# On combining wavelets expansion and sparse linear models for Regression on metabolomic data and biomarker selection

Nathalie Villa-Vialaneix, Noslen Hernández, Alain Paris, Céline Domange, Nathalie Priymenko, Philippe Besse

# On combining wavelets expansion and sparse linear models for regression on metabolomic data and biomarker selection

Nathalie Villa-Vialaneix[1,2]*, Noslen Hernández[3], Alain Paris[4], Céline Domange[5,6], Nathalie Priymenko[7], Philippe Besse[8]

[1] SAMM, Université Paris 1, 90 rue de Tolbiac, F-75013 Paris - France

[2] Université de Perpignan Via Domitia, IUT, Dpt STID, F-66860 Perpignan - France

[3] Advanced Technologies Application Center, CENATAV, Havana - Cuba

[4] INRA, Unité Mét@risk, AgroParisTech, 16 rue Claude Bernard, F-75005 Paris - France

[5] AgroParisTech, UMR 0791 Modélisation Systémique Appliquée aux Ruminants, F-75005 Paris - France

[6] INRA, UMR 0791 Modélisation Systémique Appliquée aux Ruminants, 16 rue Claude Bernard, F-75005 Paris - France

[7] ENVT, INRA, UMR 1089, Université de Toulouse, F-31076 Toulouse - France

[8] Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, F-31062 Toulouse - France

## Abstract

*nathalie.villa@univ-paris1.fr, Corresponding author

Wavelet thresholding of spectra has to be handled with care when the spectra are the predictors of a regression problem. Indeed, a blind thresholding of the signal followed by a regression method often leads to deteriorated predictions. The scope of this paper is to show that sparse regression methods, applied in the wavelet domain, perform an automatic thresholding: the most relevant wavelet coefficients are selected to optimize the prediction of a given target of interest. This approach can be seen as a joint thresholding designed for a predictive purpose.

The method is illustrated on a real world problem where metabolomic data is linked to poison ingestion. This example proves the usefulness of wavelet expansion and the good behavior of sparse and regularized methods. A comparison study is performed between the two-steps approach (wavelet thresholding and regression) and the one-step approach (selection of wavelet coefficients with a sparse regression). The comparison includes two types of wavelet bases, various thresholding methods and various regression methods and is evaluated by calculating prediction performances. Information about the location of the most important features on the spectra was also obtained and used to identify the most relevant metabolites involved in the mice poisoning.

# 1 Introduction

The recent development of high-throughput acquisition techniques in biology has brought a large amount of high dimensional data as high-resolution digitized signals. For instance, microarrays record the level of transcription of several thousands genes at the mRNA level and mass spectrometry or nuclear magnetic resonance (NMR) are used at the protein and metabolite levels. Modern biology now faces new issues related to these data: one of them is to deal with data having a high or even an extremely high dimension : typically, after a standard pre-processing, metabolomic profiles coming from NMR techniques have hundreds of variables for less than one hundred observations. In particular, the number of available samples is often much smaller than the data dimension and standard regression or classification methods are likely to overfit the data. For that reason, dimension reduction or variable selection are usually needed to improve the quality of the prediction in predictive models or to understand which features are involved in a given situation.

Dimension reduction are based on projections that usually build a small number of combinations of a large number of original features (see [Ramsay and Silverman, 1997] for examples and discussion about these approaches). Principal Component Analysis (PCA), Multidimensional scaling (MDS) [Cox and Cox, 2001] and Partial Least Squares (PLS) [Wold, 1975] are the most standard linear projection methods. Dealing with metabolomic data, a commonly used basis for projecting the data is the Wavelet Transform (WT) [Mallat, 1999]. Wavelet expansion is frequently performed to correct the baseline and to de-noise the data by removing the smallest details with a thresholding method. Then, in a second phase, a regression or a classification method is applied on the thresholded signal [Xia et al., 2007, Alexandrov et al., 2009]. On the other hand, selection methods select a small number of variables among the original ones to ensure an easy interpretation, often at the cost of deteriorated prediction performances: as an example, [Wongravee et al., 2009] used a bootstrap approach and PLS-DA to select variables in a large metabolomic dataset prior a classification. Finally, projection and variable selection are sometimes combined as in [Alsberg et al., 1998a, Kim et al., 2008].

The present paper tackles the issue of the best way to apply regression methods to metabolomic spectra. More precisely, a numerical variable of interest, that can be a phenotype or an environmental condition, is predicted from the metabolomic profile. As pointed out in [Rohart et al., 2012], the problem to predict a numerical phenotype from metabolomic data is little addressed in the literature so far, despite its numerous potential applications. Here, the focus is not merely put on achieving a good prediction accuracy but also on extracting the most influential features in the metabolomic spectra.

