INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2012/04

Advantages of the Bagidis methodology for metabonomics analyses: application to a spectroscopic study of Age-related Macular Degeneration

TIMMERMANS, C., de TULLIO, P., LAMBERT, V., FRÉDÉRICH, M., ROUSSEAU, R. and R. von SACHS

Les apports de la méthodologie Bagidis pour l'analyse de données métabonomiques : application à une étude spectroscopique de la Dégénérescence Maculaire Liée à l'Age

Advantages of the Bagidis methodology for metabonomics analyses: application to a spectroscopic study of Age-related Macular Degeneration

Catherine Timmermans[1], Pascal de Tullio[2], Vincent Lambert[3], Michel Frédérich[2], Réjane Rousseau[1] & Rainer von Sachs[1]

[1] *ISBA, Université catholique de Louvain. Voie du Roman Pays, 20, BE-1348 Louvain-la-Neuve.*
E-mail : *catherine.timmermans@uclouvain.be, rejane.rousseau@uclouvain.be, rvs@uclouvain.be*
[2] *CIRM, Université de Liège. Av. de l'Hôpital, 1, BE-4000 Liège.*
E-mail : *P.DeTullio@ulg.ac.be, M.Frederich@ulg.ac.be*
[3] *LBTD, Université de Liège et ophtalmologie, CHU de Liège. Av. de l'Hôpital, 1, BE-4000 Liège.*
E-mail : *vincent.lambert@chu.ulg.ac.be*

Résumé
La méthodologie Bagidis propose une mesure de distance entre spectres qui tient compte des variations horizontales et verticales affectant les pics spectraux, dans un cadre unifié. Cette méthode repose sur une décomposition des spectres dans une base d'ondelettes de Haar asymétriques. Ses atouts pour l'étude de spectres ^{1}H RMN en métabonomique sont illustrés ici dans le cadre d'une étude d'une maladie oculaire, la dégénérescence maculaire liée à l'âge. Une analyse visuelle, un modèle de détection de la maladie et une recherche de biomarqueurs sont proposés et comparés avec des méthodes reconnues.
Mots-clés : spectroscopie, métabolomique, non-alignement, ondelette, distance, classification

Abstract
The Bagidis methodology proposes a distance measure between spectra, that takes into account, in a unified framework, both horizontal shifts and amplitudes variations that might affect spectral peaks. The method relies on the expansion of the spectra in unbalanced Haar wavelet bases. Its opportunity for investigating ^{1}H NMR spectra in metabonomics is illustrated here in the framework of a study of an eye disease: age-related macular degeneration. Visual analysis, disease detection model and search for biomarkers are proposed here and compared with known methods.
Keywords : spectroscopy, metabolomics, misalignement, wavelet, distance, classification

1 Introduction

Metabonomics and metabolomics studies are analyses which aim at the simultaneous detection of every small weight molecule present in a biofluid, an organ or an organism (see *Nicholson and*

Lindon, 2008, for instance, for an introduction to this field). Those "small weight molecules", the molecular weight of which being typically less than 1500 daltons, are referred as *metabolites*. Numerous biological processes affect the concentrations of metabolites. Compounds of which the concentration is specifically modified by a given process are called *biomarkers* for that process. They can be seen as the *fingerprint* of the biochemical reactions underlying the given process. Identifying biomarkers leads to a better understanding of biological processes, and might help at designing efficient tools for its detection or prediction.

Metabolomics studies become more and more frequent in various scientific area (*Lindon et al, 2007*, gathers some example applications): detection of origin in the food industry, cultivars discrimination in agronomy, toxicological studies in environmental and pharmaceutical sciences, screening of drug candidates in pharmaceutical sciences, diagnostic tool in medicine, etc. The present work is concerned about metabolomics data investigation for biomedical purposes: we aim at discriminating blood serum samples between patients suffering from Age-related Macular Degeneration (AMD) and healthy patients. AMD is an ocular disease, that is a leading cause of vision loss in western countries amongst people aged fifty or older (see *Noël et al, 2007* , for instance). However, behind this specific application, the methodology we describe has a larger scope and might advantageously be applied on various metabolomics datasets.

