Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Féraud et al., J Biom Biostat 2017, 8:4
DOI: 10.4172/2155-6180.1000367

Review Article Open Access

Independent Component Analysis and Statistical Modelling for the Identification of Metabolomics Biomarkers in ^{1}H-NMR Spectroscopy

Baptiste Féraud[1,2], Réjane Rousseau[3], Pascal de Tullio[4], Michel Verleysen[2] and Bernadette Govaerts[1]*

[1]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), UCL, Belgium
[2]Machine Learning Group, UCL, Belgium
[3]Arlenda S.A., Belgium
[4]Center for Interdisciplinary Research on Medicines (CIRM), Université de Liège (ULg), Belgium

Abstract

In order to maintain life, living organism's product and transform small molecules called metabolites. Metabolomics aims at studying the development of biological reactions resulting from a contact with a physio-pathological stimulus, through these metabolites. The ^{1}H-NMR spectroscopy is widely used to graphically describe a metabolite 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 pro-cess 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 two main steps providing knowledge on specific ^{1}H-NMR spectral areas via the identification of biomarkers and via the visualization of the effects caused by some external changes. The first step implies Independent Component Analysis (ICA) in order to decompose the spectral data into statistically independent components or sources of information. The in-dependent (pure or composite) metabolites contained in bio fluids are discovered through the sources, and their quantities through mixing weights. Specific questions related to ICA like the choice of the number of components and their ordering are discussed. The second step consists in a statistical modelling of the ICA mixing weights and introduces statistical hypothesis tests on the parameters of the estimated models, with the objective of selecting sources which present biomarkers (or significantly fluctuating spectral regions). Statistical models are considered here for their adaptability to different possible kinds of data or contexts. A computation of contrasts which can lead to the visualization of changes on spectra caused by changes of the factor of interest is also proposed. This methodology is innovative because multi-factors studies (via the use of mixed models) and statistical confirmations of the factors effects are allowed together.

Keywords: Metabolomics; Multivariate statistics; Independent component analysis; Biomarker identification; ^{1}H-NMR spectroscopy; Linear mixed models

Introduction

In a metabolomics context, proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy generates spectral profiles describing the metabolite composition of collected bio fluid samples. A comparison of several spectra of metabolites in various specific states permits a preliminary graphical and qualitative investigation of changes in bio fluid metabolite composition inherent to the presence of a stressor. However, the complexity of ^{1}H-NMR spectra and the number of spectra (of samples) usually available in metabolomics studies require a semi-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 the spectra obtained in various conditions [1-5]. 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, also provides a measure of statistical significance of the identified biomarkers.

The first and the most common chemo metric tool used in preliminary metabolomics studies is Principal Component Analysis (PCA). PCA is a starting point for analysing multivariate data and can rapidly provide an overview of the hidden information. PCA produces a two-dimensional plot (score plot) where the coordinate axes correspond to the two first principal components [6]. If spectra differ according to a specific characteristic (presence or absence of a stress, for example), the score plot 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 clusters.

However, variations within groups are sometimes larger than variations between groups, resulting in a score plot with clusters that overlap or do not directly correlate to the studied characteristics. In such cases, additional information can be extracted by using more advanced 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 [7,8]. Another limitation of PCA is its high sensitivity to noise for the analysis of ^{1}H-NMR data: very small and random fluctuations within

***Corresponding author:** Govaerts B, Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), UCL, Belgium, Tel: +32 (0) 10 47 43 38; E-mail: baptiste.feraud@uclouvain.be

Received March 25, 2017; **Accepted** July 12, 2017; **Published** August 23, 2017

Citation: Féraud B, Rousseau R, de Tullio P, Verleysen M, Govaerts B (2017) Independent Component Analysis and Statistical Modelling for the Identification of Metabolomics Biomarkers in ^{1}H-NMR Spectroscopy. J Biom Biostat 8: 367. doi: 10.4172/2155-6180.1000367

noise of the ^{1}H-NMR spectrum can result in irrelevant clusters in the score plot formed by the two first principal components.

Despite these limitations, PCA often remains the main statistical standard for the analysis of ^{1}H-NMR data. In a previous work, Rousseau et al. extend the standard PCA methodology by selecting the two most discriminants factors for the score plot (instead of using systematically the two first ones), and by using statistical methods for the identification of biomarkers [9-11]. It is suggested to use PLS-DA or ICA (Independent Component Analysis) for the decomposition of spectra, resulting in improvements for the identification of biomarkers (in comparison to PCA).

