I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2013/06

Combination of Independent Component Analysis and statistical modelling for the identification of metabonomic biomarkers in ^{1}H-NMR spectroscopy (second version)

ROUSSEAU, R., FERAUD, B., GOVAERTS, B. and M. VERLEYSEN

Combination of Independent Component Analysis and statistical modelling for the identification of metabonomic biomarkers in ^{1}H-NMR spectroscopy (second version)

Réjane Rousseau[1,2], Baptiste Féraud*[1], Bernadette Govaerts†[1], and Michel Verleysen‡[1,3]

[1]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université Catholique de Louvain, Belgium

[2]Arlenda S.A., Belgium

[3]SAMOS-MATISSE Team, Université Paris I, Panthéon - Sorbonne, France

March 28, 2013

Abstract

In order to maintain life, living organisms product and transform small molecules called metabolites. Metabonomics is a recent scientific platform, studying the development of biological reactions caused by a contact with a physio-pathological stimulus, through these metabolites.The ^{1}H-NMR spectroscopy is widely used to graphically describe a metabolites composition via spectra. Biologists can then confirm or invalidate the development of a biological reaction if specific NMR spectral regions are altered from a given physiological situation to another. However, this process supposes a preliminary identification step which traditionally consists in the study of the two first components of a Principal Component Analysis (PCA). This paper presents a new methodology in four steps providing knowledge on specific ^{1}H-NMR spectral areas (and by extension on biomarkers) via the identification of biomarkers as such, and via the visualization of the effects caused by some external changes. A first step implies Independent Component Analysis (ICA) in order to decompose the spectral data into statistically independent components or sources of information. The independent (pure or composite) metabolites contained in biofluids are discovered through the sources, and their quantities through mixing weights. The advantages of independent components are described in comparison with usual PCA analysis. Specific questions related to ICA like the choice of the number of

*baptiste.feraud@uclouvain.be

†bernadette.govaerts@uclouvain.be

‡michel.verleysen@uclouvain.be

components or their ordering will be discussed here. The second step consists in a statistical modelling applied to the ICA outputs. The third step will introduce statistical hypothesis tests on the parameters of the estimated models, with the objective of selecting sources which present biomarkers or significantly fluctuating spectral regions according to our factor of interest. A panel of various statistical models is considered here, that can adapt to different possible kinds of data or different investigations. Finally, the last step proposes a computation of contrasts which can lead to the visualization of changes on spectra caused by changes of the factor of interest. The whole methodology is illustrated on two experimental datasets.

Keywords: Metabonomics, multivariate statistics, Independent Component Analysis, biomarker identification, ^{1}H-NMR spectroscopy, linear mixed models.

1 Introduction

Metabonomics, one of the most recent section in the world of "Omics", extracts biochemical information reflecting somel biological events. This science studies how the metabolic profile of a complex biological system changes in response to stresses like diseases, toxic exposures or dietary changes. In more technical terms, metabonomics is defined as *"The quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification"*.[1]

Proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy generates spectral profiles describing the composition of metabolites in collected biofluid samples. A comparison of several spectra in various specific states can permit a preliminary graphical and qualitative investigation of changes in biofluid metabolite composition inherent to the presence of a stressor. However, the complexity of ^{1}H-NMR spectra and the high number of spectra (of samples) necessary for metabonomic studies imply an automated data analysis. In addition, systematic differences between samples are often hidden behind biological noise and/or behind peak shifts.

Adequate data pre-processing and multivariate statistical methodologies are then required to extract spectral regions with stable differences between spectra obtained in various conditions. These regions, directly linked with biomarkers, are assumed to be associated with the alteration of an endogenous metabolite in reaction to the contact with a considered stressor. A biomarker can then be isolated to detect and follow changes in biological systems. Beside this goal of biomarker identification, statistical analysis, through predictive models, can also be used on ^{1}H-NMR spectra to provide the probability of the occurrence of a biological reaction.

The first and the most common chemometric tool used in preliminary metabonomic studies is Principal Component Analysis (PCA).[2] [3] This method is a starting point for analyzing multivariate data and can rapidly provide an overview of the hidden information. Data are presented as a two dimensional plot (scoreplot) where the coordinate axis correspond to the two first principal components. If spectra differ according to a specific characteristic, the scoreplot reveals the presence of natural clusters in the datasets. An examination of the loadings leads to identify biomarkers or key portions of the ^{1}H-NMR spectra giving rise to these regroupments.

Sometimes, variations within groups are bigger than variations between groups, resulting in a scoreplot with clusters that overlap or do not directly correlate to the studied characteristics. In such cases, additional information can be extracted by using other data decomposition methods such as partial least squares (PLS), discriminant PLS (PLS-DA) or orthogonal PLS (O-PLS). As PCA, these methods look for systematic variances between samples. In contrast, they use information about samples such as groups of the characteristic of interest. Therefore, these methods often allow a better separation of samples and a clearer identification of significant biomarker variables, but are biased (in contrast to PCA).

Unfortunately, in many publications, PCA is the only statistical standard while it remains a highly questionable procedure. Haluska and Powers have underlined the negative impact of the PCA sensitivity to noise for the analysis of ^{1}H-NMR data [4]: very small and random fluctuations within noise of the ^{1}H-NMR spectrum can result in irrelevant clusters in the scoreplot formed by the two first principal components. They propose to remove noisy regions by only using signals above a chosen peak intensity threshold during PCA.

In a previous work, Rousseau and al presented an improved PCA method [5]. If random noises or even unexpected variations are contained within data, clusters of interest can be presented in later factors of the PCA decomposition. They proposed a methodology to identify the two factors that discriminate the most two classes of spectra, and help them to form a proper scoreplot. In this paper [5], several statistical methods for the identification of metabonomics biomarkers in ^{1}H-NMR spectroscopy are also suggested and compared. Discriminant Partial Least Square (PLS-DA) and a method based on independent component analysis (ICA) showed good competitive performances. On the contrary, even with the proposed improvement, the PCA biomarker identification demonstrated a general low efficiency.

We propose here to expand the use of ICA for the identification of specific ^{1}H-NMR spectral regions that are discriminant for two or more categories of spectra. The previous promising results are motivating and encourage to go futher into the study of the use of ICA for the analysis of metabonomics data. Additionally, applications of ICA in contexts quite similar to metabonomics, as genomics [9] [10] and even in MASS spectroscopy metabonomics [11], have shown that ICA clearly outperforms PCA. ICA has also the advantage to share similitudes with better known PCA. Both of them are projections methods which linearly decompose data into components. As for PCA, the ICA results can then be supported by visual representations. Anyway, the ICA components have a more stringent nature than principal components: PCA decomposes data into uncorrelated components of maximal variance when ICA attempts to achieve an even greater objective by modeling the data as a linear mixture of maximally independent components. For non-gaussian data, the general structure can be more naturally explained and ICA is likely to be successful in this context because most biological variance sources have exactly non-gaussian distributions. The independence of the components is also adequate for biological interpretation because the analyzed biofluid (e.g. plasma, urine) can be seen as a mixture of unrelated metabolites and ^{1}H-NMR spectra may then be interpreted as weighted sums of ^{1}H-NMR spectra of these independent metabolites. The application of ICA should then ideally recover components which

may represent the independent metabolites contained in the media.

In this context, this paper proposes a methodology for the identification of ^{1}H-NMR metabonomic biomarkers in 4 steps (Figure 1). The first step is the implementation of ICA in order to reduce the dimension and decompose the multivariate spectral dataset into statistically independent components. We then propose solutions to select the optimal number of components to estimate and to order these components. Comparatively to usual PCA analysis, we demonstrate the usefulness of independent components to overview data and to search for outliers. The second step of this methodology consists in a statistical modelling of the ICA results. We consider a panel of various mixed linear statistical models adapted to the nature of the domain. In a third step, the model coefficients and appropriate multiplicity corrected statistical tests are used to decide which ICA sources can be considered as biomarkers of the stressor(s) of interest. Finally, in a fourth step, a method is proposed to visualize the stressor effect on ^{1}H-NMR spectra. In other words, all the models are then used to identify biomarkers and to visualize the effects of the experimental factors on these biomarkers.

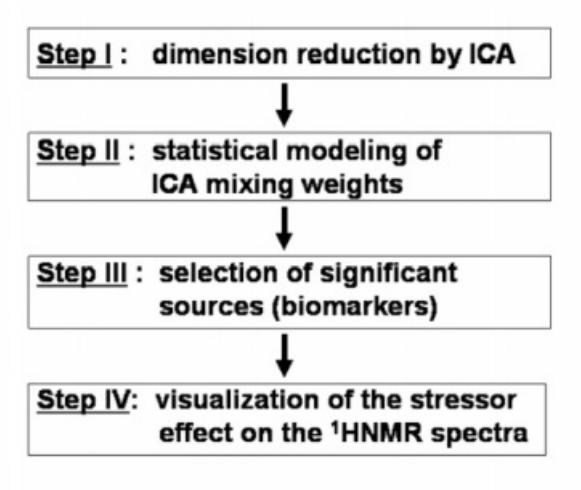

Figure 1: Methodology steps

This article is organized as follow. Section 2 provides a presentation of both typical metabonomic data and experimental data used in this paper. Section 3 presents the first part of the methodology, the ICA dimension reduction, with beforehand a general presentation of ICA. The third subsection introduces a criterion to measure the amount of information contained in the obtained components. This measure allows us to order the components in a similar way to the percentage of explained variance used in PCA. In Section 4, we propose to use the mixing weights resulting from ICA in combination with statistical linear models in order to identify (and perhaps interpret) biomarkers. Section 5 presents the third step consisting in a selection of sources that describe biomarkers based on the significance of the models estimated in step 2. In section 6, from the selected sources, we propose to compute contrasts to visualize the spectral effects when one factor of interest changes. All the concepts presented in sections 2 to 6 are illustrated on a simple experimental dataset. Finally, section 7

illustrates the methodology through an application linked with a more complex ^{1}H-NMR metabonomic dataset and section 8 introduces an example on real data related to Age related Macular Degeneration (AMD).

2 Data description

2.1 Typical metabonomic 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 ID (animal or human) and characteristics, treatment, dose or time of sampling. A ^{1}H-NMR dataset contains the spectral evaluations of biofluid samples which were 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 metabonomic 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 that each molecule can have more than one spectral peak and hence may contribute to more than one descriptor. As a large variety of dynamic biological systems and processes are reflected in spectra, a range of physiological conditions, for example the nutritional status, can also represent a source of variability into spectra. Noise and biological fluctuations are thus natural and inevitable in spectral data. Each spectrum in the ^{1}H-NMR dataset is also usually linked with 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 via the current gold-standard examinations (biological measures or hysto-pathological ones) generated for the subject for which spectra are measured.