Metabolomics datasets often consist in nuclear magnetic resonance spectra (an overview of this technique can be found in *Lindon et al, 2007*). From a statistical point of view, those spectra are curves with *sharp local patterns* ("spectral peaks"). Not only their amplitudes but also their locations and shapes are affected by noise, this noise arising from the biological variability of the samples but also from unavoidable changes of the experimental conditions of spectra acquisition. However, most multivariate statistical methods rely on the good alignment of the peaks to be compared (see *Timmermans and von Sachs, 2010*, for a discussion). Otherwise, false differences might be detected between the spectra. Realignment techniques, such as *dynamic time warping* have thus been developed, which can be applied as a preamble to the statistical analysis. Those realignment techniques are however imperfect.

In this context, the Bagidis methodology (*Timmermans and von Sachs, 2010*) aims at explicitly and simultaneously taking into account both amplitudes variations and horizontal shifts that might affect the patterns in a curve. This methodology relies on the definition of a semi-distance based upon the expansion of the curves in unbalanced Haar wavelet bases. For each curve, an unbalanced Haar wavelet basis is selected so as to hierarchically encode the patterns the curve is made of: the main patterns are supported by the first basis vectors, while subsequent basis vectors support less important ones. Every basis vector is associated to a specific level change in the curve. Such wavelet bases are associated to each of the spectra, using a sliding window to focus on successive smaller spectral zones. The distance between two spectra is measured as a weighted sum of hierarchically computed differences in both the locations and the amplitudes of the pattern from one spectra to another. Visualization tools, classification procedure and statistical tests can be used, that take into account the Bagidis semi-distance.

Given this, we investigate the AMD dataset as follows: we blindly discriminate blood serum samples from healthy and diseased patients; we build a nonparametric model for predicting the AMD health status from blood serum; we select statistically discriminative spectral peaks with a aim to identify AMD biomarkers; we discuss whether statistically discriminative spectral peaks are related to systematic amplitude changes or horizontal variations, or both simultaneously. At each step of our analysis, we discusses how Bagidis compares to a recently published statistical analysis of the AMD dataset (*Rousseau, 2011*) and show how this methodology can be used as

an useful complement to statistical tools usually used in metabolomics (*Rousseau et al, 2008*).

This paper is organized as follows. Section 2 gives an overview of the BAGIDIS methodology. Section 3 describes the AMD dataset. Section 4 discusses the statistical analysis of the AMD dataset. Section 5 concludes.

2 An overview of the Bagidis methodology

The acronym BAGIDIS stands for *BAses GIving DIStances*, as *basis expansion* is at the core of the methodology, the latest being centered on the introduction of a new *distance measure*. The BAGIDIS methodology has been introduced in *Timmermans and von Sachs, 2010*. Further investigation of its use in nonparametric functional statistics (*Ferraty and Vieu, 2006*) is provided in *Timmermans et al, 2011*. Key ideas are as follows.

As a first step of the procedure, each curve is decomposed in a set of short series using a sliding window, so that each windowed series should not contain two significant patterns of the same amplitude. The length of the window is problem-dependent an is denoted Dt. Each windowed segment x of each of the curves in the dataset is expanded in a particular wavelet basis, which is referred to as the *Best Suited Unbalanced Haar Wavelet Basis* (BSUHWB). We denote the expansion of x in this basis as $x = \sum_{k=0}^{Dt-1} d_k \psi_k$, where the coefficients d_k (hereafter the *detail* coefficients) are the projections of x on the corresponding basis vectors ψ_k. The BSUHWB basis is obtained using the *Bottom-Up Unbalanced Haar Wavelet Transform* (BUUHWT) proposed by *Fryzlewicz, 2007*.