Motivated by these previous results, and by the good results obtained with ICA in close domains such as genomics and Mass Spectroscopy metabolomics, this paper expands the use of ICA to the identification of specific ^{1}H-NMR spectral regions that are discriminant for two or more categories of spectra. PCA and ICA share common properties. 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. However, the ICA components have a more stringent nature than principal components: PCA decomposes data into uncorrelated components when ICA decomposes them into independent ones; independence is a stronger statistical concept than un-correlation for non-Gaussian data. Independence of the components is also adequate for biological interpretation because the analysed bio fluid (e.g. plasma, urine) can be seen as a mixture of unrelated metabolites. ^{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 two-steps methodology for the identification of ^{1}H-NMR metabolomics biomarkers. Having introduced a typical experimental dataset, used to illustrate the methodology throughout this paper, the first step consists in the implementation of ICA in order to reduce the dimension and decompose the multivariate spectral dataset into statistically independent components. Solutions are proposed to select the optimal number of components and to rank them by importance. The second step of this methodology consists in a statistical modelling of the ICA resulting mixing weights [12-15]. A panel of various mixed linear statistical models adapted to the nature of the domain are considered. The model coefficients and appropriate statistical tests are used to decide which ICA sources can be considered as biomarkers of the stressor(s) of interest, including a visualization of the effect of the latter on the ^{1}H-NMR spectra. In addition, contrasts are computed from the selected sources to visualize the spectral effects when one factor of interest changes. Finally, available in the Supporting Information, the methodology has been used on real medical data to successfully find biomarkers for Age related Macular Degeneration (AMD).

Materials

A simple set of metabolomics data is used as running example to illustrate the methodology detailed in this paper. This section details this dataset, including the acquisition steps.

Typical metabolomics data

A typical experimental metabolomics database is formed by three sets of data: a design, a set of ^{1}H-NMR spectra and biological and/or histopathological data. The design describes the experimental conditions underlying each available spectrum. Typical design factors are: subject ID (animal or human) and its characteristics, treatment, dose and time of sampling. A ^{1}H-NMR dataset contains the spectral evaluations of bio fluid samples which were collected according to the design. A primary data reduction ("binning") is carried out by digitizing the one-dimensional spectrum into a series of typically 250 to 3000 integrated regions or descriptor variables. However, a typical metabolomics 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 several descriptors. Moreover, 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 unavoidable in spectral data. Finally, each spectrum in the ^{1}H-NMR dataset is also usually linked with one or more variable(s), which tends to confirm the presence of a response of the organism to the stressor. This confirmation is obtained via the current gold-standard examinations (biological measures or histopathological ones) on the subject for which spectra are measured.

Experimental data

Experimental data were produced according to a specific design, in order to provide a database in which one controls the alterations of known descriptors. The next sections detail the design, the acquisition and the pre-processing steps on data. A more detailed description and analysis of these data is available.

Experimental design: Homogeneous urine samples were spiked with two products at different levels of concentration and analysed through spectroscopy [16-18]. The products are citric acid ("citrate") and hippuric acid ("hippu-rate"). They were added to urine at four levels of concentrations, respectively 0, 2, 4 and 8 mM for citric acid (Qc=8 mM), and 0, 1, 2 and 4 mM for hippuric acid (Qh=4 mM). The resulting 14 points design is illustrated in Figure 1.

As shown in Figure 2, the peaks corresponding to each product are located in distant areas. The hippurate is characterized by three peaks, with two of them in region containing a low level of noise (around 7 ppm). On the contrary, citrate peaks are located in the noisy region (around 2 ppm). Note that during the spectral pre-processing these peaks are aggregated in a single one to avoid alignment problems.

Sample preparation and acquisition of the ^{1}H-NMR data: The two products (citrate and hippurate) were first mixed with phosphate buffer containing TSP (Trisodium Phosphate). The volume of buffer

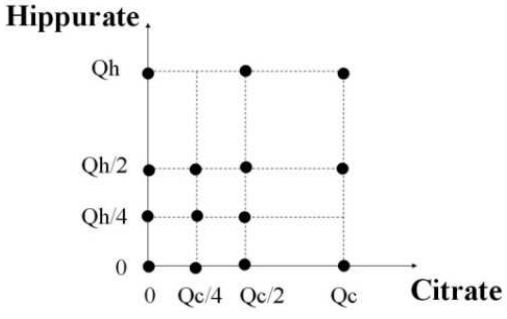

Figure 1: Experimental design.

