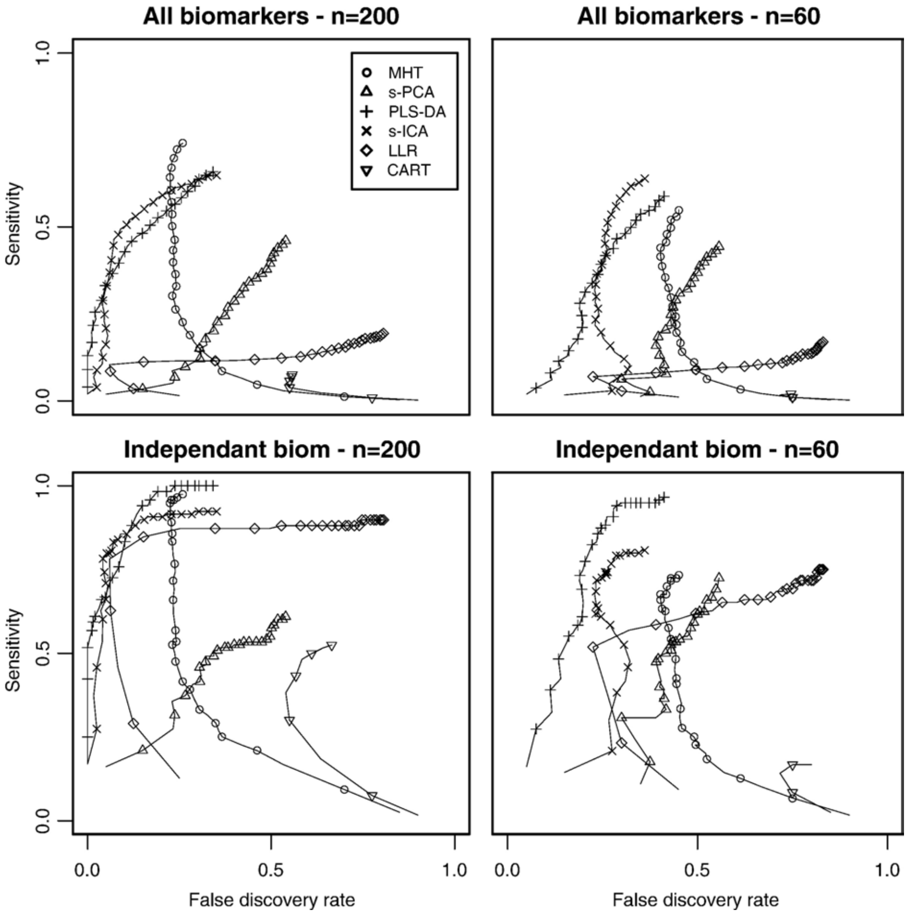

Fig. 7. Mean ROC curves (sensitivity versus false discovery rate) for the 6 methods. For clarity, curves only show symbols representing even number of identifications.

The s-PCA method, traditionally used in this context, is performing very poorly. As explained in Section 4, s-PCA only provides correct identifications in the presence of big alterations in an ideal noisy spectra. This is why in the presented more likely natural case with biggest biomarkers in a noisy part of the spectra, s-PCA is the second worst method (after CART): even if the number of biomarkers was chosen adequately (which would require a well-defined criterion), an increase of the sensitivity would be accompanied by a high FDR.

The ICA method is more natural than methods based on PCA: indeed the independence statistical criterion corresponds to the notion of independent biomarkers, contrarily to the decorrelation as in PCA-based algorithms. This is certainly the main advantage of ICA; the consequence is that the independent biomarkers can be retrieved and plotted in the form of a spectrum (see Section 4) and the metabolites playing a role as biomarker can thus potentially be identified.

Besides this interpretation power, the ICA method also gives good biomarker identification performances. As it can be seen in Fig. 6, biomarkers are correctly identified even in the noisy part of the spectra. ROC curves go also in the upper left corners of the diagrams, showing that sensitivity can become high when increasing the number of identifications without deteriorating too much the FDR. In the case of the search for independent biomarkers, the method is even more efficient. In terms of the mean number of biomarkers found, only the PLS-DA method can compete with the ICA one.

From Fig. 6, it is however visible that the good performances of PLS-DA mostly come from the less-noisy regions of the spectra, while the ICA method is more robust in the strongly noisy regions. At comparable performances in terms of ROC curves, it can be concluded that ICA is more robust to noise than PLS-DA. The PLS-DA method is however very efficient in recovering independent biomarkers (it is the only method that finds always all of them in the 20 experiments with 200 spectra).

The LLR method is not adequate to find all biomarkers. Indeed, as a purely prediction tool, it stops selecting potential biomarkers once the predictive performances are acceptable. This means that once a biomarker is found, all other ones that are dependent to the first one will not be identified. Indeed, the method gives much better performances when looking for independent biomarkers only. It is the method which reaches the highest sensitivity with the smaller mean number of identifications (0.8 with 5 identifications). However, from Fig. 6, it appears that in the 20 runs, different biomarkers are selected among each set of dependent ones. Building different models (from slightly different samples) can thus lead the biologist to find several biomarkers influenced by a single metabolite, which can be interesting in some cases.