An interesting property of the Unbalanced Haar wavelet bases expansions, is that the set of points $\{y_k\}_{k=1\ldots Dt-1} = \{(b_k, d_k)\}_{k=1\ldots Dt-1}$ determines the *shape* of x uniquely, the complete determination of x requiring the additional coefficient d_0 that encodes the mean level of the series. Furthermore, the BUUHWT induces an interesting property of hierarchy in the resulting BSUHWB expansion. The idea is that the ranking of the basis vectors of the BSUHWB reflects the decreasing importance of the patterns they encode, for the description of the global shape of x. The notion of *hierarchy* that we refer to is the hierarchy induced by the BUUHWT algorithm itself: by construction, x is encoded in its BSUHWB as the sum of a constant mean level (rank $k = 0$) and a linear combination of level changes, the few first (small rank indexes) encoding the most striking features of x, while the last ones (large rank indexes) are less important. In such a way, the *Bottom-Up Unbalanced Haar Wavelet Transform* allows for an automatic and unique hierarchical description of each of the segment into a segment-adapted orthonormal basis. The hierarchy makes the resulting bases comparable to each other, although different. Consequently, we propose to compare the segments through a weighted p-norm between their mapping $\{y_k\}$ into the location-amplitude space of their breakpoints and details coefficients:

$$d^{\text{BAGIDIS}}(x^{(1)}, x^{(2)}) = \sum_{k=1}^{Dt-1} w_k \left\| y_k^{(1)} - y_k^{(2)} \right\|_{\lambda p} = \sum_{k=1}^{Dt-1} w_k \left(\lambda \left| b_k^{(1)} - b_k^{(2)} \right|^p + (1-\lambda) \left| d_k^{(1)} - d_k^{(2)} \right|^p \right)^{1/p}$$

with $p = 1, 2, \ldots, \infty$, with $\lambda \in [0; 1]$, and where w_k is a well suited weight function. As such, this semi-distance takes advantage of the hierarchy of the well adapted unbalanced Haar wavelet bases: breakpoints and details of similar rank k in the hierarchical description of each segment are compared to each other, and the resulting differences can be weighted according to that rank. As the breakpoints point to level changes in the segments, the term $\left| b_k^{(1)} - b_k^{(2)} \right|$ can be interpreted as a measure of the difference of location of the features, along the horizontal axis.

Being a difference of the projections of the segments onto wavelets that encode level changes, the term $\left|d_k^{(1)} - d_k^{(2)}\right|$ can be interpreted as a measure of the differences of the amplitudes of the features, along the vertical axis. Such a dissimilarity d^{BAGIDIS} is shown to be a semi-distance. It is computed for each windowed segment separately, for each pair of spectra of the dataset. The semi-distance between two spectra is then defined as the average value of the distances between its windowed segments.

In a prediction setting, weights w_k should ideally be 1 at rank k if that rank carries information for discriminating the series, and 0 otherwise. This is easily obtained using a cross-validation procedure. When no prediction criterion is at hand, or in order to get a first idea of how the dissimilarities do behave, we suggest in *Timmermans and von Sachs, 2010,* to *a priori* use the weight function $w_k = \frac{\log(Dt+1-k)}{\log(Dt+1)}$. This allows to associate a large weight to the comparison of features encoded at the first rank of the hierarchy, and a decreasing weight to the smaller features at the end of the hierarchy, which is empirically what we expect. The parameter λ actually defines a scaling in the *breakpoints-details* plane, and hence in the original units of the problem. Setting λ at its extreme values 0 or 1 allows to investigate the contributions of the breakpoints differences and details differences separately. In a prediction setting, λ can easily be optimized using cross-validation. Besides, the presence of this parameter allows the semi-distance to be robust with respect to scaling effects: if λ is optimized according to a given criteria (such as the mean square error of a prediction model), the relative dissimilarities between the series of a dataset will remain the same, whatever the scales of measurements along the horizontal and vertical axes, so that the predictive qualities of the model will not be affected by such a change in the units of measurements.