A one phase approach is tested that performs a sparse or a regularized regression method on the wavelet coefficients resulting from the wavelet representation of the spectra. Contrary to thresholding methods, where the coefficients selection is not directly related to the prediction of the target variable, the introduced approach automatically selects the most relevant wavelet coefficients in relation to the target variable. The relevance of the proposal is assessed through a case study. The purpose is to recover the drug dose ingested by a mouse from its metabolomic profile, in order to prevent a possible illness. A comparison study is performed on this real world problem, that leads to several conclusions: first, as was expected, wavelet transform is well adapted to the representation of metabolomic data and leads to better predictive performances. Then, variable selection by a blind thresholding of the wavelet coefficients deteriorates the predictions contrary to a variable selection performed by means of a sparse approach. This last method leads to the most accurate prediction performances.

The remaining of the paper is organized as follows: Section 2 presents the case study. Section 3 briefly surveys the state-of-the-art methods used to handle metabolomic data in a regression framework and specifically focuses on wavelet preprocessing. In this section, our proposal is described as well as the methodology used for the comparison. Finally, Section 4 discusses the results and shows that the obtained regression model is relevant enough to extract interesting biomarkers related to the studied target. Some conclusions are given in Section 5.

# 2 Case study and material

## 2.1 Problem description

The data used in this experiment are described in [Domange et al., 2008] and stand in the framework of a toxicology experiment based on metabolomic data. The study is devoted to the metabolomic exploration on the mouse model of the disruptive effect at the metabolic side of a plant, *Hypochoeris radicata* (L.) (HR), which is toxic for horse species. It may induce severe neuropathies that bring locomotive incapacitating damages [Domange et al., 2010].

The disruptive effect of HR is studied in male and female mice ($2 \times 36$) for 21 days at most. The mice were given a diet in which HR was introduced in form of a ground dry powder at 3 or 9%; a control group with 12 animals received no HR at all. 397 metabolomic spectra were acquired in urine, at different days of the experiment. In short, the data set is $(X_i, \mathrm{HR}_i, d_i)_{i=1,\ldots,397}$ where $X_i$ is a metabolomic profile (hence a curve, as shown in Figure 1), $\mathrm{HR}_i$ is the daily dose ingested by the corresponding mouse ($\mathrm{HR}_i \in \{0, 3, 9\}$ and $d_i$ is the number of days from the beginning of the experiment up to the spectrum acquisition ($d_i \in \{1, \ldots, 21\}$). More precise information about the data can be found in [Domange et al., 2010].

The issue of interest is to predict the total dose of HR ingested, which is the daily HR dose multiplied by the number of days of ingestion, from the metabolomic data. This problem can be written as a regression problem:

$$y_i = \Phi(X_i) + \epsilon_i \tag{1}$$

where $y_i = \mathrm{HR}_i \times d_i$, $\Phi$ is the regression function to be estimated and $\epsilon_i$ is an error term. This problem is motivated by several questions that frequently arise in such an experimental settings:

- the first motivation is to know if the metabolomic profile alone is enough to predict the drug dose ingested by an animal, which can be useful to prevent an illness;

- conversely, the second motivation is to understand if the influence of the HR dose ingestion is strong enough not to be seen as an artifact: if $y_i$ can be accurately estimated from $X_i$ then this is a strong indication that the HR dose and more precisely, its cumulative effect, is really disrupting the mouse metabolomic profile;

- finally the last motivation is to use the estimated regression function to corroborate a set of relevant metabolites influenced by the HR ingestion. The chosen approach is to extract the explanatory variables (i.e., the part of the metabolomic profiles) with the strongest predictive power, from the estimated regression function.

## 2.2 Data pre-processing

The data, acquired with $^1$H NMR technique, are transformed as described in [Domange et al., 2008] to obtain 397 spectra consisting in an intensity distribution with 751 (non zero) variables. This step can be seen as a routine designed to transform the original continuous signal into a discrete one, thus to ease its analysis. An example of a resulting spectrum is given in Figure 1.

[Figure 1 about here.]

In order to recover the continuity of the signal, discrete wavelet decomposition is performed on the pre-processed spectrum: this is one of the most commonly used signal transformation approach and it is particularly well suited for uneven and chaotic signals, such as metabolomic profiles. Additionally, the normal growth of the mice influences the metabolomic profile. As this effect could be mixed with the total HR dose ingested by the mice (which also depends on the day of measurement), a correction, based on the control group's quantiles alignment, is also performed on the wavelet coefficients. This correction is based on the assumption that, other the control group, no distribution variation in the metabolomic profiles should be seen: the group's quantile alignment is a robust method leading to comparable metabolomic profiles distributions each day, in the control group. This method is quite standard in such cases (see, e.g., what is done for microarray normalization in the **R** package `limma`, for instance [Bolstad et al., 2003]).

In the remaining, the obtained wavelet coefficients are denoted by

$(W_i)_{i=1,\ldots,397} \subset \mathbb{R}^D$ where $D$ is the number of wavelet coefficients used in the regression method (it depends on the wavelet basis and also on the DWT approach as described in Section 3.3 but in any case $D < 751$).