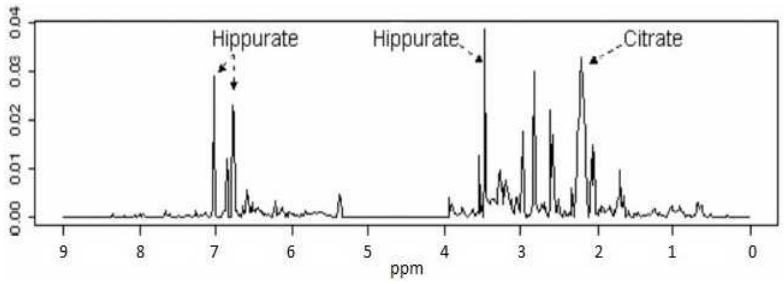

Figure 2: A Typical urine intensity spectrum with spiked citrate and hippurate.

was adapted in order to obtain a volume of 600 ml. Each urine sample came from a pool of 344 female Fischer rats and had a volume of 1200 ml. Each mixture of TSP, citrate and hippurate was added to a urine sample, centrifuged, frozen at-80°C and unfrozen at 40°C the day before the ^{1}H-NMR analysis. Each of the 14 mixtures was splitted into two parts; one was diluted on a 1-to-1 basis with water. The 28 resulting samples were then analyzed randomly within a single day of measurements [19]. NMR measurements were made with a 600 MHz Bruker spectrometer with 4 mm FI-SEI ATM probe. The spectral information is then included in 28 individual free induction decays (FIDs).

The post-acquisition treatments: Each acquired spectrum was processed using Bubble, a MATLAB tool for automatic processing and reducing NMR spectra. Bubble performs sequentially: suppression of the water resonance, apodisation (with a line broadening factor of 1 Hz), Fourier transform and phase correction, baseline correction using a Whithaker smoother, 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 the spectra prepared by Bubble [20]. 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 large shifts. Finally, we normalized again 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 constant sum normalization: each spectrum is divided by the sum of intensities on all its ppms values.

Notations

Let X be the (m n) matrix of spectral data containing n spectra, each of them being described by m descriptors. Y is a (n l) matrix of design data describing each sample or spectrum by l variables. In our experimental data, n=28, m=600 and the l=2 design variables correspond to the citrate and hippurate concentrations. This dataset is used for illustrating the methodology developed in the following sections. Some of the steps are illustrated on a 24-spectra dataset only. The latter results from removing 4 spectra from the original dataset, corresponding to the two replicates (with and without water dilution) of the samples with maximum concentration of hippurate and without citrate, and vice-versa. The 24-spectra dataset has the advantage to correspond to a non-orthogonal design of experiments and will allow to emphasize the differences between PCA and ICA results.

Methods

Dimension reduction and signal decomposition by Independent Component Analysis and biomarkers identification by statistical modelling of the ICA weights.

The basic idea of Independent Component Analysis (ICA) is to recover unobserved multidimensional independent signals from linearly mixed observed ones [21]. In the metabolomics context, it is used to extract metabolite profiles of potential biomarkers from available ^{1}H-NMR spectra.

The ICA methodology

ICA was originally developed for signal processing to solve the problem of blind source separation (BSS). In the basic noiseless ICA model, each observed signal is a mixture of unknown statistically independent signals (named sources or components):

$$X=SA^{T} \quad (1)$$

Where X denotes the (m n) matrix that contains n original signal vectors of m observations (x_i), S denotes the (m q) matrix that contains q unknown source vectors s_j, and A is a mixing matrix. Both S and A are unknown. The "unmixing" problem considered by ICA is to find an unmixing matrix such that the sources can be estimated by $\hat{S} = XW$, where $\hat{S}$ denotes the matrix formed by q estimations of scaled independent source vectors s_j (as columns). The ICA model introduces an undetermination in the scale of the recovered sources. Indeed, scaling a source by a factor l is exactly compensated by dividing the corresponding column of the mixing matrix by l. 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 are 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 that maximizes the non-gaussianity of the estimated sources, under the constraint that their variances are constant. The non-gaussianity may be estimated by the negentropy, as in the FastICA algorithm used in this work. Other ways of estimating the sources exist. Often, data are pre-processed before applying ICA. First, mixtures are reduced to zero mean without loss of generality. The second steps consist in 'whitening', i.e. applying PCA. This reduces roughly by half the number of parameters to be estimated by ICA, therefore facilitating the task of the latter. In addition using PCA allows us to reduce the number of mixtures to be used by ICA; the number q of sources to be computed can be fixed in this step via a method discussed.

ICA on metabolomic data and algorithm application: In the context of metabolomics ^{1}H-NMR data, the analyzed biofluid (e.g. plasma, urine) can be seen as a mixture of individual metabolites; 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. The Fast ICA algorithm recovers sources and linked weights from the spectral matrix through following steps:

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

$$X^C = X - 1_m.\widetilde{X}$$

Where $\widetilde{X}$ is the 1 × n vector of spectral means and 1 m a m 1 unit vector.