2.2 Experimental data

2.2.1 The experimental plan

An experimental plan (see Figure 2) was designed in order to provide a database in which one controls the alterations of known descriptors. In this experiment, homogeneous urine samples were spiked with two products at different levels of concentration and analyzed through spectroscopy. These products are citric acid ("citrate") and hippuric acid ("hippurate"). They were added to urine at four levels of concentrations, respectively 0, 2, 4 and 8 mM for citric acid, and 0, 1, 2 and 4 mM for hippuric acid.

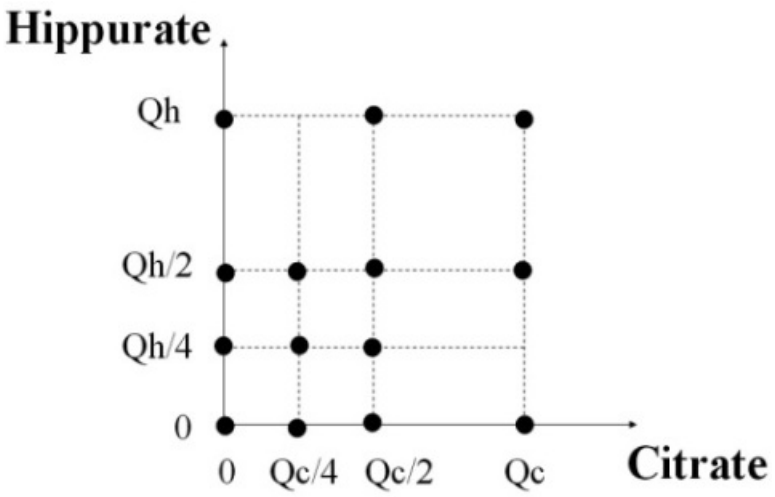

Figure 2: Experimental design

As shown in Figure 3, the peaks corresponding to each product are located in distant areas. The hippurate is characterized by three peaks, with two of them in the low field region containing a low noise level. On the contrary, citrate peaks are located in the noisy region. In spectral pre-processing, note that these peaks are aggregated in one to avoid alignment problems (see Section 2.2.4).

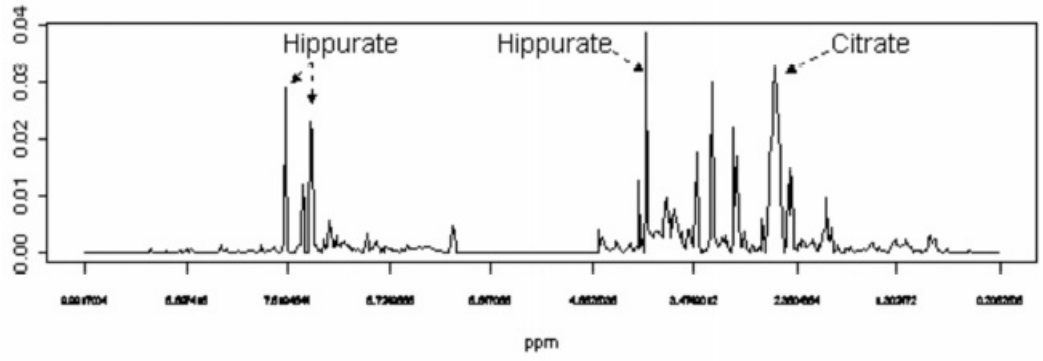

Figure 3: A typical urine spectrum with spiked citrate and hippurate

This experimental design was repeated several times with different conditions (two water dilutions, five days, two replicates per day) but only 28 spectra are used in this paper. It corresponds to two replicates of the 14 point design of one day of experiment and in only one water dilution. All spectra are available from the first author under request.

2.2.2 Hypothetical metabonomic study

In this paper, the spectra obtained from the designed experiment are used to mimic a typical metabonomic study. We will suppose that 28 subjects having four different levels of age (the four levels of citrate) have

received four different doses of a drug (the four levels of hippurate). The hypothetical goal of the study is to find a critical spectral region or a biomarker which synthesize the drug effect on the urine. The "discovered" biomarker will hopefully have the shape and position of the hippurate peaks but obviously no information on these peaks is provided in the methodology.

2.2.3 Sample preparation and acquisition of the ^{1}H-NMR data

The two products (citrate and hippurate) were first mixed with phosphate buffer containing TSP. The volume of buffer was adapted in order to obtain a volume of 600μl to add to a urine sample. Each urine sample came from a pool of 344 female Fischer rats and had a volume of 1200μl. Each mixture was added to a urine sample, centrifuged, frozen at 80 ^{o}C and unfrozen at 40 ^{o}C the day before the ^{1}H-NMR analysis. Samples were then analyzed randomly within each day of experiment. NMR measurements were made with a 600 MHz Bruker spectrometer with 4mm FI-SEI ATM probe. The spectral information is then included in 28 individual free induction decays (FIDs).

2.2.4 The post-acquisition treatments

Each acquired spectrum was processed using Bubble, a MATLAB tool for automatic processing and for reducing NMR spectra. [17]. Bubble performs in sequence : suppression of the water resonance, apodisation (with a line broadening factor of 1Hz), Fourier transform, phase correction, baseline correction using a Whithaker smoother [16], median normalization and warping in order to align shifted peaks. The last step of the Bubble process reduces, by simple integration, the part of the spectrum situated between 0.2 and 10 ppm to 600 descriptors. We manually added several pre-processing tools to spectra prepared by Bubble. First, we replaced all the negative values by zero. Secondly, we set to zero the ppm values corresponding to the large non-informative urea peak and to the already treated water peak (4.5-6.0 ppm). Then, the spectral region around the citrate resonances (2.56-2.72 ppm) was integrated and summarized in just one peak to suppress high shifts. Finally, we normalized for a second time the dataset. Indeed, the effect of the first normalization by the median, necessary to realize an accurate warping, is cancelled due to the reduction. The second normalization consists in a constant sum normalization : each spectrum is divided by the sum of intensities for all its ppms values.

2.3 Notations

Let X be the $(m \times n)$ matrix of spectral data containing n spectra, each of them being described by m descriptors. Y is a $(n \times l)$ matrix of design data describing each sample or spectrum by l variables. One of these variables describes the characteristic related to the biomarkers: y_k. In our experimental data, n=28, m=600 and two of these design variables correspond to the citrate and hippurate concentrations, later assimilated as subject age and drug dose.

3 First step of the methodology: Independent Component Analysis

3.1 The ICA theoretical principles

The basic idea of Independent Component Analysis (ICA) is to reconstruct from observation sequences original sequences that are assumed to be independent. ICA is a multivariate analysis tool which aims at separating or recovering unobserved multidimensional independent signals from linearly mixed observed ones [7] .

ICA was originally developed for signal processing to solve the problem of blind source separation (BSS) [12]. In this context, the aim of BSS is the recovery of a number of original signals when only a mixture of them is available.

In the basic noiseless ICA model, each observed signal is a mixture of unknown statistically independent signals (named sources or components):

$$X = SA^T \tag{1}$$

with X denoting the $(m \times n)$ matrix that contains n original signal vectors of m observations (x_i), S denoting the $(m \times q)$ matrix that contains q unknown source vectors s_j . The relative contribution of each component to the expression profile for a given sample is determined by the coefficients of the unknow $(n \times q)$ mixing matrix A^T. Finally, the "unmixing" problem considered by ICA is to recover S. The goal of ICA is to find a demixing matrix W such that the sources can be estimated by $\hat{S} = X.W$ where $\hat{S}$ denotes the matrix formed by q estimations of scaled independent source vectors s_j (as columns).

The ICA model introduces an ambiguity in the scale of the recovered sources. It results from the fact that scaling a source by a factor λ is exactly compensated by dividing the corresponding column of the mixing matrix by λ. A natural way for fixing the magnitudes of independent components is thus to assume that each component has unit variance. It should be noted that the ambiguity of the sign remains as we can multiply any component by -1 without affecting the model.

The key assumption of ICA is that the sources have to be statistically independent. Under the ICA model, the observed data tend to be more gaussian than the independent components due to the Central Limit Theorem (the distribution of a sum of independent random variables is generally more gaussian than the summands). Thus, the independence of random variables can be reflected by non-gaussianity. Solving the ICA problem aims then at finding a matrix W and maximising the non-gaussianity of the estimated sources, under the constraint that their variances are constant.

Two classical measures of non-gaussianity are the kurtosis (the fourth-order cumulant) and the negentropy. Although the idea of maximising the kurtosis seems quite simple, it can be very sensitive to outliers.[13] In this paper, we used an algorithm based on the maximisation of the negentropy, the FastICA algorithm proposed by Hyvärinen.[8] The entropy of a random variable Y, which is the basic concept of information theory, is defined

as follow :

$$H(Y) = -\int f_Y(y) log(f_Y(y)) dy \tag{2}$$

In information theory, among all random variables of equal variance, the normal one has the largest entropy. The FastICA algorithm uses a contrast function called the negentropy J ,defined by :

$$J(Y) = H(Y_{gauss}) - H(Y) \tag{3}$$

where Y_{gauss} is a gaussian random variable with the same covariance matrix as Y. The main disadvantage of using negentropy is that it is computationnally intensive because it requires the estimation of the probability density function. Therefore, a simpler approximation of negentropy is used in FastICA. [?]

Before applying this algorithm to the data, some pre-processing steps are necessary. First, to simplify the theory and the algorithm, one assumes, without loss of generality, that both the mixture variables and the independent components have zero mean. This assumption is achieved by centering each observed signal vector. The second step, called "Whitening", allows the ICA algorithm to transform and reduce the dimension of the signal matrix X to a $(m \times q)$ matrix of orthogonal vectors T in order to reduce the number of parameters to be estimated. Columns of T are linear combinations of original signal vectors and obtained by PCA with unit variances. The number q of sources to be computed can be fixed in this step via a method discussed in Section 3.3.

3.2 ICA on metabonomic data

In the context of metabonomic ^{1}H-NMR data, the analyzed biofluid (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 components included in the mixture, interpretable as spectra of pure or complex metabolites.

3.2.1 Algorithm application

The FastICA algorithm is applied to the spectral matrix as follows:

- Pre-processing step 1: center X by columns:

  $$X^c = X - \mathbf{1_m} \cdot \tilde{X}$$

  where $\tilde{X}$ is the $1 \times n$ vector of spectral means and $\mathbf{1_m}$ a $m \times 1$ unit vector.