# 3 Methodological proposal

## 3.1 State-of-the-art on using DWT in regression problems

Wavelet transforms are often applied to signals as a pre-processing step before the statistical analysis [Davis et al., 2007, Xia et al., 2007]. A thresholding approach on the discrete wavelet transform is then generally performed in order to remove the smallest (and most irrelevant) detailed coefficients from the spectra representation. Standard thresholding strategies are the so-called "hard thresholding" that simply removes the smallest coefficients and leaves the others unchanged and the "soft thresholding" that removes the coefficients smaller than a given threshold and reduces the others from the value of this threshold. Of course, the choice of the threshold is very important and several solutions have been proposed: for instance, the SURE and Universal policies are calculated from an estimation of the level of noise and justified by asymptotic properties (see [Donoho and Johnstone, 1994, Donoho, 1995, Donoho and Johnstone, 1995, Donoho et al., 1995]). Also, [Nason, 1996] suggests to use a cross-validation criterion to choose the threshold and [Johnstone and Silverman, 1997] to rely on a different threshold for each level. More recently, [Gonzàlez et al., 2013] shows that keeping solely the finest details coefficients at the lowest decomposition level produces a representation of the data having the ability to correct a putative baseline default.

A natural approach to predict a phenotype from metabolomic profiles expressed in the wavelet domain would then be to perform a thresholding prior to the application of a well chosen regression method (see, e.g., [Xia et al., 2007]). But this methodology does not link the wavelet coefficients selection to the prediction purpose. An alternative solution is to perform a variable selection method, that takes into account the target variable, before learning the regression or the classification function. In this direction, [Alexandrov et al., 2009] uses a multiple testing approach with a Benjamini & Hochberg adjustment to select the relevant wavelet coefficients in relation to a target factor variable before building a classification model (based on SVM) to predict it. [Saito et al., 2002] proposes to select the wavelet coefficients that maximize the Kullback-Leibler divergence between estimated densities obtained for the various levels of a factor target variable before learning a classification function on the basis of the selected coefficients. Also, [Jouan-Rimbaud et al., 1997] uses a “Relevant Component Extraction” that thresholds the less informative wavelet coefficients from a PLS between the spectra and a target variable of interest. These latter approaches explicitly focused on wavelet coefficients selection but any feature selection method is expendable for such a task (see [Liu and Motoda, 1998, Guyon and Elisseeff, 2003] for reviews about feature selection). Feature selection algorithms can be time consuming and it has also been pointed out in [Raudys, 2006] that they can lead to feature *over*-selection that hinders the prediction performances.

Another approach is to simultaneously select the variables and optimize the prediction error: [Alsberg et al., 1998b] select the wavelet coefficients that minimize the cross validation error of a PLS regression. Model selection methods penalize the prediction error with a quantity depending on the number of variables involved in the regression (see, i.e., [Biau et al., 2005, Rossi and Villa, 2006] for examples in a similar framework where the signal is projected onto an orthogonal basis for classification purpose where the data are functions). However, model selection requires the definition of a relevant penalty term that can be hard to choose effectively, as pointed out in [Fromont and Tuleau, 2006].

## 3.2 A sparse one-phase approach

More recently, sparse methods [Tibshirani, 1996] have been intensively developed because they allow the selection of the relevant predictors during the learning process in an efficient and elegant way. The prediction error is penalized by the $L^1$ norm of the parameters of a linear model and it can be proved that this leads to nullify some of the parameters in an optimal way.

Our proposal is to use penalized regression methods to simultaneously define a regression function and select the most important wavelet coefficients involved in the definition of this regression function. More precisely, the numerical variable of interest (here, the total HR dose ingested by the mice, $(y_i)_i$) is predicted from the metabolomic spectra through a penalized linear model where the predictors are all the wavelet coefficients (without prior thresholding). More precisely, the regression function $\Phi$ in Equation 1 is estimated by a penalized linear regression on the wavelet coefficients (used instead of $X_i$ as predictor variables): $\hat{\phi}(W_i) = W_i^T \hat{\beta}$ where

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^D} \frac{1}{397} \sum_i \|y_i - W_i^T \beta\|^2_{\mathbb{R}^D} + \lambda p(\beta)$$

where $\|z\|^2_{\mathbb{R}^D} = \sum_{j=1}^D z_j^2$.

Depending on the form of the penalization, $p(.)$, the method is likely to perform a rough or less rough variable selection:

- if $p(\beta) = \|\beta\|_{L^1} = \sum_{j=1}^D |\beta_j|$, the linear regression is a sparse linear regression also named LASSO [Tibshirani, 1996]. It selects wavelet coefficients, in the set of $D$ original coefficients, in a optimal way for prediction purpose;

- if $p(\beta) = \|\beta\|^2_{\mathbb{R}^D}$, the linear regression is a ridge regression which tends to produce $\beta$ with small norms but does not perform a selection of the wavelet coefficients;

- if $p(\beta) = (1 - \alpha)\|\beta\|^2_{\mathbb{R}^D} + \alpha\|\beta\|_{L^1}$, $\alpha \in ]0, 1[$ the linear regression is the so-called “elasticnet” method [Zou and Hastie, 2005], proposed in an attempt to use the advantages of the two previous penalties. As LASSO, it selects a reduced number of wavelet coefficients involved in the regression function but this number is usually larger than the one obtained when using the LASSO method.