- Pre-processing step 2 ("whitening"): reduce by PCA the (m × n) matrix X^c to a (m × q) matrix of scores T (q £ min(n; m)):

X^c=T*P*=T P + E.

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

- Extraction of S and A^T from T with the fastICA algorithm

The (m × 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 (n q) matrix. Each column a_j is then a (n × 1) vector containing the weights (or contributions) of the corresponding source s_j in 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 $|a_{ij}|$.

Choice of the number of sources to estimate: One important parameter that may influence the ICA results is the number q of estimated components. The effective number of independent sources contributing to the signal is obviously un-known. ICA algorithms make the fundamental assumption that the number of sources q is less than or equal to the number of observed mixtures n. Moreover, to make the implementation of the fast ICA algorithm effective, the maximal value for q is the smallest dimension of its input matrix T, i.e. q min (n, m). In ^{1}H-NMR metabolomics datasets, the resolution of a spectrum m is typically higher than the number of spectra n. The maximal value for q is then the number n of observed spectra. Anyway, when n is large, choosing q=n can produce convergence problems or very high computational costs. Choosing q<n by discarding some score vectors obtained via the whitening matrix T helps convergence and discards noise. PCA provides a natural ordering of the columns of T* according to the eigenvalues l_j of X^{cT} X^c. The q first score vectors associated with the largest eigenvalues are then selected to form the reduced matrix T. Let us define D_q the proportion of the variation of X^c explained by the first q principal components:

$$D_q = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \quad (2)$$

We propose to choose q on the basis of a scree plot in order to guarantee the preservation of most of the original information.

Measure of the information contained in ICA sources and sources ordering: ICA does not provide a natural ordering of the computed sources. This section presents a possible solution to this limitation. Given a set of q estimated sources s_j, we can reconstruct the data as $\hat{X} = SA^T$. Let us define the error when X^c is reconstructed with source s j only:

$$E_j = \left(X^c - s_j.a^T{}_j\right) = S_{\neq j} A^T{}_{\neq j}.$$

This error is equivalent to the data reconstructed with all the other sources contained in the (m (q1)) matrix $S_{6=j}$. 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:

$$R_j^2 = 1 - \frac{tr(\mathrm{E}_j^T \mathrm{E}_j)}{tr\left(\hat{X}^T \hat{X}\right)} = \frac{\sum_{i=1}^{n} a_{ij}^2}{TR(\mathrm{A}^T\mathrm{A})}$$

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

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

With D_q the proportion of variance explained by the q scores in T (see eqn. (4)). This measure of importance finally allows ordering the sources s_j according to their respective C_j.

Example: In this section, the ICA procedure is applied on the n=24 spectra with m=600 ppms dataset described. Figure 3 shows the expected improvement of ICA over PCA for this specific dataset: PCA will produce principal components of maximum variance, while ICA should provide independent directions, corresponding to the sources of interest. As the experimental samples are mixtures of three products, we ideally expected to find three independent sources of variation: the urine, the citrate and the hippurate.

Of course, in the data analysis, it is supposed that we do not have the information on the sources and expect to recover them blindly according to our methodology. Based on the screeplot (Figure 4), we first chose to estimate q=6 sources. The percentage of explained variance with these first six PCs is D_6=0.9796.

A discussion about the three more important ICA sources derived by the ICA algorithm and the first three sources (or loadings vectors) obtained by applying classical PCA to the spectral matrix are detailed in the Supporting Information additional figure file.

Biomarkers identification by statistical modelling of the ICA weights

The second step of the methodology fits a statistical model in order to identify metabolomics 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 metabolomics study. This modelling step will provide a list of sources which significantly influence the spectra when the level of a factor of interest changes [22]. The profiles of these sources will then help the biologist to identify corresponding metabolites and designate them as candidate biomarkers.

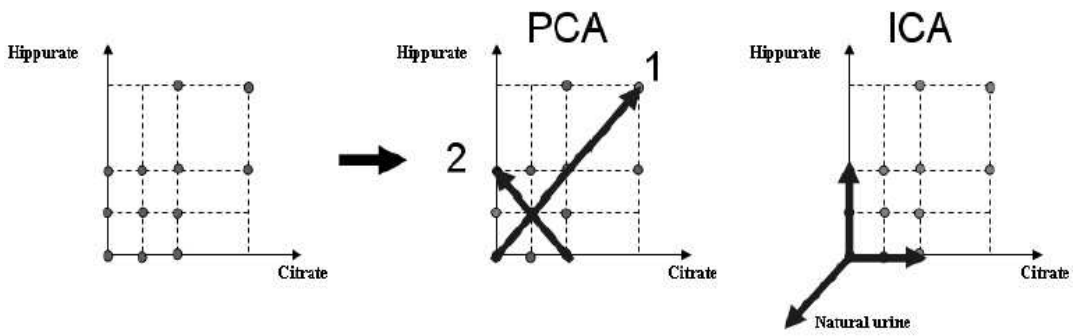

Figure 3: Components expected to be found by PCA and ICA on an illustrative experimental design.