- Preprocessing step 2 ("Whitening"): reduce by PCA the $(m \times n)$ matrix X^c to a $(m \times q)$ matrix of scores T $(q \leq \min(n, m))$:

  $$X^c = T^\star P^\star = TP + E$$

The column vectors of the full score matrix $T^\star$ are centered, uncorrelated and their variances are equal to one. In other words, the variance-covariance matrix of $T^\star$ equals the identity matrix: $Var(T^\star) = I_n$. $P^\star$ is a $(n \times n)$ matrix defined on the basis of the eigenvectors of the covariance matrix $(X^{cT}X^c)/n$. Note that this PCA differs from usual PCA for metabonomic biomarkers identification as the resulting components are linear combinations of observations (spectra) and not of variables (spectral descriptors), and centering is done by spectra and not by descriptor. The number of sources q to be estimated can be fixed to less than $\min(n, m)$. This is performed here by selecting the q first scores vectors (columns) of $T^\star$ in order to build the $(m \times q)$ matrix T. P is then defined as the q first lines of $P^\star$ and E is the error matrix. The choice of q is discussed in section 3.3.

- ICA based on T and calculation of S and A^T. The fastICA algorithm, with parallel extraction of components, proceeds in the following steps:
  - compute a $(q \times q)$ unmixing matrix W such that $TW = S$ where S is the $(m \times q)$ matrix of independent sources. W is chosen to maximize the negentropy of the columns of S.

    As the variance of TW must be equal to one, this is equivalent for the whitened data to constrain the norm of W to be unity .
  - define the mixing matrix A as $A = W^{-1}P$ in order to obtain the ICA decomposition $X^c = SA + E$.

The $(m \times q)$ matrix S contains q estimated independent components (IC),s_j. Each s_j has a zero mean and a unit variance, and at least $(q-1)$ sources are non gaussian. The A mixing matrix is a $(q \times n)$ matrix. Each column a_j is then a $(n \times 1)$ vector containing the weights or contributions of the corresponding source s_j during the construction of the n observed spectra. A source s_j playing a major role in the contribution of an observed spectrum x_i has then a potentially large absolute value $|a_{ij}|$.

3.3 Choice of the number of sources to estimate

One important parameter that may change ICA results is the number q of estimated components. The real number of independent sources contributing to the signal is obviously unknown and has to be guessed. In the ICA theory, it is supposed that the number of sources is less than or equal to the number of observed mixtures: $q \leq n$. This represents a required condition to avoid overlearning effects. Moreover, to make the implementation of the fastICA algorithm effective, the maximal value for q is the smallest dimension of its input matrix. Indeed, the data matrix used as input in the algorithm is the whitened matrix. The maximal number of sources to be computed is then fixed by the score matrix T, with $q \leq \min(n, m)$. In ^{1}H-NMR metabonomic datasets, the resolution of a spectrum m is typically higher than the number of spectra n. The maximal value q will then correspond to the number n of observed spectra: $q \leq n$. Anyway, when n is large, the choice of $q = n$ can

produce convergence problems or very high computational costs. On the other hand, q should be large enough to allow sufficient freedom or richness of choice for the feature selection algorithms to be powerful.

To avoid convergence problems, the number of sources is here limited to a chosen value $q < n$ by discarding some score vectors obtained via the whitening matrix $T^\star$. The selection of these vectors is based on the PCA natural ordering of the columns of $T^\star$ according to the eigenvalues λ_j of $X^{cT}X^c$. The q first vectors of scores associated with the largest eigenvalues are selected to form the matrix T of reduced dimensions $(m \times q)$ with $q < n$. This keeps only components which explain most of the variance in the data and discards those describing noise. Let us define D_q the proportion of the variation of X^c explained by the first q principal components:

$$D_q = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{n} \lambda_j}$$

We propose to choose q on the basis of a screeplot in order to be quite sure to preserve enough original information.

3.4 Measure of the information contained in ICA sources

In ICA, there is no natural ordering of the computed sources. This section presents a possible alternative. Given a set of q estimated sources s_j, we can reconstruct the data as $\hat{X}^c = SA^T$. Let us define the error (in the reconstruction) observed with only the source s_j by:

$$E_j = (X^c - s_j.a_j^T) = S_{\neq j}A_{\neq j}^T$$

This error is equivalent to the data reconstructed with all the other sources contained in the $(m \times (q-1))$ matrix $S_{\neq j}$. For sources with zero mean and unit variance, it can be shown that a measure of the proportion of the variation in T explained by s_j is:

$$R_j^2 = 1 - \frac{tr(E_j^T E_j)}{tr(\hat{X}^T \hat{X})} = \frac{\sum_{i=1}^{n} a_{ij}^2}{tr(A^T A)}$$

The proportion of the variance of signals in X^c explained by a source s_j is then defined by:

$$C_j = \frac{\sum_{i=1}^{n} a_{ij}^2}{tr(A^T A)} \times D_q$$

with D_q the proportion of variance explained by the q scores in T. Below, the s_j are ordered according to their respective C_j.

3.5 Example

In this section, this ICA procedure is applied on the dataset described in section 2.2.1. It involves n=28 spectra with m=600 ppms, each corresponding to the two replicated samples of each of the 14 mixtures of urine. As our samples are mixtures of three products, we ideally expected to find three independent sources of variation:

the variation of the urine spectra and the respective variations due to the citrate and hippurate peaks. Anyway, we supposed that we do not know that data come from an experimental design, and have based the number of calculated sources on the percentage of explained variance of the principal components (PCs) in the whitening stage. Based on the screeplot (see Figure 4), we choose to calculate q=6 sources. The percentage of explained variance with these first six PCs is of the order of $D_j = 0.9796$.

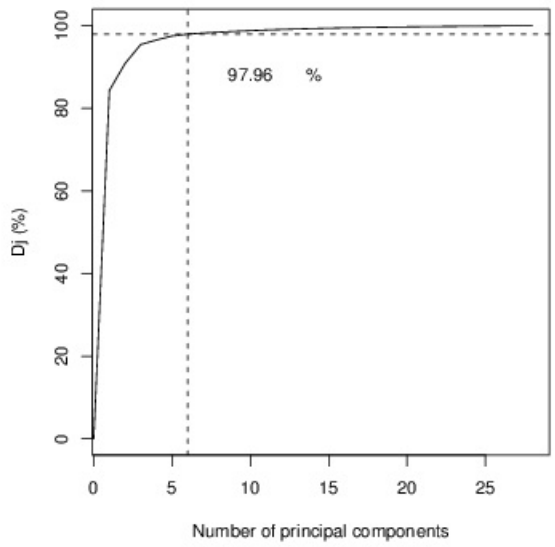

Figure 4: Screeplot of the % of explained variance with the first q PCs (from the "PCA-whitening")

The FastICA algorithm was then applied to the (600×6) T matrix. Figure 5 presents the six computed ICA sources and one can directly see that the goal is reached. Source 1 (38.18% of the information) represents a typical urine spectrum, source 2 (29.62%) the spectrum of pure citrate and source 3 (27.86%) the spectrum of hippurate. The three last sources explain a very low amount of information and may be attributed to noise. Note that, out of product peaks, sources 2 and 3 present very low noise. This is an advantageous characteristic of ICA compared to PCA (see below).

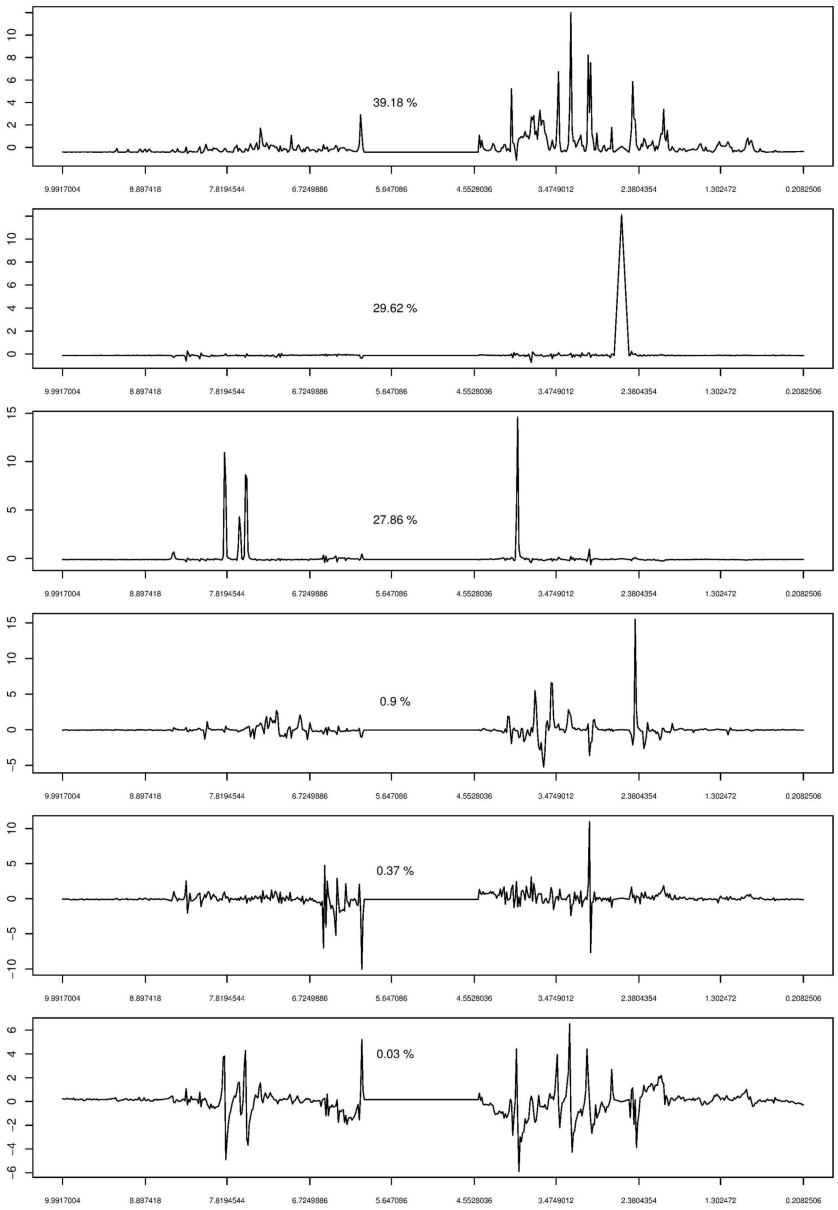

Figure 5: the $q = 6$ ICA sources

Figure 6 presents, on a scatter plot, the values of the mixing weights a_{ij} for sources 2 and 3 and for the 14×2 experiments. The shape of the experimental design can directly be recognized. This illustrates that mixing weights give a direct idea of the amount of each metabolite in the mixture. The diamond shape of the design is due to the fact that citrate and hippurate quantities have been added to a constant quantity of urine and are consequently not real proportions. The positive values of all weights take into account the fact that pure urine already contains a certain amount of citrate and hippurate.