Using a sparse linear regression method, such as LASSO or elasticnet, then leads to perform a thresholding that is adapted to the regression task. Moreover, the thresholding is made in a joint way, leading to select a common set of wavelet coefficients for all the spectra (contrary to standard thresholding that nullify a different set of wavelet coefficients for each spectrum). This property is likely to help prevent overfitting. Finally, sparse regressions lead to the selection of a very limited number of coefficients that can, eventually, help the interpretation (see Section 4 for a discussion and a comparison of the different numbers of selected wavelet coefficients according to both methods).

## 3.3 Comparison methodology

The comparisons aim at understanding how the different approaches performs in predicting the total dose of HR ingested by mice. Different wavelet approximations and regression methods are combined. More precisely,

- the possible wavelet approximations applied to the pre-processed data (as described in Section 2.2) are raw spectra (no wavelet approximation), wavelet coefficients (Haar or D4 bases), thresholded wavelet coefficients (D4), undecimated wavelet detailed coefficients (D4). "*thresholded wavelet coefficients*" correspond to the wavelet coefficients that remain positive after a soft threshold with SURE policy and "*undecimated wavelet detailed coefficients*" correspond to the union of the finest details coefficients of the original spectra with the finest details coefficients of the shifted spectra (obtained using the approach of [Beylkin, 1992, Gonzàlez et al., 2013]). When using the full wavelet decomposition or the "undecimated wavelet" approach, the dimensionality of the original problem, $D = 751$ is left unchanged whereas the "thresholded wavelet" approach leads to a dimensionality reduction ($D = 71$ for D4 DWT), which is a standard way to handle large dimension regression tasks.

- the possible regression method applied to the wavelet coefficients are sparse or regularized regression methods as described in Section 3.2 (LASSO, ridge regression and elasticnet), PLS regression, which is a standard approach when dealing with a large number of variables and random forest [Breiman, 2001], as a basis for a comparison with non-linear methods.

For a sake of simplicity, only the following combinations are compared:

- any wavelet approximation is combined with the elasticnet regression. Our proposal is to use the full wavelet decomposition (without thresholding) with a sparse regression method. To enlighten the uselessness of the thresholding when using a sparse regression method, thresholding is also combined with elasticnet in the comparison;

- the full wavelet decomposition is also combined with any regression method described above.

A total of 9 combinations are thus compared, summarized in Table 1.

[Table 1 about here.]

In order to train and to evaluate each of these combinations, the following methodology is applied:

**Wavelet transform** First, the data are or are not preprocessed by a DWT. The obtained coefficients are also scaled (each coefficient is centered to a zero mean and scaled to a standard deviation equal to 1).

**Split** The observations (i.e., the pairs $(W_i, y_i)_i$) are randomly split into a training set $\mathcal{S}_T$ and a test set $\mathcal{S}_V$ with balanced sizes (approximatively 200 observations each) taking into account the proportion of observations in the groups defined by sex, dose (including the control group to train the regression function so that it can predict when the animal is not affected by HR ingestion) and day of measure. To estimate the methods variability, this step is repeated 250 times giving 250 training sets and the corresponding test sets.

**Train** The regression method is then applied to each training set. Several methods involve hyper-parameters that have to be tuned: for random forest, the hyper-parameters are the number of trees, the number of variables selected for a given split, ... They are set to the default values, coming from useful heuristics; the stabilization of the out-of-bag error is achieved using that strategy.

For sparse and regularized linear regressions, an optimal $\lambda$ is automatically selected through a regularization path algorithm (see, e.g., [Efron et al., 2004] for the LARS algorithm in the case of LASSO). Additionally, for elasticnet, the mixing coefficient $\alpha$ is set to 0.5 which was the best choice according to other experiments in which $\alpha$ was varied in $\{0.1, 0.25, 0.5, 0.75\}$ (not shown in this paper for a sake of simplicity).

Finally, for PLS, the number of kept components (between 1 and 40) is tuned by a 10-fold cross-validation strategy performed on the training set.

**Test** The root mean square error (RMSE) is calculated for each approach involved in the comparison and for all the corresponding test sets:

$$RMSE_V = \sqrt{\frac{1}{n_V} \sum_{i \in \mathcal{S}_V} (y_i - \hat{y}_i)^2}$$

where $n_V$ is the number of observations in the test set and $\hat{y}_i$ is the estimation of the total dose of HR ingested.

The methodology described above is illustrated in Figure 2. It leads to obtain nine sets of 250 test errors, one for each combination of a wavelet transform and regression algorithm.

[Figure 2 about here.]

All the simulations are performed using **R** free software [R Development Core Team, 2012] and the packages `wavethresh` [Nason and Silverman, 1994] (for wavelet facilities), `glmnet` [Zou and Hastie, 2005] (for sparse and regularized linear methods), `mixOmics` [Lê Cao et al., 2009] (for PLS) and `randomForest` [Liaw and Wiener, 2002] (for random forest).