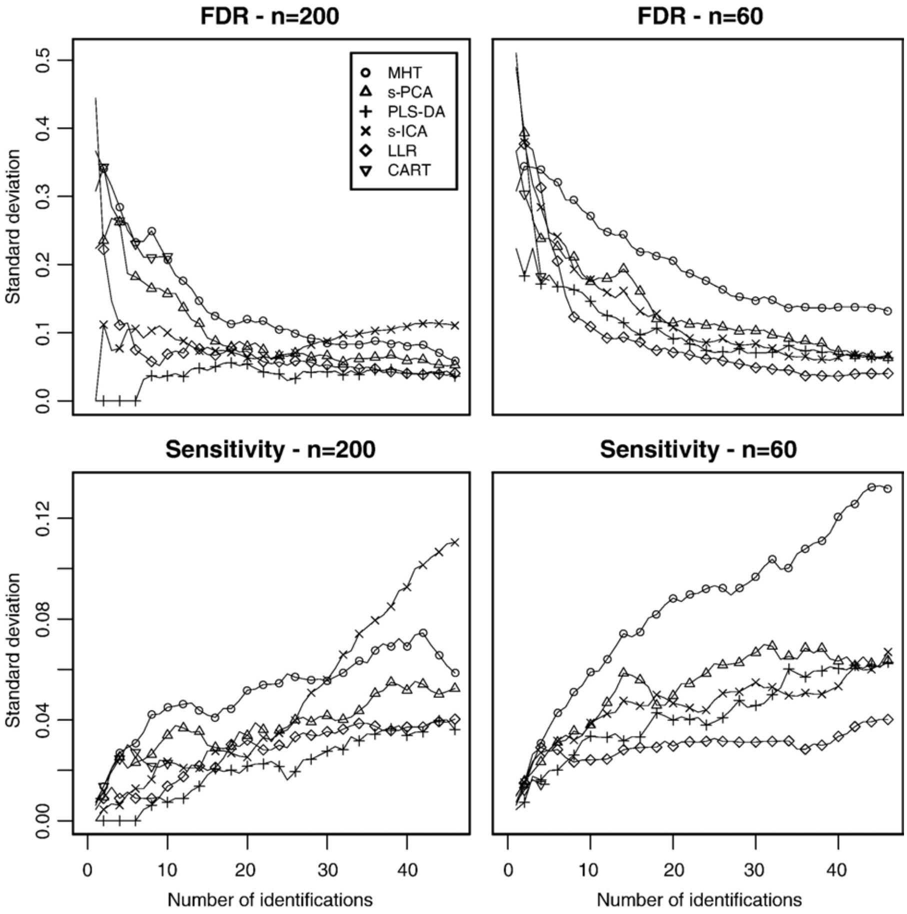

Fig. 8. Standard deviation of FDR (top) or sensitivity (bottom) versus number of identifications for the 6 methods. For clarity, curves only show symbols representing odd number of identifications.

The CART method does not lead to good performances. As a predictive tool, it identifies only independent biomarkers. Moreover, CART only succeeds to identify independent biomarkers in low-noise regions. The advantage of the CART method resides in its tree representation that is easily interpretable, but coupled to low performances and to be used in non-noisy problems only, i.e. non realistic situation.

Finally, let us remind that PLS-DA, LLR and CART are the only predictive methods among the six ones, providing them a further advantage when prediction is also an objective of the study.

5.4. Variability of the results

In addition to the comparison of the mean performances of the methods over 20 runs, it is important to characterize the variability of the results among the runs. A high variability can be considered as a drawback as it makes results less repeatable, but can also be exploited to extract additional information (as detailed for instance in the LLR paragraph above).

Fig. 8 shows the standard deviations of the sensitivity and of the FDR (top and bottom respectively), in the 200 and 60 spectra cases (left and right respectively) among the 20 datasets. As it can be seen, the standard deviation of both the sensitivity and the FDR are high in the MHT and s-PCA methods. In the MHT case, it even increases in the 60-spectra case, proving a low robustness to small samples. On the other hand, PLS-DA and LLR are the more stable methods, LLR being clearly the winner in small sample. In the ICA case, the surprising result that the variance is higher in the 200-spectra case than in the 60-spectra one comes from the difficulty that the ICA method has to handle high-dimensional signals [21]. Coupled to the fact that the ICA method is more robust to noise than other ones, it can be concluded that the best situation to use ICA is with noisy small samples.

6. Conclusions

Metabonomics is emerging as a valuable tool in a number of biological applications. Althought, the choice of efficient chemometric methods for biomarkers identification in ^{1}H NMR based metabonomic remains an important research topic. This paper proposes to revisit the traditionally used PCA method and to explore more advances chemometrics and statistical tools to identify biomarkers from ^{1}H NMR spectra classified in two groups according to a stressor factor of interest. Each proposed method delivers biomarker scores to indicate which metabolites of the analyzed biofluid are affected by the stressor factor. The application of each method to samples of 60 and 200 spectra issued from a semi-artificial database has allowed to observe the following properties: easiness of interpretation of the results, robustness to noise, ability to identify biomarkers. ROC curves have been used to represent method false discovery rates and sensitivities.