3 The AMD dataset

AMD is an ocular pathology, that can lead to rapid vision loss. It affects central fine vision, needed for reading, driving and face recognition, for instance. This disease exists in two distinct forms, one of which arising from an uncontrolled formation of new blood vessels (*angiogenesis*) under the *macula*, a part of the retina at the rear of the eye. The misknowledge of AMD anthology motivates the search for biomarkers in blood serum samples through a metabolomics approach (*Noël et al, 2007*). The AMD database was originally collected for a study lead by *de Tullio, Frédérich and Lambert* (Université de Liège). It consists in 200 blood samples, 100 of which arising from AMD patients and the other 100 arising from non-AMD patients ("control" patients), the AMD health status being diagnosed by an ophthalmologist. All AMD patients are aged over sixty and are followed by an ophthalmologist at *Centre Hospitalier Universitaire* in Liège, Belgium. Control patients are aged-matched patients in the same hospital, without any sign of ocular disease and not having a known history of AMD. The database also contains some additional general and clinical information. A complete description of this database can be found in *Rousseau* (2011).

A one-dimensional ^{1}H NMR spectrum was acquired from each blood sample, using a 500 MHz Bruker Avance spectrometer. A CPMG sequence with water pre-saturation was applied to attenuate broad signals arising from protein and water. Due to spectral acquisition problem, 6 AMD samples and 1 control sample were removed from the study. The resulting product of an ^{1}H NMR spectrum acquisition is a time signal called *Free Induction Decay* (FID). In order to chemically interpret the signal, each FID is converted in a spectral signal using a Fourier transform. Before and after this Fourier transform, several other pre-treatment of the

signals are also needed for the data to be statistically exploitable. In this study, we used the automatic pre-treatement procedure for metabolomics data which is advised and validated by *Rousseau, 2011*. It includes first order phase correction, suppression of the solvent, apodization by a scale function, apodization by an exponential function, Fourier transform, zero-order phase correction, setting to zero of negative values, warping, conversion in ppm scale, spectral window selection, bucketing, removal of undesired regions, spectral zone aggregation and normalization. More details on the acquisition procedure and on the pre-treatments can be found in *Rousseau, 2011*.

As a result of this procedure, the AMD dataset contains 193 spectra, of which 94 comes from AMD patients. Each spectra is a curve of 600 consecutive intensity measures in a spectral range going from 10 to 0.2 ppm.

4 Statistical Analyses

Except if mentioned otherwise, we make use of the Bagidis semi-distance with parameters $Dt = 25$, $p = 2$, $\lambda = 0.5$ and the default value of w_k. Results we obtain are compared with the recent results obtained by *Rousseau, 2011*. We see that additional insight into the data is gained by using the Bagidis methodology.

4.1 Visual analysis

Figure 1, *top left*, provides with the projections of the spectra on the first plane of a principal component analysis (PCA). This representation is to be compared with a *multidimensional scaling* (MDS) representation of the dataset, based upon the Bagidis semi-distance (with $\lambda = 0.5$) in Figure 1, *top right*. Multidimensional scaling is a projection technique that aims at preserving given distances between the observations in the dataset, so that the proximities of the data in the plane of projection can be interpreted -up to a certain degree- as "real" proximities of the data according to the chosen distance. MDS used jointly with the Euclidean distance corresponds to a PCA. In both representations, points are colored in different values according to their AMD health status.

We see that using Bagidis allows for a nearly optimal discrimination between AMD and non-AMD serum spectra, the distinction being essentially encoded by one single axis, while PCA detects an effect of the AMD health status but does not achieve such a clear discrimination. Furthermore, four outliers were detected in the data by *Rousseau, 2011* using visual inspection of this PCA representation, and a PCA for group-centered data. Those spectra are marked by triangles instead of points in Figure 1. They were removed of the dataset in *Rousseau, 2011*. We do not observe such a aberrant behavior for those spectra when using Bagidis. This might indicate that a problematic warping, resulting in misalignment, may be the cause of the aberrant behavior observed in the PCA. In this study, we do not discard those spectra from the database.

Figures 1 *bottom left and right* are MDS representations, obtained using a balance parameter λ fixed as 0 and 1 respectively. This allows to diagnose the effect of detail differences and breakpoint differences in the distance separately. Although some information on the AMD health status is clearly contained in the detail differences, we observe the major role of breakpoints location for discriminating the spectra. This might indicate a systematic peak appearance, shape modification of a peak, horizontal shift, or change of sign in the difference of amplitudes of neighbor peaks.