# 4 Results and discussion

This section presents the results of the experiments described in Section 3. Section 4.1 is devoted to the comparison of the numerical performances of the various combinations. The differences between the approaches (including the number of wavelet coefficients selected) are discussed. Then, Section 4.2 extracts relevant features from the best combination of wavelet preprocessed and regression method and compares it with a previously known list. This provides another point of view on the relevance of the combination of the DWT with sparse and regularized linear models for metabolomic data analysis, this time as a feature selection method. The biomarkers that are the most involved in the prediction of the total dose of HR ingested are selected using an importance measure. The overall methodology is general enough to be expandable for any regression method.

## 4.1 Numerical performances comparison

The averaged RMSE over the 250 test sets as well as their standard deviations are reported in Table 2.

[Table 2 about here.]

In addition, the boxplot of the $R^2$ over the 250 test sets[1] are given in Figure 3 for the case where the data are expanded on the D4 basis and where all wavelet coefficients are kept.

[Figure 3 about here.]

For the best method (combination of a DWT on a D4 basis with elasticnet), the mean $R^2$ is equal to 89.00% which is quite satisfactory. Thus, the accuracy of the prediction on the sample test is good enough to be used as a relevant method to estimate the total dose of HR ingested by the animal from the metabolomic profile alone.

Conversely, being able to predict the HR ingestion from the metabolomic profile is a proof that the disrupting effect of HR on the metabolism is not an artefact because an accurate relation between both variables is established. Contrary to a test approach, that would have lead to test each part of the metabolomic profile, this approach enlighten the strength of the relation between the whole metabolomic spectrum and the target variable, here the HR dose. Moreover, it does not even require the use of a control group.

[1] $R^2 = 1 - \frac{\sum_{i=1}^{n_{\text{Test}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{\text{Test}}} (y_i - \bar{y})^2}$ where $\bar{y} = \frac{1}{n_{\text{Test}}} \sum_{i=1}^{n_{\text{Test}}} y_i$.

### 4.1.1 Comparison of the wavelet transforms

The first conclusion arising from Table 2 is that the wavelet transform effect is stronger than the choice of the regression method. In particular, using the wavelet coefficients remaining after a soft thresholding results in less accurate predictions than using all the wavelet coefficients or even than the direct use of the raw data.

Moreover, using all the wavelet coefficients in combination with a sparse approach (elasticnet or LASSO) is the most accurate method; the impact of the basis choice (D4 or Haar) is almost negligible. Undecimated wavelet transform is the second most accurate wavelet transform approach: this may be the indication that the coefficients with the finest details contain most of the useful information for the prediction task. Maybe, an optimal trade-off would be to select wavelet coefficients at several scales, leaving only the coefficients at the crudest scales.

To assess the significance of these conclusions, paired t-test were computed to compare the RMSE of the various wavelet transforms: the differences between the use of Haar or D4 wavelets are not significant (at level 1%) but the differences between the use of all D4 wavelet coefficients and the use of either the raw spectra, the D4 undecimated wavelet approach or the D4 thresholded coefficients are all significant. Note that, even if the differences between the averaged RMSE seem to be small, they are calculated over 250 replica which is a large enough number to provide confidence in these conclusions.

### 4.1.2 Comparison of the regression methods

Comparing the regression methods, those that are (at least partially) based on a sparse regularization, such as elasticnet and LASSO, obtain the best results. Ridge regression is not as accurate as the methods based on a sparse regularization but its variability is lower. Actually, combining a ridge and a sparse penalty in the elasticnet seems to slightly decrease the variability of the elasticnet results compared to those of the LASSO (except for two outlier samples). Moreover, the influence of the mixing parameter $\alpha$ is not really strong: test errors for elasticnet with $\alpha = 0.1$, 0.25 or 0.75 are not shown in the paper but would have mostly lead to the same conclusion: $\alpha = 0.1$ or 0.25 has slightly deteriorated (but comparable) test errors, whereas $\alpha = 0.75$ has test errors closer to the LASSO.

Finally, PLS, that is probably better suited for explanatory purpose, does not give very satisfactory predictive performances in this case study but also has a low variability. Here, random forest is the method that gives the worst accuracy and also the largest variability of the performances over the 250 test sets.

Once again, the significance of these conclusions can be assessed by paired t-tests: the differences between RMSE obtained by elasticnet and RMSE obtained by ridge regression are significant. Of course, the same remark holds for the comparison between elasticnet and any method performing worse than ridge regression. This leads to the conclusion that the combination of a DWT and a sparse linear method, such as elasticnet, is indeed a good choice to handle regression problems where the predictors are metabolomic data.

### 4.1.3 Number of selected wavelet coefficients

Section 3.2 explains that using a sparse method on all the wavelet coefficients can be seen as a joint thresholding adapted to the target variable. Then, it is interesting to compare the numbers of coefficients selected by sparse methods to the number of coefficients selected by a classical thresholding approach. For D4 basis, 71 wavelet coefficients remain after the soft thresholding phase. The numbers of selected coefficients over the 250 regression functions provided by elasticnet and lasso are given in igure 4.