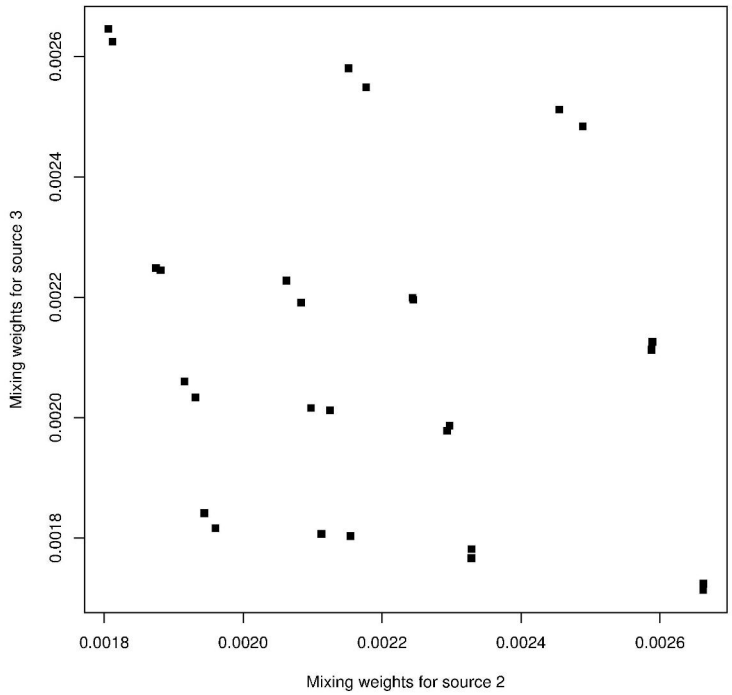

Figure 6: Mixing coefficients for sources 2 and 3

3.6 Comparison between ICA and PCA

ICA and PCA are both methods allowing dimensional reduction but using a different principle to choose the directions of their components. PCA results in uncorrelated axes with directions computed from the second order statistics while ICA provides statistically independent axes with directions actually based on the second and higher orders. The statistical independence of ICA sources is a stronger concept than the non-correlation of the principal components from PCA. If the variables are independent, they are uncorrelated, while uncorrelatedness does not imply independence. For this reason, ICA can then be seen as a generalization of PCA which can highlight high-order dependencies in addition to correlations. ICA also provides more natural and biologically meaningful representations of the data. The independent components are also more suitable for our study than uncorrelated components: in metabonomic data, the components (metabolites) of interest are not systematically in the direction of the maximal variance.

Figure 7.a shows an experimental design allowing to illustrate the advantage of ICA on PCA. In these data (24 extracted spectra), PCA should ideally choose the two first directions shown in Figure 7.b. These directions represent a variation of both citrate and hippurate. As it can be seen in Figure 8.a, each of the two corresponding loading vectors includes spectral representations of both products. Figures 7c and 8b respectively present the three directions ideally chosen by ICA and the corresponding sources. Each ICA direction corresponds to one of the three products with independent concentration contained into the samples. Moreover, PCA loadings

contain much more noise than ICA sources, thus increasing the confusion when searching for biomarkers. This noise induces a worse projection of the design in the score plots (this can be seen on the replicates).

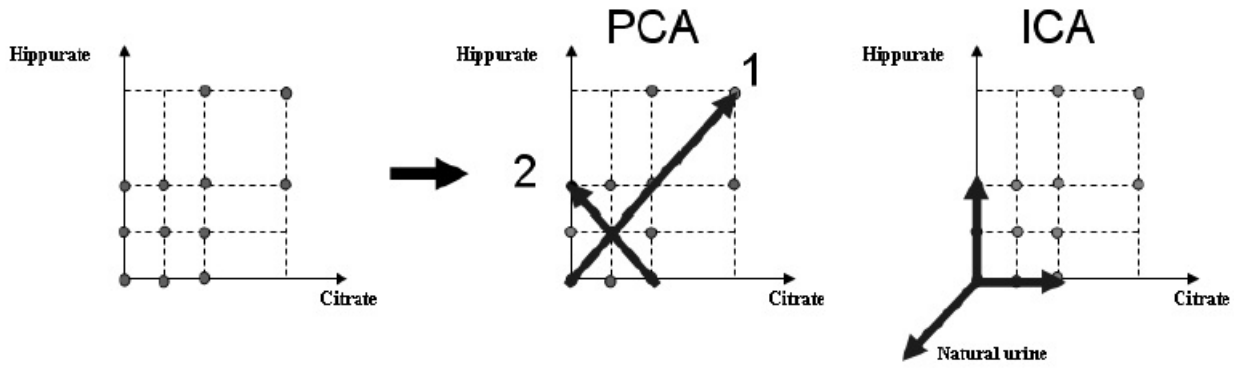

Figure 7 a b c: Component directions chosen by PCA and ICA on an illustrative experimental design

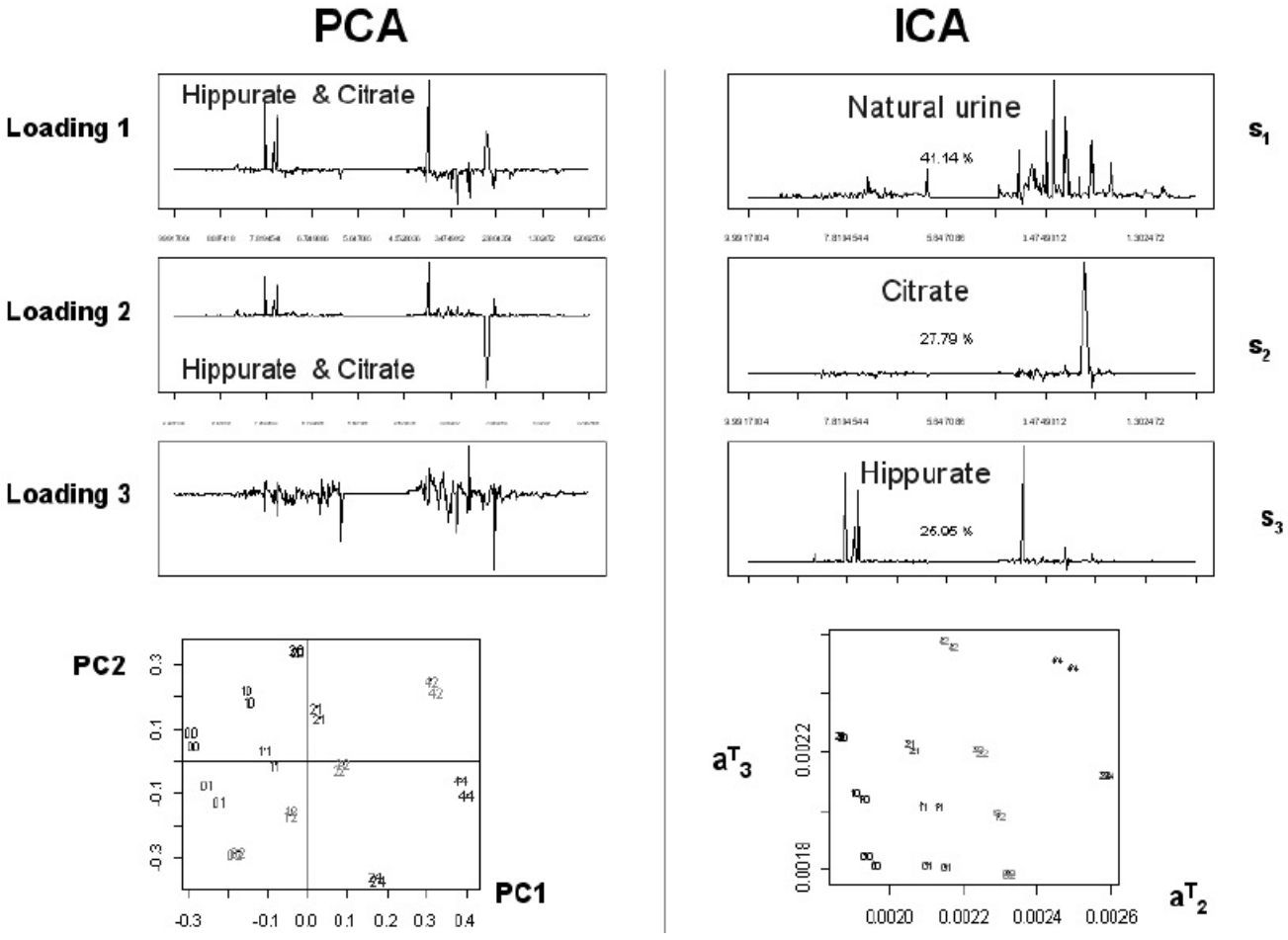

Figure 8 a b: Loadings and scores for PCA - sources and mixing weigths for ICA

Another advantage of ICA over PCA in metabonomics is the fact that ICA searches for non-Gaussian sources, and biological sources are typically non-Gaussian. They have either sub -or super-Gaussian distributions (thicker or thinner tails than Gaussian). PCA is most successful in case of Gaussian cases only.

ICA has of course some drawbacks which must be emphasized. ICA requires to choose the number of components to compute. The dimensions of the unmixing matrix to estimate can be fixed to obtain a number of sources equal or less than the number of variables, and the final independent components depend on this postulated number. This is not the case in PCA. In a lot of situations, as in metabonomic studies, the real number of independent contributions to the signal is unknown.

Another difference is the ordering of the components. In PCA, the components are naturally ordered from the singular values of the data matrix and are used to decrease the dimension of the problem (by considering only the first components which explain most of the variance in the data). In ICA, the sources have no order and the order in which the sources in S are listed by the algorithm is irrelevant to their independence. In section 3.3, we proposed a measure of the amount of information contained in each estimated source, giving rise to some ranking.

Finally, in contrast to PCA, all ICA algorithms face the problem of convergence to local optima, thus slightly different components will be produced when the same data is reanalyzed. It is then recommended to run ICA algorithm several times and check the stability of the results when using different values of q.

4 Step II: Statistical modeling

4.1 Goal and principle

The second step of the methodology aims to fit a statistical model in order to identify metabonomic biomarkers from ICA results. More precisely, the model will search for a link between the ICA mixing weight matrix A and the design factors of the metabonomic study.

The logic underlying this approach is the following. An ^{1}H-NMR spectrum reflects the concentrations of pure or complex metabolites contained in the analyzed sample. The design factors, as for example the dose of an administrated drug, can influence these concentrations and consequently modify the spectra in a specific way. The methodology presented in this paper supposes that the q sources recovered by ICA are the spectral images of pure or complex metabolites that are influenced by the (observed or unobserved) variables underlying the study. Under this assumption, the mixing weights a_{ij} should be proportional to the concentrations of the identified metabolites in the samples. Finally, our statistical models will search for an effect of the design variables of interest on these concentrations (quantified via the mixing weigths).

4.2 Linear mixed model specification and estimation

Let a_j be the $(n \times 1)$ vector of mixing weigths corresponding to the j^{th} ICA source and Y the $(n \times l)$ experimental design matrix. In order to specify the model to be estimated, two model matrices have to be built from Y:

- Z^1, a $(n \times p_1)$ incidence matrix containing the fixed effects of the model: typically a constant term, coded categorical design variables, continuous variables and interactions or other high-order terms.
- Z^2, a $(n \times p_2)$ incidence matrix containing the random effects of the model: typically coded random design variables as subject, batch, day and interactions between fixed and random variables.