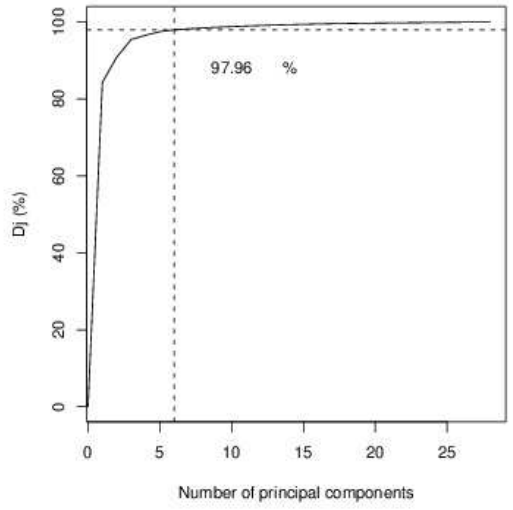

Figure 4: Scree plot of the % of explained variance with the first q PCs (from the PCA whitening).

Principle: The fundamental principle underlying this step of the methodology is the following. A ^{1}H-NMR spectrum reflects the concentrations of pure or complex metabolites contained in the analysed sample. The design factors, as for example the dose of a 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.

This step aims then at finding, through the mixing weights, which sources affect significantly the spectra when the factor of interest of the study changes (e.g. presence/absence of a disease, dose of a drug...) in spectral matrices where several other noise or controlled factors potentially affect the spectra (e.g. subject, nutrition status,...). Linear mixed statistical models are used in this context to (1) allow to decorrelate the effect of the factor of interest on the mixing weights from the effects of the other covariables and noise factors, (2) allow to take into account the random character of some covariables and (3) provide a measure of statistical significances of the link between the factor of interest and the sources.

Linear mixed model specification, estimation, testing and interpretation: Let a_j be the (n 1) vector of mixing weights corresponding to the j^{th} ICA source and Y the (n l) experimental design matrix containing the variables of the study (as the factor of interest for which biomarkers are searched for) and other covariables which may have affected the spectra. In order to find how these variables are linked to each vector of weights a_j, fitting a linear mixed statistical model is a flexible solution and a very classical approach in the context of biomedical studies. For each of the q sources s_j, the following linear mixed model can be written as follows:

$$a_j = Z^1\beta_j + Z^2\gamma_j + \varepsilon_j \,. \tag{3}$$

Where,

Z^1 is 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.

β_j, a $(p_1\ 1)$ 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 e_j is a $(n \times 1)$ vector of residuals distributed as a multivariate normal $N(0 \times R)$.

Different specific cases of this general model are possible according to the inclusion of Z^1 or Z^2 or both. Models using only Z^1 are also called GLM models in the statistical literature, and include ANOVA and regression models depending of the categorical or continuous nature of the variables included in the model. In most cases, the variable of interest of the study will be included in this matrix as a dose of a drug or a treatment versus placebo or the presence/absence of a disease. It can also include covariables that are not directly of interest but may greatly affect the spectra as the age or sex of a patient. 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, hospital, operator, batch…), but this is not yet common in metabolomics. Complex metabolomics 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.

Depending of the generality of the specified model, the estimated parameters and related significance measures will be provided by basic statistical softwares or by more advanced ones like the PROC MIXED procedure in SAS or *lme* function in R.

Testing the significance of the factor(s) of interest is a key step in the methodology and is typically neglected in most metabolomics studies. It will allow more generalized and powerful conclusions to the population of interest from which the data are issued. 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.

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:

$$pjk = 2 \times P(t_{(n-p)}) \geq |t_{(j,k)}|) \quad (4)$$

$$t(j,k) = \hat{\beta}_{jk} / s\left(\hat{\beta}_{jk}\right) \quad (5)$$

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 b and $t_{(n-p)}$ is a t random variable with (n-p) degrees of freedom.