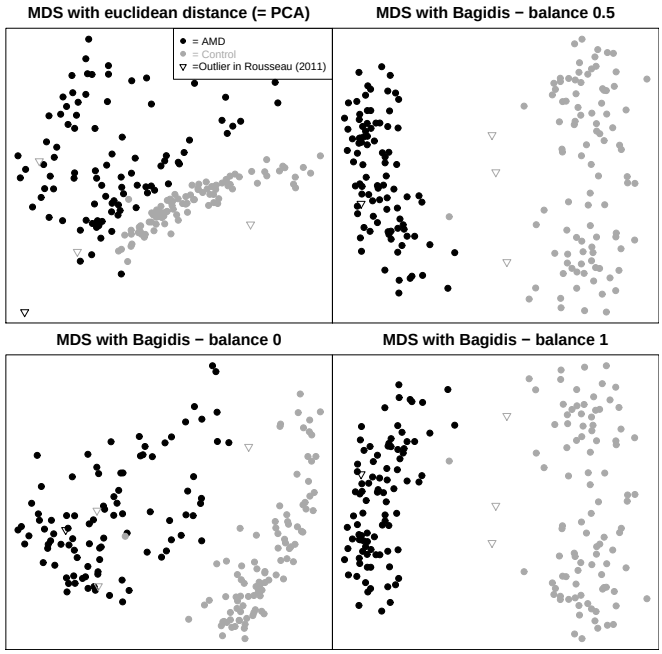

Figure 1: PCA representation, as compared with MDS representation using BAGIDIS with balance parameter $\lambda = 0.5\,, 0$ and 1 respectively. Points are colored according to the AMD health status of the corresponding projected spectrum.

4.2 AMD detection

We aim at predicting from the spectrum if a patient is affected by AMD or not. A training set of 150 spectra is randomly selected and a functional nonparametric discrimination model is adjusted (*Ferraty and Vieu, 2007*). This model is a k-nn predictor relying on a matrix of semi-distances between the spectra, with k being cross-validated. We consider the adjustment of such a model using BAGIDIS and compared its performances with those obtained using the same model with several other semi-distances: the Euclidean distance, the PCA-based distance, a derivative-based semi-distance, a semi-distance that realigns before computing an Euclidean distance (see *Ferraty and Vieu, 2007* or *Timmermans et al, 2011* for definitions). In each case, the number of misclassification observed on the remaining 43 spectra is recorded. This test for the prediction of the AMD health status from the spectra is repeated 80 times, with different randomly selected training sets. Results are summarized in Table 1, for BAGIDIS and its best competitor, being a PCA-based semi-distance with at least 6 components. We observe that the non-optimized BAGIDIS obtains *no error* 10% more often than the PCA-based semi-distance. Furthermore, we can optimize the weights and the λ parameter of the BAGIDIS semi-distance using a cross-validation procedure within the training set, and the resulting model is tested on the remaining 43 series. This test is repeated 18 times on different randomly selected training sets, and no prediction error occurs. At each repetition, only 1 non-zero weight is selected. We observe no prediction error in every case, indicating a risk of misclassification that is probably smaller than 0.05. This indicates a very good capacity of discriminating the serum spectra from AMD and healthy patients.

	Occurrences of 0 error out of 43 predictions	Occurrences of 1 error out of 43 predictions
PCA-based semi-distance with $q \geq 6$	40 times out of 80 **50%**	40 times out of 80 **50%**
Non-optimized BAGIDIS semi-distance with prior weights and $\lambda = 0.5$	48 times out of 80 **60%**	32 times out of 80 **40%**
Optimized BAGIDIS semi-distance (1 non zero weight is selected)	18 times out of 18 **100%**	0 times out of 18 **0%**

Table 1: Summary results for the prediction of the AMD health status from the spectra.