For each of the q sources s_j, the following linear mixed model is then defined:

$$a_j = Z^1 \beta_j + Z^2 \gamma_j + \epsilon_j \tag{4}$$

where β_j is a $(p_1 \times 1)$ vector of constant parameters to be estimated, γ_j is a $(p_2 \times 1)$ vector of random effects distributed as a multivariate normal $N(0, G)$ and ϵ_j is a $(n \times 1)$ vector of residuals distributed as a multivariate normal $N(0, R)$ [14].

Different specific cases of this general model are possible according to the inclusion of both Z^1 and Z^2, only Z^1 or only Z^2 in the model. Models using only Z^1 can be separated into two categories according to the nature of the fixed effects covariates. In the case of categorical covariates, the model is an ANOVA one. In the case of quantitative covariates, a linear regression model is defined. And when both types of variables are included a (fixed) GLM model is concerned. A quantitative variable is typically the dose of a drug and a categorical variable can be different levels of a treatment (e.g. placebo versus a low and a high dose of a drug). Note that, in many medical studies, quantitative variables are often categorized before being introduced in statistical models.

Models using only Z^2 are variance components models including only random factors. This arises when one is interested by the effect of various populations (or analytical factors) on the spectrum variability (e.g. subject, operator, batch,...), but this is not yet common in metabonomics. Complex metabonomic studies will typically include both fixed and random effects as for example in longitudinal studies where n subjects belonging to p categories of treatments are followed over time. The next subsection will illustrate a simple ANOVA and regression cases on the hypothetical metabonomic data. Section 7 will illustrate the complete methodology on a more complex metabonomic dataset.

4.3 Example

We consider here the modelling step on the experimental data in the case where only fixed effects are present. In section 3.5, six ICA sources were identified from the spectral matrix and the mixing weigths gathered in a

(28×6) matrix A. The design matrix Y contains two fixed covariates: the drug dose y_1 (or hippurate) and the age of the subject y_2 (or citrate). y_1 is the covariate of interest for which biomarkers are investigated.

These variables can be introduced either as continuous or as categorical variables in the linear model. In the first case, matrix Z^1 will be, at least, a (28×3) matrix with a constant term as first column and y_1 and y_2 as second and third columns. And, for each source s_j, the following linear model is of application:

$$a_j = Z^1 \beta_j + \epsilon_j = \beta_{j0} + \beta_{j1} y_1 + \beta_{j2} y_2 + \epsilon_j \tag{5}$$

The β's estimated by linear regression for the six sources are given in Table 1. We will interpret these coefficients in the next step. Note that higher order terms (quadratic or interaction terms) could also have been taken into account in this model.

If the two covariates are introduced as categorical variables in the model, Z^1 becomes a (28×7) matrix with a constant term as first column and two blocks of three columns corresponding to the binary coding of the 4-levels categorical variables. Such model can then be estimated by regression but corresponds also to a two ways ANOVA model which can be fitted through classical ANOVA formulae when the design is balanced [15]. The model would be written in the ANOVA literature as:

$$a_{jih} = \beta_0 + \beta^i_{j1} + \beta^h_{j2} + \epsilon_{jih} \tag{6}$$

where indices i and h refer to the levels of the two variables y_1 and y_2 and β^i_{j1} and β^j_{j2} to the corresponding main effects according to source s_j. Note that one could also introduce an interaction term in this model. ANOVA model results will be provided in the next section.

5 Step III: Biomarker identification

5.1 Goal and principle

The third step of the procedure aims at finding which of the q ICA sources vary significantly in the observed spectra when values of the covariates of interest are modified. These sources or combinations of them will be considered as potential biomarkers.

Practically, the choice of the significant sources is based on the statistical significance of the terms or effects included in the mixed models estimated in Step II. The adequate statistical tests depend on the model structure and must include a multiplicity correction when the number of sources is large. This step produces r_k significant sources for each covariate or more complex effect of interest in the study. These sources form the input of the next step of the procedure.

5.2 Selection of significant sources

In general mixed models, several common procedures exist to test the significance of model terms. They are different for fixed and random effects, depend on the method applied to estimate the model and may be controversial when complex random effects occur. To keep things simple, this paper treats only some simple cases and the reader is invited to consult related literature [14] and software for general situations (e.g. PROC MIXED in SAS or *lme* function in *R*).

Let us suppose that the model contains only fixed continuous and categorical effects and that the effect of interest is the main effect of a continuous covariate y_k. The significance of y_k is derived for each source s_j through the p-value related to a t-statistic and calculated as follows:

$$p_{jk} = 2 \times P(t_{(n-p)}) \geq |t(j,k)|) \tag{7}$$

with

$$t(j,k) = \hat{\beta}_{jk}/s(\hat{\beta}_{jk}) \tag{8}$$

and where $\hat{\beta}_(jk)$ is the coefficient of y_k in the fitted model on a_j,$s(\hat{\beta}_{jk})$ is the standard error associated with $\hat{\beta}_(jk)$, n is the number of observations (spectra), p is the number of parameters into the model and $t_{(n-p)}$ is a t random variable with $(n-p)$ degrees of freedom.

If one supposes now that the effect of interest is a categorical covariate with q levels, the significance of y_k is then derived for each source s_j through a F statistic as follows:

$$p_{jk} = P(F_{q-1,n-p} \geq F(j,k) \tag{9}$$

with

$$F(j,k) = MSy_j^k/MSR_j \tag{10}$$

and where MSR_j is the mean square of model residuals for source s_j, MSy_j^k the mean square related to y_k effect and $F_{q-1,n-p}$ a F random variable with $(q-1)$ and $(n-p)$ degrees of freedom.

If such procedure is applied on K variables with more complex effects of interest in the model (and for each of the q sources), $(K \times q)$ tests are performed and the decision of significance via the p-values must take into account the multiplicity situation. If $(K \times q)$ remains reasonably small, a simple Bonferroni correction is still applicable and the significance of the effect of y_k for source s_j is confirmed if $p_{jk} \leq \alpha/(K \times q)$, where α is a chosen total error rate (e.g. α=0.05). For larger $(K \times q)$, procedures like False Discovery Rate (FDR) could be a solution [20].

5.3 Example

In the example discussed in Section 4.3, if the dose effect y_1 is the only effect of interest and is treated as continuous in the model, the p-values associated with the t-tests are given in Table 1 (second column). These p-values will be declared significant if smaller than $\alpha/6 = 0.00833$ with $\alpha = 0.05$. One can notice that four sources

are significant (three of them are extremely significant), the most significant one being source s_3 corresponding to the hippurate spectrum (as expected).

Sources	$\hat{\beta}_{j1}$	Linear Regression p-values	$F(j,1)$	ANOVA p-values
s_1	-6.6^{-7}	1.94^{-15}	105.46	8^{-13}
s_2	-5.52^{-7}	4.77^{-16}	152.71	2.04^{-14}
s_3	$2.64.6^{-6}$	8.30^{-35}	4468.90	1.31^{-29}
s_4	-1.07^{-7}	0.27	0.83	0.50
s_5	2.21^{-07}	0.004	2.86	0.06
s_6	3.70^{-9}	0.96	0.02	0.99

Table 1: Results of Linear Regression and ANOVA models.

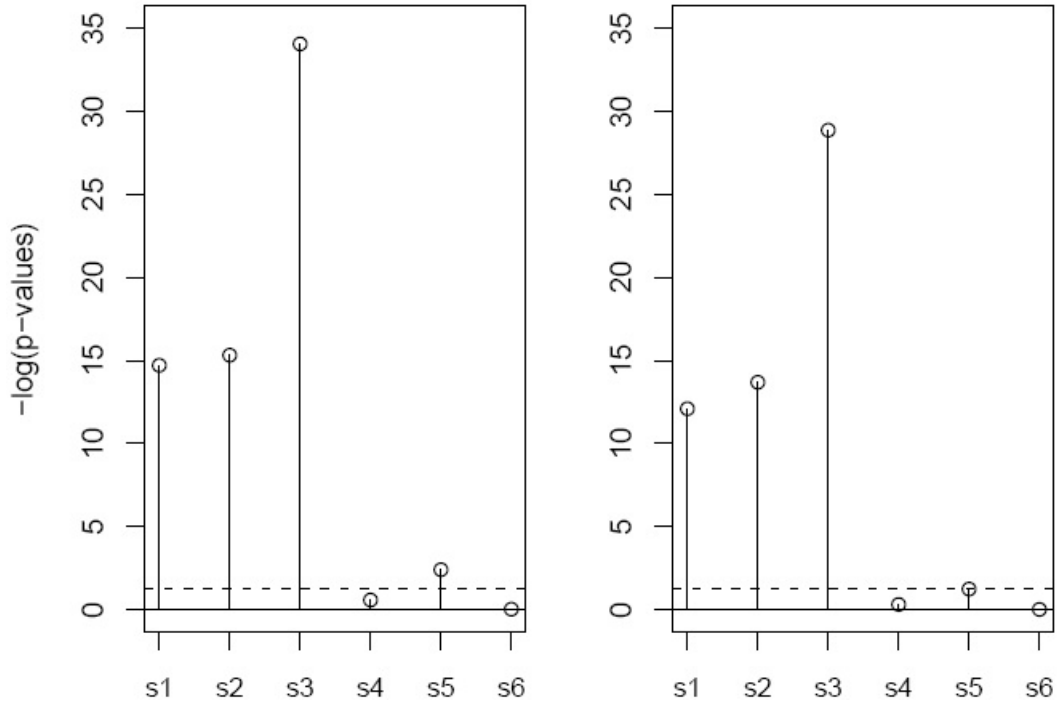

Figure 9: p-values corresponding to each sources for the regression models (left) and for the ANOVA models (right). These p-values are expressed as -log(p-value) here. Significant p-values are over the dotted line which represents the levels of significance after Bonferroni correction along with the -log transformation

If y_1 is introduced as a categorical variable in the model (see Equation 6), the two last columns of Table 1 provide the F-statistics and related p-values for the six sources and Figure 9 (right) shows that three sources are

significant, the most significant still being s_3. Regression and ANOVA approaches select then the same more significant source: s_3.

Spectral regions represented in s_3 may then represent biomarkers or spectral expression of metabolites significantly affected by a change of the factor of interest y_1. As expected in this example with y_1 being the hippurate dose, s_3 presents as biomarkers the peaks in the spectral zone of the hippurate molecule.

Additionally, both linear regression and ANOVA models select s_1 (spectral profile of pure urine) and s_2 (spectral profile of pure citrate). In linear regression models, the signs of each estimated $\hat{\beta}_{j1}$ effect of y_1 on the modelised vector of weight could also be inspected. For a selected source, a positive $\hat{\beta}_{j1}$ indicates that the contribution of this source to the observed spectra significantly increase when y_1 increases. In other words, biomarker peaks presented in this source increase when y_1 increases.