If one supposes 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:

If one supposes 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

$$p_{jk}=P(F_{q-1,n-p}) \geq F(j,k) \quad (6)$$

With,

$$F(j, k)=MSy^k_j/MSR_j, \quad (7)$$

And where MSR_j is the mean square of model residuals for source s_j, MSy^k_j 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 (and for each of the q sources), (K × q) tests are performed and the decision of significance via the p-values must take into account the multiplicity situation. If (K × 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}£/(K × q), where a is a chosen total error rate (e.g. a=0.05). For larger (K × q), procedures like False Discovery Rate (FDR) can be used. Through these testing procedures, the modelling step can be summarized into a table containing for each mixing weight vector (and related source) a measure of significance for each factor of interest (Table 1). This result is the basis of biomarker identification and interpretation.

Let us define S^* as the (m × r) matrix of the r significant sources identified for the (or a) factor of interest in the study. A first way to extract biomarkers from these sources consists in examining their profile and identifying the known pure or complex metabolites with close profiles. This approach is appropriate when r is quite small and has "clean fingerprints". Also, sources do not provide quantification or a direction of the metabolite effect.

An additional and more informative approach is then proposed. It is a generalization of the concept of contrast estimation in classical linear models and gives an answer to 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)?

Let us introduce y^1_k and y^2_k, two levels of interest for a quantitative covariate y_k (e.g. two drug doses). Let us then define $\Delta\hat{a}^*_{2-1} = \hat{a}^*_{2-}\hat{a}^*_1$ as the vector of the differences of predictions for these two covariate levels and for the r identified sources. For models without interaction, 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:

$$C_{2-1} = S^* \Delta \hat{a}^* \quad (8)$$

Where C_{2-1} is a (m × 1) vector and can be drawn as a 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 if $\hat{\beta}^*_k$ is the vector of the coefficients for y_k and the r identified sources, the expected change between the spectra for the two levels y^1_k and y^2_k is given by $C_{2-1} = S^* \hat{\beta}^*_k \left(y^2_k - y^1_k\right)$. If y_k is introduced as a categorical variable in the model and if $\hat{\beta}^{*1}_k$ 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 $C_{2-1} = S^* \hat{\beta}^*_k \left(\hat{\beta}^{*2}_k - \hat{\beta}^{*1}_k\right)$.

Example: This section illustrates the modelling step on the experimental data presented. All 28 spectra are used in this section and the two design factors (hippurate and citrate levels) are used as variables. With 28 spectra, the screeplot suggests to calculate six ICA sources. The profiles of the three first sources are very similar than those obtained with the reduced design. Let us define y_1 as the hippurate dose and take it as the factor of interest for which biomarkers are investigated, and y_2 as the citrate level supposed to be an additional covariate in the study.

These variables can be introduced either as continuous or as categorical variables in the linear model. In the first case, matrix Z will be a (28 3) matrix with a constant term as first column and y_1 and y_2 as second and third columns. For each source s_j, the following linear model is written as:

$$a_j = Z^1\beta_j + \varepsilon_j = \beta_{j0} + \beta_{j1}y_1 + \beta_{j2}y_2 + \varepsilon_j \quad (9)$$

Sources	$\hat{\beta}_{j1}$	Linear Regression p-values	*F(j,1)*	ANOVA p-values
S_1	-6.60e^{-7}	1.94e^{-15}	105.46	8e^{-13}
S_2	-5.52e^{-7}	4.77e^{-16}	152.71	2.04e^{-14}
S_3	2.65e^{-6}	8.30e^{-35}	4468.90	1.31e^{-29}
S_4	-1.07e^{-7}	0.27	0.83	0.50
S_5	2.21e^{-7}	0.01	2.86	0.06
S_6	3.70e^{-9}	0.96	0.02	0.99

Table 1: Results of Linear Regression and ANOVA models.

The β_1's estimated by linear regression for the six sources and the corresponding p-values are given in Table 1 (the β_2's are not provided since this covariate is not considered of interest). Note that higher order terms (quadratic or interaction terms) could be included in such model. If the two covariates are introduced as categorical variables in the model, Z 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. This model is written in the ANOVA literature as:

$$a_{jih} = \beta_0 + \beta_j^i 1 + \beta_j^h 2 + \varepsilon_{jih} \quad (10)$$

Where indices i and h refer to the levels of the two variables y_1 and y_2, and β_{j1}^{i} and β_{j2}^{i} to the corresponding main effects according to source s_j. Note that one could also introduce an interaction term in this model. Table 1 provides the F statistic and corresponding p-values for the effect of the first factor on the six sources. If y_1 is considered as the only variable of interest, a p-value will be declared significant if smaller than a/6=0:00833 with a= 0:05 according to the Bonferroni correction.