[Figure 4 about here.]

The average number of selected coefficients is often much smaller than the one obtained with the classical thresholding approach. For instance, the best method (elasticnet) selects 46.5 wavelet coefficients on average. Hence, not only are the "one-phase" approaches faster and more accurate, they also select less (but more relevant, according to the increase in accuracy) wavelet coefficients.

## 4.2 Important biomarkers extraction

The relevance of the application of elasticnet on all the wavelets coefficients is assessed by using the learned regression function, obtained in the previous section, in order to extract the most important features related to the total dose of HR ingested. A natural approach would be to directly analyze the variables selected by the sparse regression but, because of the wavelet transform preprocessing, these are not directly linked to the spectra locations that are of interest.

Alternatively, a standard approach, for linear models, is to select the most important variables by the p-values of the coefficients associated to the variables; this approach is not reliable in our context, both because it only selects the most important wavelet coefficients (and, once again, not the spectra locations) and also because if the explanatory variables are highly correlated, the results of such tests are strongly related to the variables that are used in the model. A small change in the list of explanatory variables can lead to a very different list of significant variables and thus, the approach is not really reliable in the case of a large number of explanatory variables.

To overcome these difficulties and to achieve the study of the influence of the original variables (and not of the wavelet coefficients) in the prediction, we used a generalization of the importance measure originally designed for random forest [Breiman, 2001]. This approach provides a way to assess the relevance of biomarkers, to quantify their respective implications in the biological phenomenon and thus to corroborate a list of biomarkers already extracted elsewhere. In the following, Section 4.2.1 describes our approach whereas Section 4.2.2 analyzes the results.

### 4.2.1 A measure of the importance of the variables

L. Breiman proposes the calculus of an “importance” measure to assess the relevance of each explanatory variable in a random forest [Breiman, 2001]. This measure is based on the observations that are not used to train a given tree (out-of-bag observations): the values of the explanatory variable under study are randomly permuted and the importance is defined as the decrease of the accuracy (in terms of increased mean square error for a regression problem) between the predictions made with the real values and those made with the randomly permuted values. The more the MSE increases, the more important the variable is for prediction. This approach was proven to be successful in variable selection in [Archer and Kimes, 2008, Genuer et al., 2010].

We propose to use a similar approach to describe the way a wrong value for a given variable (here a given value in the spectrum) propagates through the wavelet transform and the regression function and affects the accuracy of the final prediction of the total dose of HR ingested ingested by the mouse. This analysis is focused on the best regression approach, i.e., the use of all wavelet coefficients coming from a D4 basis expansion combined with elasticnet. As in the approach proposed in [Breiman, 2001], the importance is calculated from observations that are not used during the training process. More precisely, the 250 test samples described in Section 3.3 are used to calculate importance measures: the “importance” of a variable is the mean rate (over the test sets) of MSE increase after a random permutation of its values among the individuals (the other variables remaining with their true values). The idea is to assess the prediction power of a variable by means of the prediction accuracy disruption when this variable is given false values. The process is repeated for the 751 variables corresponding to spectra locations, as described in Algorithm 1. It can handle the way a given part of the spectra affects the quality of the prediction of the total dose of HR ingested. It thus gives an assessment to the most relevant features in metabolomic spectra (i.e., the features that contribute the most to an accurate prediction of the HR dose), despite the series of transformations done.

**Algorithm 1** Variables importance calculation

```
for each explanatory variable, $v$ of the data set do {Variable loop}
    Randomization Randomize the values of $v$ for the 397 observations. The new explanatory variables (spectra) with randomized values for $v$ are denoted by $(X_i^v)_i$;
    Wavelet expansion Calculate the wavelet coefficients with a D4 expansion for $(X_i^v)_i$. These are denoted by $(W_i^v)_i$;
    for each test set, $\mathcal{S}_V$ do {Test set loop}
        Mean square error calculation Calculate the MSE based on the explanatory variables $(W_i^v)_{i \in \mathcal{S}_V}$, $\text{MSE}_{v,\mathcal{S}_V}$;
        Importance calculation for $\mathcal{S}_V$ Compare $\text{MSE}_{v,\mathcal{S}_V}$ to the original MSE obtained for the test set $\mathcal{S}_V$, $\text{MSE}_{\mathcal{S}_V}$: $\mathcal{I}_{v,\mathcal{S}_V} = 1 - \frac{\text{MSE}_{\mathcal{S}_V}}{\text{MSE}_{v,\mathcal{S}_V}}$;
    end for
    Importance calculation for variable $v$ Average over the $T = 250$ test samples: $\mathcal{I}_v = \frac{\sum \text{Test sets}\, \mathcal{I}_{v,\mathcal{S}_V}}{T}$.
end for
```

### 4.2.2 Results of the biomarkers extraction and comments

Figure 5 gives the importance of the 751 original variables (spectra locations) ranked by decreasing value.