Table 1 shows that, on the contrary of $\hat{\beta}_{31}$, $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$ are negative: an increase in y_1 is followed by an increase in the spectral peaks of hippurate (regions in s_3) and by a decrease in the spectral peaks corresponding to natural urine (regions in s_1) and peaks corresponding to citrate (regions in s_2). This can be easily explained by the fact that each observed spectrum is normalized to have a sum equal to one (constant sum normalization).

Comparisons between linear regression and ANOVA p-values highlight the fact that source s_5 is only selected by linear regression. Although ANOVA analysis has the advantage to account for slightly more of the variation, the ANOVA method is evaluated on more degrees of freedom than the regression and has greater p-values. This can lead to a risk of missing significant effects as for s_5. Treating independent variable as continuous should then be the choosing method in the first instance, with ANOVA being used if regression analysis is not appropriate (e.g. if the relationship between the variables is not linear enough). Figure 10 shows that in this example the relationship between the weight vectors and y_1 can be considered as linear, and this for each of the level of the other covariate y_2.

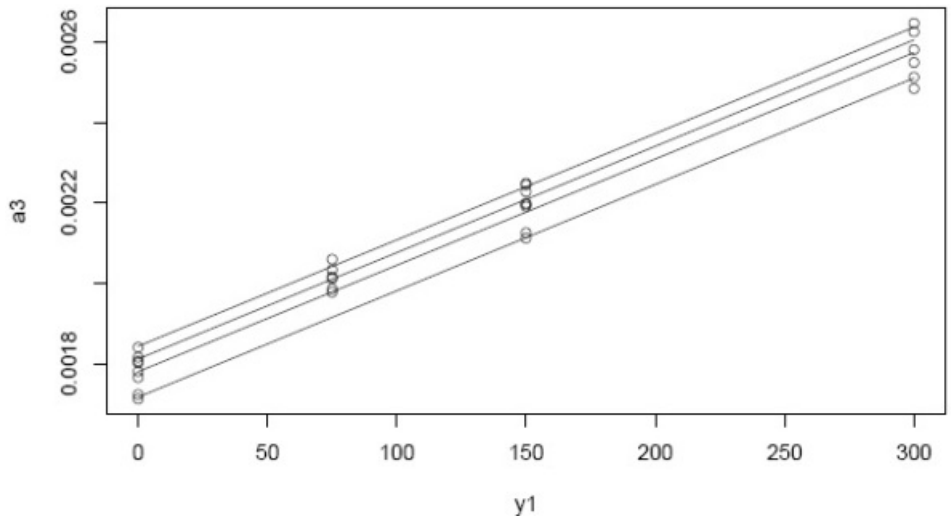

Figure 10: relationship between the hippurate dose (y_1) and the vector of mixing weights a_3. Each of the four

lines represents the estimated linear regression models for one fixed value of y_2

6 Step IV: Visualization of biomarkers and factor effects

6.1 Goal and principle

For each covariate of interest in the metabonomic study, Step III provides a list of r significant sources. Step IV proposes then a simple tool to visually interpret these sources as potential biomakers. It aims to answer the following question: which average change is expected in the spectrum when the covariate of interest changes from one level to another (e.g. if a patient is or is not affected by a disease or if the dose of a drug is increased).

6.2 Contrast calculation

Let us define S^* as the $(m \times r)$ matrix of significant sources identified in Step III. Let's then y_k^1 and y_k^2 be two levels of interest for covariate y_k (e.g. two drug doses). Let us finally define $\Delta\hat{a}^*_{2-1} = \hat{a}^*_2 - \hat{a}^*_1$ as the vector of differences of model predictions for these two covariate levels and for the r identified sources. For models without interations, these differences are only influenced by the terms in y_k. For models with interactions, the values of the other factors should be fixed to chosen levels.

Consequently, the expected change in spectra can simply be obtained via the following contrast:

$$C_{2-1} = S^* \Delta\hat{a}^* \tag{11}$$

where C_{2-1} is a $(m \times 1)$ vector and can be drawn as spectrum to visualize the spectral zones which are affected by the covariate. In particular, if y_k is introduced as a continuous variable in the model and $\hat{\beta}^*_k$ is the vector of the coefficients for y_k and the r identified sources, the expected change between the spectra at the two levels y_k^1 and y_k^2 is given by: $C_{2-1} = S^* \hat{\beta}^*_k (y_k^2 - y_k^1)$.

If y_k is introduced as a categorical variable in the model and $\hat{\beta}^{*1}_k$ and $\hat{\beta}^{*2}_k$ are the vectors of the estimated effects for the two levels of interest for the r sources, the change in spectra is provided by $C_{2-1} = S^*(\hat{\beta}^{*2}_k - \hat{\beta}^{*1}_k)$.

6.3 Example

In the design matrix Y, the hippurate dose y_1 is observed at the following values: 0, 75, 150 and 300 mg. Three contrasts C_{2-1}, C_{3-1}, C_{4-1} respectively describe the expected changes in spectra when the drug dose goes from 0 to 75 mg, 0 to 150 mg and 0 to 300 mg. Figure 11 presents the three contrasts obtained when y_1 is introduced as a continuous variable in the model, while Figure 12 covers the case where y_1 is used as a categorical variable.

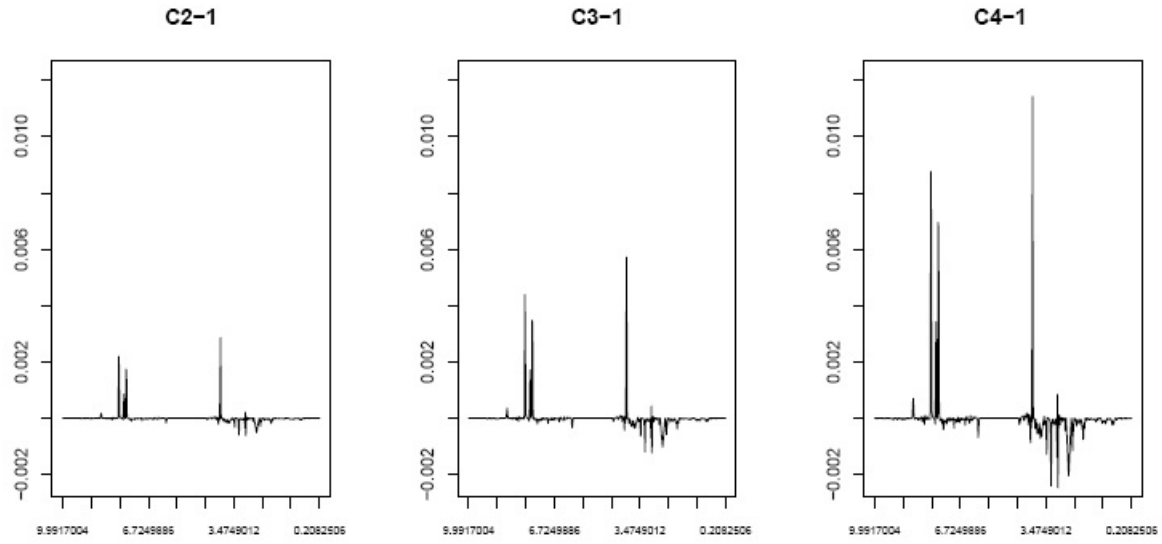

Figure 11: the three contrasts obtained when y_1 is introduced as a continuous variable

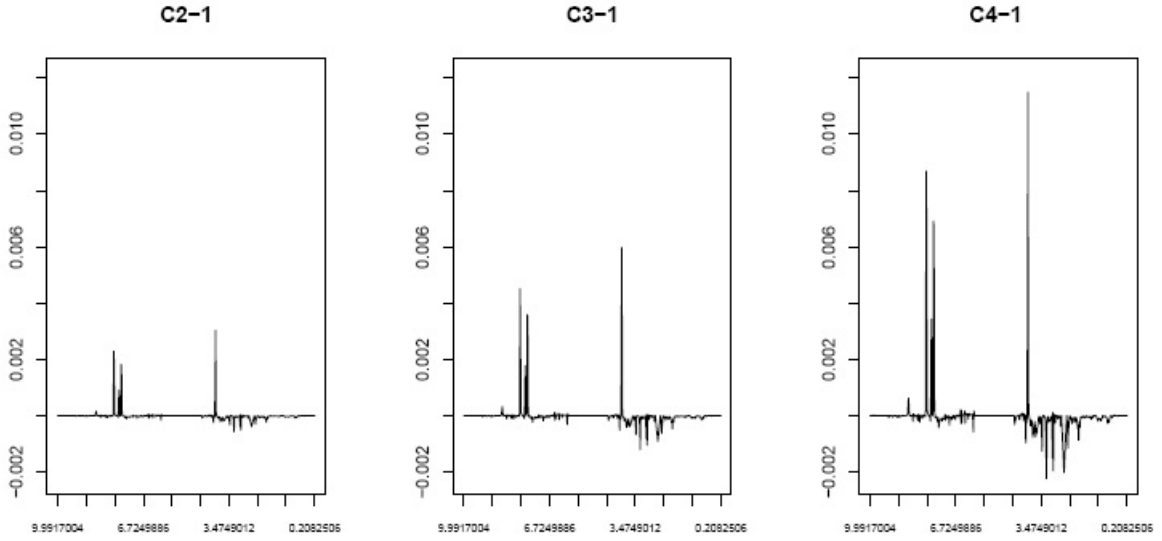

Figure 12: the three contrasts obtained when y_1 is introduced as a categorical variable

As the dose goes from 0 to a positive value in each of the three contrasts, the hippurate peaks increase. On the contrary, negative values find themselves everywhere else. It indicates that when the drug dose increases, peaks corresponding to other metabolites decrease. This can be explained by the fact that each spectrum is normalized to have a total concentration equal to one.

In addition, it is very important to see that comparisons between C_{2-1}, C_{3-1}, C_{4-1} show that when the drug dose is increased by a factor of two (75 to 150 mg, 150 to 300 mg), hippurate peaks are expected to increase in the same proportion.

Step I to IV were also applied on these data with the estimation of respectively four and height sources. The three contrasts obtained in each situation demonstrate the stability of the contrasts according to these different number of sources.

7 Application to a more complex dataset

In this section, data extracted from the experimental design described in Section 2.2.1 will allow to illustrate the methodology presented in this paper in a more complex hypothetical situation. As shown in Figure 13, nine experimental mixtures will be considered and grouped in three classes corresponding, for example, to three hypothetical disease states: group 0 corresponds to spectra from healthy subjects, group 1 to spectra from subjects with a first kind of disease, group 2 to subjects with a second kind of disease.