Four sources can then be declared as significant in the regression model and three sources in the ANOVA model. The most important source seems to be s_3. Spectral regions linked with 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. Logically, the model recovers that a change of concentration of hippurate in the mixture introduces a signal corresponding to the third source in the resulting spectra. However, when other covariables affect the spectra, note that the methodology presented here is able to extract from the signal the effect of the variable of interest and keep in other possible non-orthogonal sources of variability. This is a crucial property in biological and medical applications, where controlled or noise covariables can greatly affect the signal and hide the effect of the variable of interest.

When a source is declared as significant, the model parameters also provide a quantification of the effect of the variable of interest on the spectra through the mixing weights. Figure 5 illustrates the linear effect of the hippurate dose on the mixing weights for the four levels of citrate. The slope of the line is 2:6410^{-}6, the parameter β_1 of the linear model for s_3.

Additionally, both linear regression and ANOVA models select

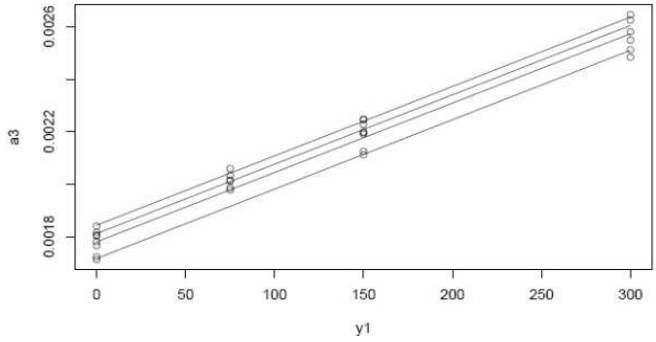

Figure 5: Relationship between the hippurate dose (y_1) and the vector of mixing weights a_3. Each of the four lines represents the estimated linear regression models for one fixed value of y_2.

s_1 (spectral profile of pure urine) and s_2 (spectral profile of pure citrate) as significant sources. In linear regression models, the sign of corresponding parameters $\hat{\beta}_{j1}$ and $\hat{\beta}_{j2}$ is negative, indicating a negative contribution of these sources to the observed spectra. This is easily explained by the constant sum normalization pre-processing applied to the spectra: if the peak heights corresponding to one metabolite in the spectra increase, the peaks corresponding to all other products (urine and citrate) decrease accordingly.

When more sources are declared as significant (but with less interpretable sources), the methodology presented in this paper allows to reconstruct the effect of the change of one factor level on the spectra independently to the effect of possible confounding model factors. 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} and 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 6 presents the three contrasts obtained when y_1 is introduced as a continuous variable in the model. This figure shows that, 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 appear everywhere else due to the normalization. The corresponding figure when y_1 is introduced as a qualitative variable is very similar.

More discussion concerning this data set and a generalization to more complex mixed models may be found in ref. [18].

Conclusions

The biomarker identification in ^{1}H-NMR based metabolomics is traditionally realised, with some limitations, via the examination of the two first components of a PCA, but without any statistical testing confirmation of factors effects. In this paper, we presented a new methodology providing three kinds of knowledge on ^{1}H-NMR metabolomics 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. This paper first presents a process to decompose by ICA the spectral data into statistically independent components and shows, on experimental data, that ICA allows to visualize, through the resulting sources, the spectral profile of independent metabolites contained in the studied bio fluid and their quantity through the corresponding mixing weights. Then, linear mixed statistical modelling is applied on ICA results to select the sources or spectral regions changing significantly according to the factors of interest. Finally, the selected sources are used to reconstruct the spectra and to compute contrasts presenting the alterations in specific regions caused by different changes of the factor of interest. Beside their discovery, contrasts also allow to visualize the alterations of potential biomarkers for defined changes of covariate conditions or context.

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.