[Figure 5 about here.]

One variable is clearly much more important than all the other ones because random permutations of its values cause an increase of almost 80% in MSE. Three other variables seem to be important (with importance greater than 20%) and another group of 5 variables are also important to a lesser extent (between the yellow line and the orange line in Figure 5).

The list of the "most important" spectra locations and the names of the associated metabolites (when it is known) are given in Table 3. Moreover, the location of these metabolites in a $^1$H NMR spectrum is shown in Figure 6.

[Table 3 about here.]

[Figure 6 about here.]

The most important metabolite is the *scyllo*-inositol which was also identified as an important metabolite in [Domange et al., 2008]. The other metabolites emphasized by the variable importance (creatinine, hippurate, valine) were also present in the original work: this confirms the reliability of our proposal. Other spectra locations, that do not correspond to known metabolites, are also identified by the variable importance. Noticing the relevance of the most important metabolites found by our approach, these unknown peaks are indications for further biological analysis to find new metabolites involved in the poisoning process.

Also, some differences arise when comparing this list with the list of biomarkers identified in [Domange et al., 2008]. Part of these differences may be explained by the fact that the dependent variable in [Domange et al., 2008] is the daily HR dose ingested (i.e., a factor variable with 3 levels) whereas, here, the total ingested dose was used in order to take into account the cumulative effect of the ingestion. But it is also the positive counterpart of not using a test approach and thus avoiding the standard false positive issue that comes with them. As the extracted spectra locations are directly related to the quality of the prediction, they are more reliable, even if not so well theoretically justified.

Finally, not only does this approach give a list of important spectra locations (corresponding to the total dose of HR ingested) but it also provides a quantification of the influence of the spectra location on the accuracy of the prediction. In our problem, *scyllo*-inositol therefore appears as the most important metabolite affected by HR ingestion because its randomization causes an 80% increase of the average MSE.

# 5 Conclusion

Wavelet transformation is commonly used to deal with spectrometric data in biology, especially for de-noising purposes. Moreover, this paper shows that, associated with a convenient learning method, it improves the understanding of the relation between metabolomic spectrum and a phenomenon of interest (for instance, metabolic disruptions linked to HR ingestion). It is also shown that using a de-noising approach, not related to the variable to be predicted, can lead to a dramatic loss of information. More precisely, some important variables seem to be located in parts of the spectra that could be seen as “minor” details. It is thus important to combine the wavelet transform and de-noising with the purpose of the study. Sparse methods, that combine a regression model and a variable selection seem to be well suited to this task: they perform a kind of joint thresholding of the wavelet coefficients that is directly related to the target variable. In particular, elasticnet gave the best performance in prediction and was also able to provide a relevant list of biomarkers, linked to the target variable, in our case study.

In conclusion, the combination of DWT with elasticnet can be used to accurately predict a numerical variable of interest from the metabolomic profile. It is also useful to identify and confirm the most important features involved in the biological process under study thanks to the importance measure introduced in this article.

# 6 Acknowledgements

The authors are grateful to Matthias De Lozzo and Cindy Hulin for providing help support in programming during their Master project in INSA Toulouse.

# References

[Alexandrov et al., 2009] Alexandrov, T., Decker, J., Mertens, B., Deelder, A., Tollenaar, R., Maass, P., and Thiele, H. (2009). Biomarker discovery in maldi-tof serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25(5):643–649.

[Alsberg et al., 1998a] Alsberg, B., Kell, D., and Goodacre, R. (1998a). Variable selection in discriminant partial least-squares analysis. *Analytical Chemistry*, 70:4126–4133.

[Alsberg et al., 1998b] Alsberg, B., Woodward, A., Winson, M., Rowland, J., and Kell, D. (1998b). Variable selection in wavelet regression models. *Analytica Chimica Acta*, 368:29–44.

[Archer and Kimes, 2008] Archer, K. and Kimes, R. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, 52:2249–2260.

[Beylkin, 1992] Beylkin, G. (1992). On the representation of operators in bases of compactly supported wavelets. *SIAM Journal on Numerical Analysis*, 29:1716–1740.

[Biau et al., 2005] Biau, G., Bunea, F., and Wegkamp, M. (2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172.

[Bolstad et al., 2003] Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Cox and Cox, 2001] Cox, T. and Cox, M. (2001). *Multidimensional Scaling.* Chapman and Hall.

[Davis et al., 2007] Davis, R., Charlton, A., Godward, J., Jones, S., Harrison, M., and J.C., W. (2007). Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems*, 85:144–154.

[Domange et al., 2008] Domange, C., Canlet, C., Traoré, A., Biélicki, G. Keller, C., Paris, A., and Priymenko, N. (2008). Orthologous metabonomic qualification of a rodent model combined with magnetic resonance imaging for an integrated evaluation of the toxicity of hypochoeris radicata. *Chemical Research in Toxicology*, 21(11):2082–2096.