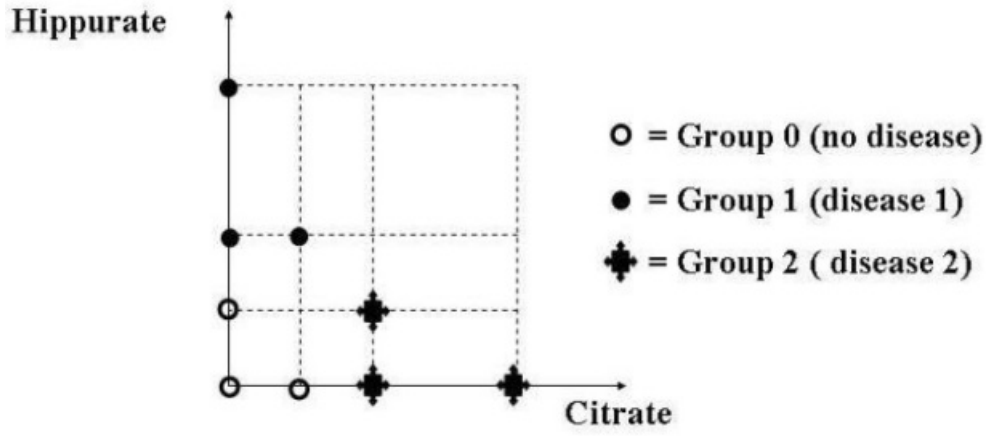

Figure 13: experimental design of a more complex dataset divided in three groups

For each experimental condition, the spiked urine samples were analyzed at eight times: as such or in diluted water (1/1), over two consecutive days and with two replicates per day per media. 72 spectra were then used and can be described in the design matrix according to the following categorical factors:

y_1: disease group (G1, G2, G3).

y_2: media (diluted or non diluted urine).

y_3: day of measurement (day1 or day2).

y_4: replicate within each day (1 or 2).

Pre-treatments described in Sections 2.2.3 and 2.2.4 were applied on the spectral data and produced a (72×600) spectral matrix X. ICA was then applied, as described in Section 3.2, in order to obtain $q = 5$ sources ($D_j = 89\%$) and the associated mixing weight vectors a_j (see Figure 14).

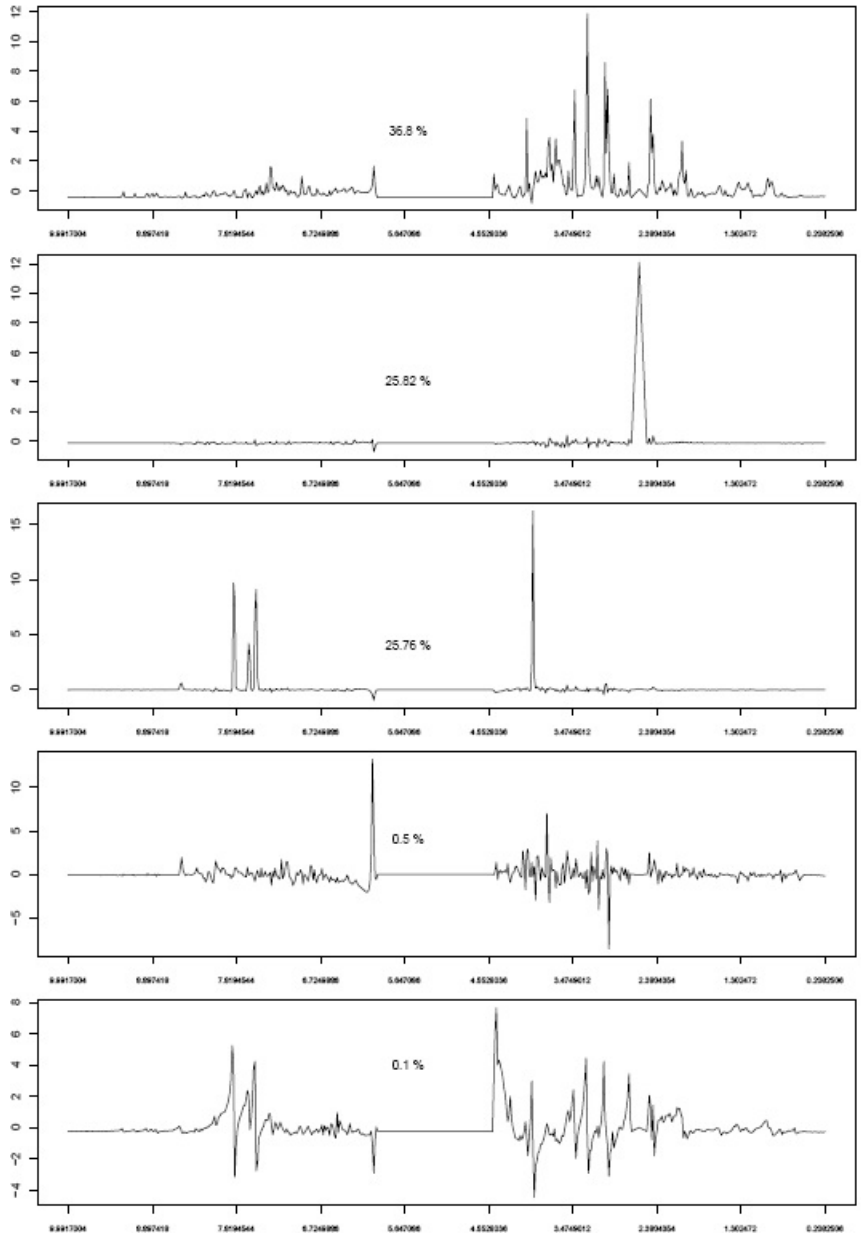

Figure 14: the $q = 5$ ICA sources for the complex data example

A mixed model was then fitted on each vector a_j with y_1, y_2, y_3 as fixed factors and with all possible interactions of first and second order. Table 2 provides the F statistics and p-values corresponding to the main effects of these three factors on the mixing weights. Interaction effects are not reported in the table because they are all not significant. The Bonferroni corrected p-value threshold was taken as $0.05/15 = 0.0033$.

This table shows that the experimental group has a very significant effect on the three first ICA sources. These sources were then used to compute three contrasts in order to illustrate the effect on the spectra when passing from one group to the other (see Figure 15). C_{1-0} presents the average expected change when a subject

Sources	$F(j,1)$	$p_{j,1}$	$F(j,2)$	$p_{j,2}$	$F(j,3)$	$p_{j,3}$
s_1	35.51	6.67^{-11}	0.09	0.77	0.01	0.92
s_2	17.55	9.98^{-7}	0.22	0.64	0.07	0.79
s_3	18.70	4.87^{-7}	0.02	0.90	0.07	0.79
s_4	0.12	0.89	269.35	1.07^{-30}	0.59	0.44
s_5	1.17	0.32	0	0.994	0.50	0.48

Table 2: Linear Regression and ANOVA models: results in a more complex situation

with no disease gets disease 1: citrate peak increases. C_{2-0} illustrates the fact that, for a healthy subject that becomes a subject with disease 2, hippurate peaks increase. Finally, C_{1-2} shows that evolving from disease 2 to disease 1 leads to a decrease of hippurate peaks and an increase of the citrate peak. All these results are of course expected.

Table 2 allows us to see that the sources do not significantly change from one day to the other (y_3) but that the media (water dilution - y_2) has a significant effect on source 4. This source is characterized by a peak on the left side of the water peak region (set to zero during pre-processing). This peak is in fact a side effect of the original water peak but the percentage of variation of the source (0.5%) shows that the different pre-processing steps used to remove media systematic effect were efficient.

None of the factors has a significant effect on source 5 but its mixing weights vector indicates that this source is mainly influenced by one outlier spectrum.

It is also interesting to compare the ICA results obtained in this example and in the example of Section 5.3. Here, 89% of the total variance of X is explained by the five first sources. In the previous example, 98% of the variation was explained by the same number of sources. This difference can be explained by the fact that replicating the measures over several days and with different dilutions introduces random noise in the data which can not be catched by ICA sources.

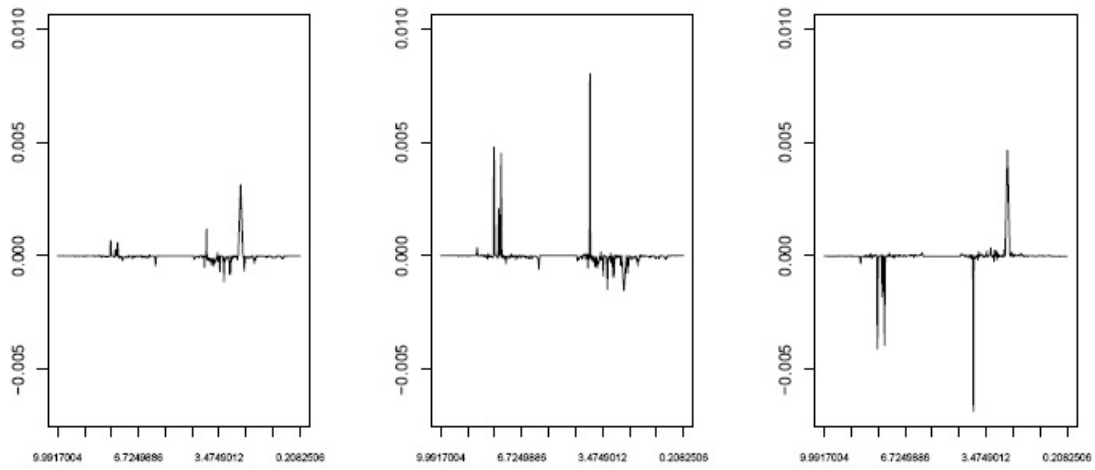

Figure 15: contrasts C_{1-0}, C_{2-0}, C_{1-2}

8 Application to real data: metabonomic study of Age related Macular Degeneration (AMD)

8.1 Goals and design

This study has been realized on serum samples due to vascular hypothesis about the AMD. Indeed, ninety percent of all vision loss due to AMD results from the exudative form, which is characterized by choroidal neovascularization, defined as newly formed blood vessels arising from choriocapillaries. Age-related changes that induce pathologic neovascularization are incompletely understood.

The goal of this research is to discover metabonomic biomarkers making the distinction between healthy and diseased subjects. These molecular biomarkers will be used to develop knowledge about the AMD pathological mechanisms. Subsequently, their corresponding spectral biomarkers could be used in the future as diagnostic tool.

The AMD study is a qualitative outcome two class problem disease metabonomic study. In this study, each spectrum is characterized by the AMD status of its corresponding subject through a binary qualitative outcome: a medical examination declares if the subject has or not the AMD.

This observational case-control study was designed according to a protocol approved by the ethical committee of the University Hospital of Liege, Belgium. Cases for the study were defined as patients of the University Hospital of Liege Belgium affected by AMD over the age of sixty. Controls are age-matched patients in the same hospital without any sign of macular disease and not having a known family history of AMD. Cases as controls were said eligible on the basis of an examination realized by a trained ophthalmologist. Informed consent was required from all study subjects before participation.