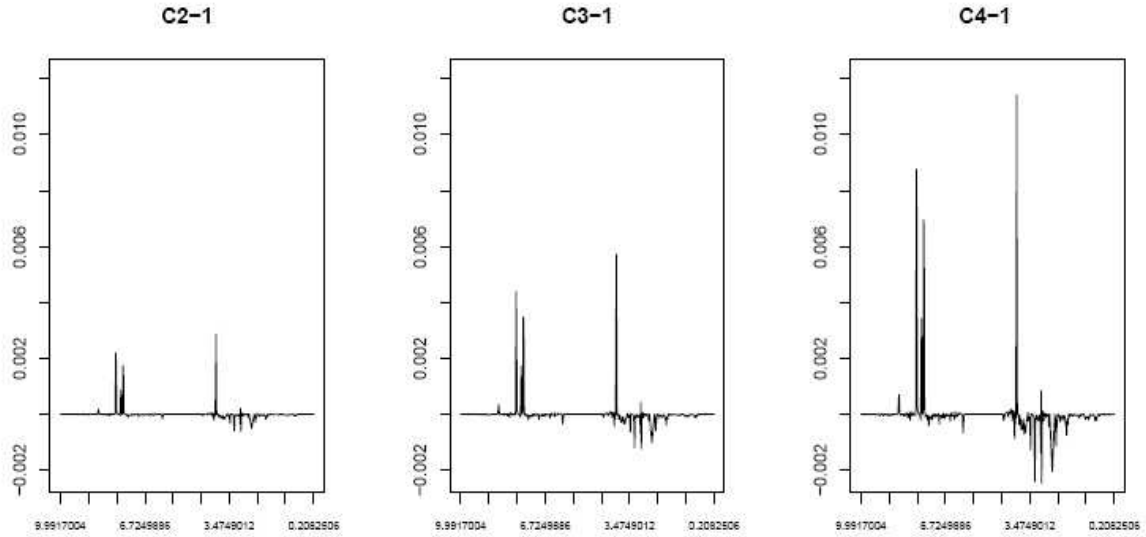

Figure 6: The three contrasts obtained when y_1 is introduced as a continuous variable.

In this paper, source selection is based on t-statistics computed on the weight vectors without using their significance levels. An accurate source selection is provided, 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 us to apply it to a large variety of complex metabolomics 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.

Furthermore, this methodology has been applied on a real metabolomic AMD dataset (see Supporting Information). The spectral biomarkers linked with this disease correspond to a metabolite supporting biological explanation of the setting of AMD.

Acknowledgement

The authors are grateful to the Centre Intrafacultaire de Recherche du Médicament, Laboratoire de Pharmacognosie et de Chimie Pharmaceutique, ULg, for providing data. Support from the IAP Re-search Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged.

References

1. Benjamini B, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29: 1165-1188.
2. Brown H, Prescott R (2006) Applied Mixed Models in Medicine, (2ndedn), John Wiley and Sons.
3. Common P (1994) Independent Component Analysis, a New Concept? Signal Processing 36: 287-314.
4. De Tullio P (2012) Biomedical application of NMR metabolomics: study of Age-related Macular Degeneration (AMD) ULg.
5. Eilers PHC (2005) Baseline correction with asymmetric least squares smoothing (Discussion Paper), Department of Medical Statistics. Leiden University Medical.
6. Friebolin H (1998) Basic one and two-dimensional NMR spectroscopy (3rdedn), Chapters 1 and 2.
7. Halouska S, Powers R (2006) Negative impact of noise on the principal components analysis of NMR data. J Magn Reson 178: 88-95.
8. Hyvärinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Network 10: 626-634.
9. Hyvärinen A, Oja E (2000) Independent Component Analysis: algorithms and applications. Neural Networks 13: 411-430.
10. Jolliffe I (1986) Principal Component Analysis, Springer-Verlag. New York.
11. Lee S, Batzoglou S (2003) Application of independent component analysis to microarrays. Genome Biology 4: R76.1-R76.21.
12. Liebermeister W (2002) Linear modes of gene expression determined by independent component analysis, Bioinformatics 18: 51-60.
13. Nicholson J, Connelly J, Lindon JC, Holmes E (2002) Metabolomics: a generic platform for the study of drug toxicity and gene function. Nature Reviews Drug Discovery 1: 153-161.
14. Noël A, Jost M, Lambert V, Lecomte J, Rakic J (2007) Anti-angiogenic therapy of exudative age-related macular degeneration : current progress and emerging concepts. Trends in Molecular Medicine 13: 345-352.
15. Nowak M (2005) Changes in lipid metabolism in women with age-related macular degeneration. Clinical and Experimental Medicine 4: 183-187.
16. Pinheiro J, Bates D (2000) Mixed-effects models in S and S-plus.
17. Rousseau R, Govaerts B, Verleysen M, Boulanger B (2008) Comparison of some chemometric tools for metabolomics biomarker identification, Chemometrics and Intelligent Laboratory Systems 91: 54-66.
18. Rousseau R (2011) Statistical contribution to the analysis of metabolomic data in ^{1}H-NMR spectroscopy UCL.
19. Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J (2004) Metabolite fingerprinting: detecting biological features by independent component analysis. Bioinformatics 15: 2447-2454.
20. Searle SR, Henderson CR, Amer J (1979) Dispersion Matrices for Variance Components Models. JASA 74: 465-470.
21. Trygg J, Holmes E, Lundstedt T (2007) Chemometrics in metabolomics. Journal of Proteome Research 6: 469-479.
22. Vanwinsberghe J (2005) Bubble: development of a Matlab tool for automated ^{1}H-NMR data pro-cessing in metabolomics. Master's thesis, Université de Strasbourg.