4.3 Search for biomarkers

As a last step of the analysis, we aim at identifying AMD biomarkers in the spectra. Six advanced statistical methods for the discovery of of metabolomics spectral biomarkers from ^{1}H NMR spectra have been identified in *Rousseau et al, 2008*: multiple hypothesis testing (MHT), supervised principal component analysis (s-PCA), supervised independent component analysis (s-ICA), discriminant partial least squares (PLS-DA), linear logistic regression (LLR) and classification and regression tree (CART). A description of those methods can be found in *Rousseau et al, 2008*, as well as an assessment of their relative performances: recommendation is given to use s-PCA with caution due to its low general efficiency; use of CART is discouraged due to its noise sensitivity; the other four methods are diagnosed promising. All those methods have been applied to the AMD database (with outliers excluded) by *Rousseau, 2011*. For each method, the 20 most significant biomarkers have been identified. Results are presented in Figure 2. Here, we compare those results to the ones we obtain using *double geometrical t-tests* (*Timmermans and von Sachs, 2010*) based upon the BAGIDIS semi-distance.

The idea of *double geometrical t-tests* is as follows. We first restrict the dataset to sliding segments of the data located in a given range of ppm values, and compute the BAGIDIS distance matrix between those segments. Then, we test for the equality of the means of the distances between two AMD segments (*intra-group distances*) and the distances between one AMD and one non-AMD segment (*inter-group distances*). We also test for the equality of the means of the distances between non-AMD segments (*intra-group distances*) and between one AMD and one non-AMD segment (*inter-group distances*). In both case, the alternative is that intra-group distances are lower than inter-group distances. These tests are performed using Welch t-tests, assuming independence and normal distribution of the distances about their group mean. Combining the results of both tests allows to deduce the relative positions of AMD segments and non-AMD segments. Only if both t-tests significantly reject their null hypotheses are the two groups statistically different in mean. In this way, we detect if the selected sliding segment is significantly discriminant with respect to the AMD health status, by requiring a significance $\alpha = 1e - 10$ for both t-tests. This test is actually performed for each sliding segment of the dataset. A Bonferroni correction is thus applied to each p-value to account for the 576 simultaneous comparisons.

Detected differences are differences in the windowed spectral segments, and not at a specific spectral location. This is a difference with competitor methods. A given spectral location contributes therefore to a number of segments equals to Dt. For each spectral location, the number of significantly discriminant segment at which it contributes is computed and reported in Figure 3 for different parametrization of the BAGIDIS semi-distance. This number of significances

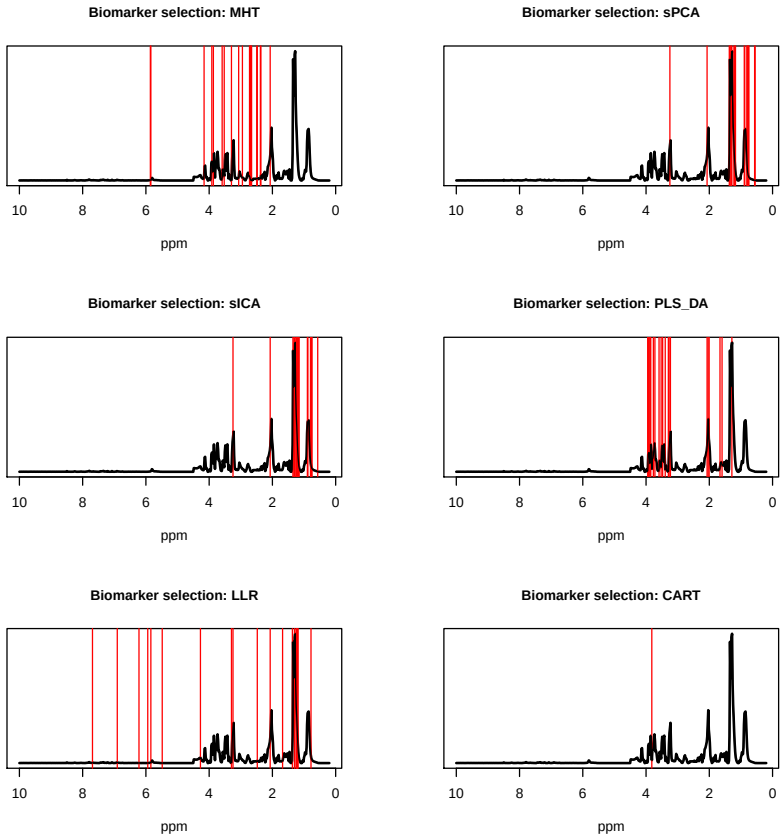

Figure 2: Search for AMD biomarkers using MHT, s-PCA, s-ICA, PLS-DA, LLR and classification and CART. For each method the 20 most significant biomarkers identified by *Rousseau, 2011* are marked by a vertical line. A typical spectrum of the AMD database is superimposed to ease the interpretation of biomarker detection.