Due to their high sensitivities to noise, the CART and the improved PCA methods have shown bad performances in comparison to the other methods. They are then not recommended in spectral databases where the signal to noise ratio and the number of spectra are low. ROC curves of PLS-DA and s-ICA methods have shown good and competitive biomarker identification performances. However, each of them presents specific relevant characteristics. The s-ICA method is robust to noise and more interpretable as it is able to recover independent metabolites from complex spectra. The PLS-DA method is very easy to apply and is efficient in recovering independent biomarkers. As it identifies only independent biomarkers, the LLR method can not be directly compared to the others. Nevertheless, it has shown to be very efficient in the context by providing automatically the smallest number of identifications for an already satisfying proportion of correct independent biomarkers. The main advantage of the last tested method, the MHT method is its simplicity coupled with overall acceptable performances especially in large samples.

This work motivates numerous further developments. First, from the application side, the methods are currently tested by the authors on real biological ^{1}H NMR databases. Their goal is to ensure that the identifications coming out of the methods correspond to metabolites present in the biofluids analyzed. Moreover, they want to verify if the ability of s-ICA to recover independent biomarkers can also be observed on real data. On the methodological side, the good results of PLS-DA and LLR motivate to explore the performances of two tools: the Penalized Logistic Regression [22] and the LLR-PLS [23]. The high sensitivity to noise of CART suggests exploring more robust related methods as the Random Forests [24]. Finally, the FDR based criterion used in MHT must clearly be improved.

Acknowledgements

The authors are grateful to Eli Lilly and the Consortium for Metabonomic toxicology, especially Jeanmarie Collet, for providing data. They also thank Paul Eilers for stimulating discussion.

References

[1] J. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Metabonomics: a generic platform for the study of drug toxicity and gene function, Nature Reviews Drug Discovery 1 (2002) 153–161.

[2] E. Holmes, H. Antti, Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra, Analyst 127 (2002) 1549–1557.

[3] J.C. Lindon, E. Holmes, J. Nicholson, Metabonomics techniques and applications to pharmaceutical research and development, Pharmaceutical Research 23 (2006) 1075–1088.

[4] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, The Annals of Statistics 29 (2001) 1165–1188.

[5] V. Tusher, R. Tibshiriani, G. Chu, Significant analysis of microarray applied to the ionising radiation response, PNAS 98 (2001) 5116–5121.

[6] J. Storey, A direct approach to false discovery rates, Journal of the Royal Statistical Society. Series B 64 (2002) 479–498.

[7] J. Storey, D. Siegmund, Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach, Journal of the Royal Statistical Society. Series B 66 (2004) 187–205.

[8] I. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.

[9] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (2000) 411–430.

[10] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17 (2003) 166–173.

[11] H. Martens, T. Næs, Multivariate Calibration, Wiley, Chichester, UK, 1989.

[12] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometrics and Intelligent Laboratory Systems 18 (1993) 251–253.

[13] B. Mevik, H.R. Cederkvist, Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR), Journal of Chemometrics 18 (2004) 422–429.

[14] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley, New York, 1989.

[15] H. Akaike, A new look at the statistical model identification, IEEE Tras. Automat. Consr. AC, vol. 19, 1974, pp. 716–723.

[16] L.J. Breiman, R. Friedman, R. Olsen, C. Stone, Classification and Regression Trees, Wadsworth, Pacific Grove, CA, 1984.

[17] F. Esposito, D. Malerba, G. Semeraro, A comparative analysis of methods for pruning decision trees, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 476–491.

[18] J.C. Lindon, J.K. Nicholson, E. Holmes, H. Antti, M.E. Bollard, H. Keun, O. Beckonert, T.M. Ebbels, M.D. Reily, D. Robertson, G.J. Stevens, P. Luke, A.P. Breau, G.H. Cantor, R.H. Bible, U. Niederhauser, H. Senn, G. Schlotterbeck, U.G. Sidelmann, S.M. Laursen, A. Tymiak, B.D. Car, L. Lehman-McKeeman, J.M. Colet, A. Loukaci, C. Thomas, Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project, Toxicology and Applied Pharmacology 187 (2003) 137–146.

[19] Vanwinsberghe J. Bubble: development of a matlab tool for automated ^{1}H NMR data processing in metabonomics, Master's thesis, Université de Strasbourg, 2005.

[20] J.P. Egan, Signal Detection Theory and ROC Analysis, Academic Press, New York, 1975.

[21] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, USA, 2001.

[22] P. Eilers, J. Boer, G. Van Ommen, H. Van Houwelingen, Classification of microarray data with penalized logistic regression, Proceedings of SPIE Progress in Biomedical Optics and Images 4266 (2001) 187–198.

[23] V. Nguyen, D. Rocke, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics 18 (2002) 39–50.

[24] LJ. Breiman, Random forests, Machine Learning 45 (2001) 5–32.