The protocol also foresaw the additional biochemical and health state information to collect on the serum samples and subjects. Additionally, AMD patients had to be categorized into active or not active phenotypic categories depending of the bleeding of lesions on the basis of the Optical Coherence Tomography (OCT) examination. A design matrix Y_A was formed during the study design step. This matrix contains the following information:

an identifiant for each blood sample providing a spectrum

the qualitative outcome describing the presence or not of AMD according to the examination (0 = yes, 1 = no).

the activity phenotypic character of AMD: 0 = AMD subject in active phase, 1 = AMD subject in inactive phase, 2 = control (no phase of activity).

In the end, the spectral data matrix X involves 193 spectra; 94 FIDs from AMD subjects and 99 FIDs from control subjects. All acquisition and pre-processing steps are then available in [6] (chap 6).

8.2 Use of ICA in the AMD study

After the detection phase of outliers, the study of the variability within the data and the implementation of several statistical methods to find biomarkers, combination of ICA with these statistical models was finally used in the objective to explore the spectral changes related to the activity state of the disease.

Stationary or evolving states of AMD is described by a variable "AMD activity". Patients in inactive phase have stationary lesions while patients in active phase have evolving lesions presenting bleeding and neovascularization. The AMD activity description is available for 75 AMD spectra, including 51 active phase and 24 inactive phase. Available values of this variable allow to list 171 of the spectra in the three following categories: AMD active phase (51 spectra), AMD inactive phase (24 spectra), no phase or control (96 spectra). The ICA with statistical models were applied on the basis of eighteen sources previously recovered (for details, see [6]) and the mixing weights of the 171 spectra. An ANOVA model involving one fixed factor, the "AMD activity", was fitted on each weight vector corresponding to one of the the q = 18 recovered sources.

The phase of activity has a significant effect on fourteen sources with a Bonferroni corrected p-value threshold taken as $0.05/18 = 0.0028$. These fourteen sources were then used to compute the contrasts presented in Figure 16. The first graph represents the changes occurring in the spectra when an healthy or control subject becomes an AMD subject in inactive phase. The second graph is the contrast computed between AMD subject in active phase and control. Last graph shows the changes occurring in the spectra when the stationary or inactive AMD disease turns to an active or evolving disease.

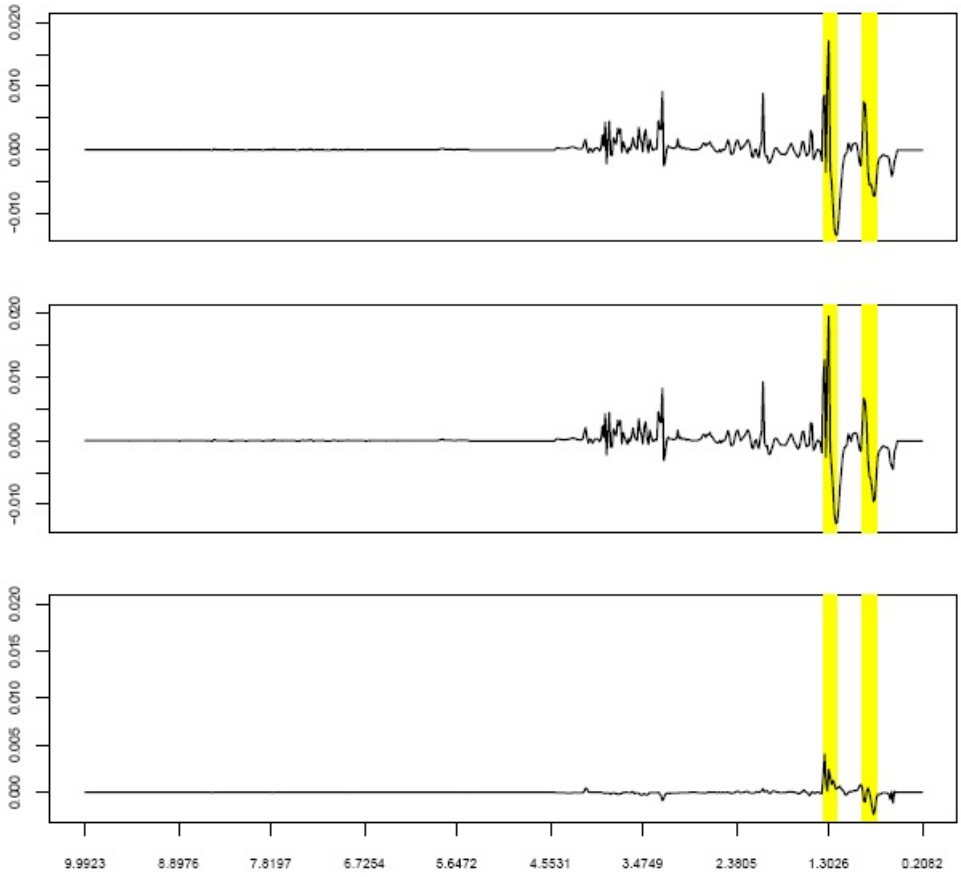

Figure 16: contrasts in the context of the AMD study: control-inactive (top), control-active (middle), active-inactive (bottom). Colored spectral zones represent the lactate and lipoproteins zones.

8.3 Molecular interpretation

Spectroscopists have identified the metabolites corresponding to proposed spectral biomarkers. Among these lists, they found descriptors corresponding to two metabolites of interest, the lipoproteins (LDL, VLDL, HDL) and the lactate. The discovery of lipoproteins as biomarker supports previous published biological hypothesis about their role in the onset of AMD [21] . Lactate as biomarker has generated from the ULg researchers a theory about AMD pathological mechanisms involving an increase of lactate.

The evolution of the lipoproteins and lactate spectral zones in the different states of the disease is viewable in the ICA contrasts (Figure 16), indicated by the yellow zones. The lactate increases with the emergence of the disease as in the transition from an inactive to an active state of the disease. On the opposite, the lipoproteins decrease in this situation. It is shown that the most informative discovered predictors belong to these zones [6].

9 Conclusions

Metabonomics is emerging as a valuable tool in a number of biological applications. Although, the biomarker identification in ^{1}H-NMR based metabonomics is traditionally realised, with some limitations, via the examination of the two first components of a PCA. In this paper, we presented a four steps methodology providing three kinds of knowledge on ^{1}H-NMR metabonomic data: the identification of biomarkers, a statistical confirmation of the significance of these biomarkers and the visualization of the effects on the biomarkers caused by factor changes.

The methodology involves a dimension reduction by ICA followed by statistical modelling approaches. We first presented a process to decompose by ICA the spectral data into statistically independent components. We exposed on experimental data that ICA allows to visualize, through the resulting sources, the spectral profile of independent metabolites contained in the studied biofluid and their quantity through the corresponding mixing weights. From Steps II and III, various linear mixed statistical models were applied on ICA results to select the sources. These sources present spectral regions changing significantly according to the factors of interest. In the final step, the selected sources were used to reconstruct the spectra and to compute contrasts presenting the alterations in specific regions caused by different changes of the factor of interest.

As exposed on experimental data, the ICA solves the weaknesses of the PCA dimension reduction by providing more natural and also more biologically meaningful representations of the data. Additionally, the combination of ICA with statistical models has the advantage to base the component selection on an inferential criterion: biomarkers are identified from components for which the covariate of interest shows a significant effect. In the usual PCA, biomarkers are identified from the component with the largest percentage of variance, without any inferential information.

In this paper, source selection was based on t-statistics computed on the weight vectors without using their significance levels. We also provided a more accurate source selection due to its inferential character but also to the fact that models give the possibility to include all the design covariates jointly with the covariate of interest. The large diversity of statistical models accepted by this methodology allows to apply it to a large variety of more complex metabonomic situations: models can include quantitative and qualitative design variables as well as combinations of fixed and random effects (linear mixed models). As a result, additionally to the proposed biomarker search, the methodology provides information on spectral regions affected by other factors of the study.

Finally, the methodology goes further than the usual search for metabonomic biomarkers: beside their discovery, contrasts also allow to visualize the alterations of potential biomarkers for defined changes of covariate conditions or context. The methodology has been also applied on a real metabonomic dataset (AMD). The spectral biomarkers linked with this disease correspond to a metabolite supporting biological explanation of the setting of AMD.

10 Acknowledgements

The authors are grateful to the Centre Intrafacultaire de Recherche du Médicament, Laboratoire de Pharmacognosie et de Chimie Pharmaceutique ULg, especially Dr. P. De Tullio and Dr. M. Frederich for providing data.

References

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

[2] Trygg J., Holmes E., Lundstedt T., Chemometrics in Metabonomics, Journal of Proteome Research 6,(2007),469-479.

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

[4] Halouska S., Powers R., Negative impact of noise on the principal components analysis of NMR data, J.Magn.Reson 178 (2006), 88-95.

[5] Rousseau R., Govaerts B., Verleysen M., Boulanger B., Comparison of some chemometric tools for metabonomics biomarker identification, Chemometrics and Intelligent Laboratory Systems 91(2008), 54-66.

[6] Rousseau R., Statistical contribution to the analysis of metabonomic data in ^{1}H-NMR spectroscopy, Thesis, UCL, 2011

[7] Hyvärinen A, Oja E.,Independent Component Analysis: algorithms and applications. Neural Networks 13 (2000), 411-430.

[8] Hyvřinen A. , Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Network 10 (1999), 626-634.

[9] Liebermeister W., Linear modes of gene expression determined by independent component analysis 18 (2002), 51-60.

[10] Lee S., Batzoglou S., Application of independent component analysis to microarrays, Genome Biology 4 (2003), R76.1-R76.21.

[11] Scholz M., Gatzek S., Sterling A., Fiehn O., Selbig J., Metabolite fingerprint: detecting biological features by independent component analysis 15 (2004), 2447-2454.

[12] Common P. , Independent Component Analysis, a New Concept?, Signal Processing 36 (1994), 287-314.

[13] Hubber P. , Projection pursuit, The Annals of Statistics13 (1985), 435-475.

[14] Brown H., Prescott R., Applied Mixed Models in Medicine, John Wiley and Sons, 2006.

[15] Searle S.R., Henderson C.R., Amer J.. Statist. Assoc. 74 (1979), 465.

[16] Eilers P.H.C., Baseline correction with asymmetric least squares smoothing, DP of Department of Medical Statistics, Leiden University Medical, 2005.

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

[18] Laird N.M., Ware J.H., Ramdom effects models for longitudinal data, Biometrics 38 (1982), 963-974.

[19] Pinheiro J.C., Mixed-effects models in S and S-plus, Springer, New York.

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

[21] Nowak M., Changes in lipid metabolism in women with age-related macular degeneration. Clinical and Experimental Medicine, 4:183 - 7, 2005.