can be used to search for biomarkers. The higher the number of significances, the higher the indication that the related spectral location might be a biomarker. A number of significances equals to Dt ($Dt = 25$ here) for a given spectral location indicates that each segment where this location contributes is significant with respect to the AMD health status. This clearly indicates for a biomarker.

Spectral zones from 3.99 to 3.06 ppm, as well as 2.48 and 2.27 ppm are strongly identified as biomarkers in Figure 3 (*top, left*). Those spectral zones are also detected by MHT and PLS-DA. A contribution in this zone, located at 3.24 is also detected by LLR, s-ICA and s-PCA, while CART has its only detection at 3.82. A highly significant detection of Bagidis also takes place in the spectral zone 7.24 to 6.89 ppm, which also contains a significant detection of LLR at 6.90 ppm. Some detection, although slightly less significant, is also found in the spectral zone 0.64 to 0.39, which is also detected by s-PCA, s-ICA, and, at one single location, by LLR. A detection at 2.07 also occur, which is also identified by all competitor methods except for CART. Possible, less significant, biomarkers are pointed out around 8.47 (no detection by other methods), 5.75 (also with LLR and MHT), 4.27 (also with LLR), 4.13 (also with MHT), 1.67 (also with LLR), 1.59 (also with PLS-DA), 1.36 (also with LLR and s-ICA), 1.22 (also with

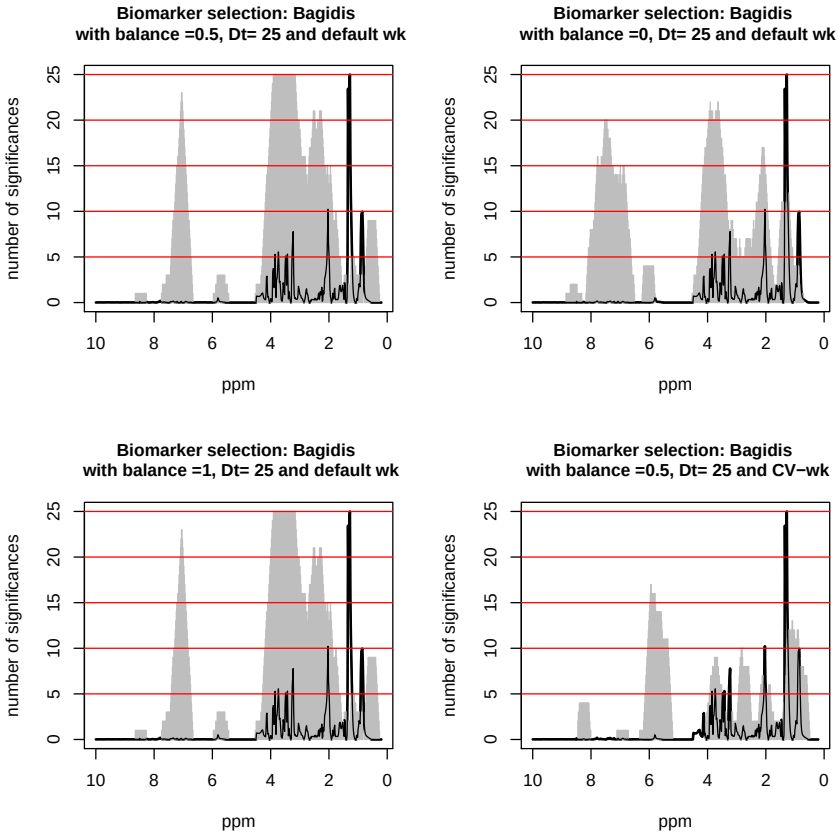

Figure 3: Search for AMD biomarkers using BAGIDIS with different parametrizations. For each spectral location, the number of detection of a spectral segment significantly discriminative for AMD which covers this location is provided. A typical spectrum of the AMD database is superimposed to ease the interpretation of biomarker detection.

s-PCA, s-ICA and LLR) and 0.90 ppm (also with s-ICA and s-PCA). Setting the λ balance parameter to 1 (breakpoints only) in the BAGIDIS semi-distance does not significantly modify these results. This is in accordance with the visual analysis in Subsection 4.1. Setting $\lambda = 0$ (amplitudes only) suppresses the detection of the spectral zone at ppms lower than 0.64, while the detection of the spectral zone at 1.22 ppm becomes more clear. The relative contributions of the spectral peaks around 5.75 and 2.07 ppm increase. This helps gaining an insight in the way the detected spectral zones do differ in AMD and non-AMD spectra. Finally, Figure 3 (*bottom, right*) identify spectral zones of significant differences between AMD and non-AMD spectra when w_k is set to its cross-validated value, as obtained in Subsection 4.2. This allows for finding a discriminant, sufficient but not exhaustive, set of biomarkers for AMD, as those biomarkers are the only one which contributes to the AMD detection model calibrated in Subsection 4.2.

We summarize this analysis by observing that BAGIDIS detects in one single study nearly all the spectral zones which were detected as biomarkers by at least one of the competitor methods. This emphasizes its consistency and its large scope of detection, which might be valuable for reducing the number of different statistical tools needed in a metabolomics study. Very few detection occur that are not detected by at least one other method, which might be an indication

that the method does not increase false detection rate relative to the combined use of competitor methods. From a biological point of view, this study has identified some biochemical pathways that could be implied in the anthology of AMD. Some additional biological experiments are required in order to validate these results.

5 Conclusion

This metabolomics study of the AMD dataset using the Bagidis methodology has been shown to be a useful complement to recent statistical analyses in the same field (*Rousseau, 2011*). It provides more informative visual discrimination of AMD and non-AMD blood samples which does not highlight outliers with respect to the semi-distance used. It allows for building a detection model for the AMD health status whose performances are shown to be really good. Finally, it allows for detecting - in a single analysis - a large set of biomarkers, this detection otherwise requiring the combination of six advanced statistical methods for biomarker search.

This analysis was performed using the *R software for statistical computing* and the library *Bagidis* which will be publicly available soon.

Bibliography

Ferraty F. & Vieu P. (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics, Springer.

Fryzlewicz P. (2007). Unbalanced Haar Technique for Non Parametric Function Estimation. Journal of the American Statistical Association, 102, pp. 1318-1327.

Lindon J.C., Nicholson J.K. & Holmes E. (2007). The Handbook of Metabonomics and Metabolomics. Elsevier.

Nicholson, J.K. & Lindon J.C. (2008). System biology: Metabonomics. Q & A. Nature, 455, pp. 1054-1056.

Noël A., Jost M., Lambert V., Lecomte J. & Rakic J.-M. (2007) Anti-angiogenesis therapy of exudative age-related macular degeneration: current progress and emerging concepts. Trends in Molecular Medicine, 13, 345-352.

Rousseau R., Govaerts B., Verleysen M. & Boulanger B. (2008), Comparison of some chemometric tools for metabonomics biomarkers identification. Chemometric and intelligent laboratory systems, 91, pp. 54-66.

Rousseau R. (2011). Statistical contribution to the analysis of metabonomics data in ^{1}H NMR spectroscopy. PhD thesis. Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Belgium.

Timmermans C. & von Sachs R. (2010). Bagidis, a new method for statistical analysis of differences between curves with sharp discontinuities. Submitted.
URL: `http://www.stat.ucl.ac.be/ISpub/dp/2010/DP1030.pdf`

Timmermans C., Delsol L. & von Sachs R. (2011). Using Bagidis in nonparametric functional data analysis: predicting from curves with sharp local features. Submitted.
URL: `http://dial.academielouvain.be/handle/boreal:81200`