[Domange et al., 2010] Domange, C., Casteignau, A., Pumarola, M., and Priymenko, N. (2010). Longitudinal study of australian stringhalt cases in France. *Journal of Animal Physiology and Animal Nutrition*, 94(6):712–720.

[Donoho, 1995] Donoho, D. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

[Donoho and Johnstone, 1994] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

[Donoho and Johnstone, 1995] Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

[Donoho et al., 1995] Donoho, D., Jonhstone, I., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B*, 57(2):301–369.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.

[Fromont and Tuleau, 2006] Fromont, M. and Tuleau, C. (2006). Functional classification with margin conditions. In *Proceedings of the 19th Annual Conference on Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 94–108.

[Genuer et al., 2010] Genuer, R., Poggi, J., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31:2225–2236.

[Gonzàlez et al., 2013] Gonzàlez, I., Eveillard, A., Canlet, A., Paris, T., Pineau, P., Besse, P., Martin, P., and Déjean, S. (2013). Undecimated wavelet transform to improve the classification of samples from metabolomic data. *JP Journal of Biostatistics*. Forthcoming.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

[Johnstone and Silverman, 1997] Johnstone, I. and Silverman, B. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 59:319–351.

[Jouan-Rimbaud et al., 1997] Jouan-Rimbaud, D., Walczak, B., Poppi, R., de Noord, O., and Massart, D. (1997). Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration. *Analytical Chemistry*, 69(21):4317–4323.

[Kim et al., 2008] Kim, S., Wang, Z., Oraintara, S., Temiyasathit, C., and Wongsawat, Y. (2008). Feature selection and classification of high-resolution NMR spectra in the complex wavelet transform domain. *Chemometrics and Intelligent Laboratory Systems*, 90(2):161–168.

[Lê Cao et al., 2009] Lê Cao, K., González, I., and Déjean, S. (2009). *****Omics: an R package to unravel relationships between two omics data sets. *Bioinformatics*, 25(21):2855–2856.

[Liaw and Wiener, 2002] Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

[Liu and Motoda, 1998] Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*, volume 454 of *The Springer Series in Ingineering and Computer Science*. Springer, Nowell, MA, USA.

[Mallat, 1999] Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press.

[Nason, 1996] Nason, G. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 58:463–479.

[Nason and Silverman, 1994] Nason, G. and Silverman, B. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, 3:163–191.

[R Development Core Team, 2012] R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0.

[Ramsay and Silverman, 1997] Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer Verlag, New York.

[Raudys, 2006] Raudys, S. (2006). *Structural, Syntactic and Statistical Pattern Recognition*, volume 4109 of *Lecture Notes in Computer Science*, chapter Feature over-selection, pages 622–631. Springer-Verlag, Berlin/Heidelberg, Germany.

[Rohart et al., 2012] Rohart, F., Paris, A., Laurent, B., Canlet, C., Molina, J., Mercat, M., Tribout, T., Muller, N., Iannuccelli, N., Villa-Vialaneix, N., Liaubet, L., Milan, D., and San Cristobal, M. (2012). Phenotypic prediction based on metabolomic data on the growing pig from three main European breeds. *Journal of Animal Science*, 90(12).

[Rossi and Villa, 2006] Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*, 69(7-9):730–742.

[Saito et al., 2002] Saito, N., Coifman, R., Geshwind, F., and Warner, F. (2002). Discriminant feature extraction using empirical probability density estimation and a local basis library. *Pattern Recognition*, 35:2841–2852.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, series B*, 58(1):267–288.

[Wold, 1975] Wold, H. (1975). *Soft modeling by latent variables; the nonlinear iterative partial least square approach.* J. Gani, Academic Press, London.

[Wongravee et al., 2009] Wongravee, K., Heinrich, N., Holmboe, M., Schaefer, M., Reed, R., Trevejo, J., and Brereton, R. (2009). Variable selection using iterative reformulation of training set models for discrimination of samples: application to gas chromatography/mass spectrometry of mouse urinary metabolites. *Analytical Chemistry*, 81(13):5204–5217.

[Xia et al., 2007] Xia, J., Wu, X., and Yuan, X. (2007). Integration of wavelet transform with PCA and ANN for metabolomics data-mining. *Metabolomics*, 3(4):531–537.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, series B*, 67(2):301–320.

# List of Figures

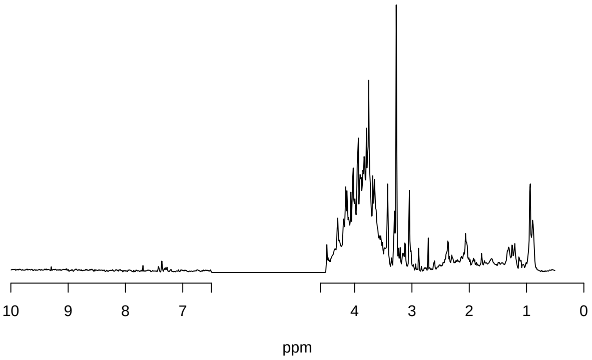

Figure 1: An example of metabolomic spectra from data discussed in Section 2 (female mice of the control group at day 0).

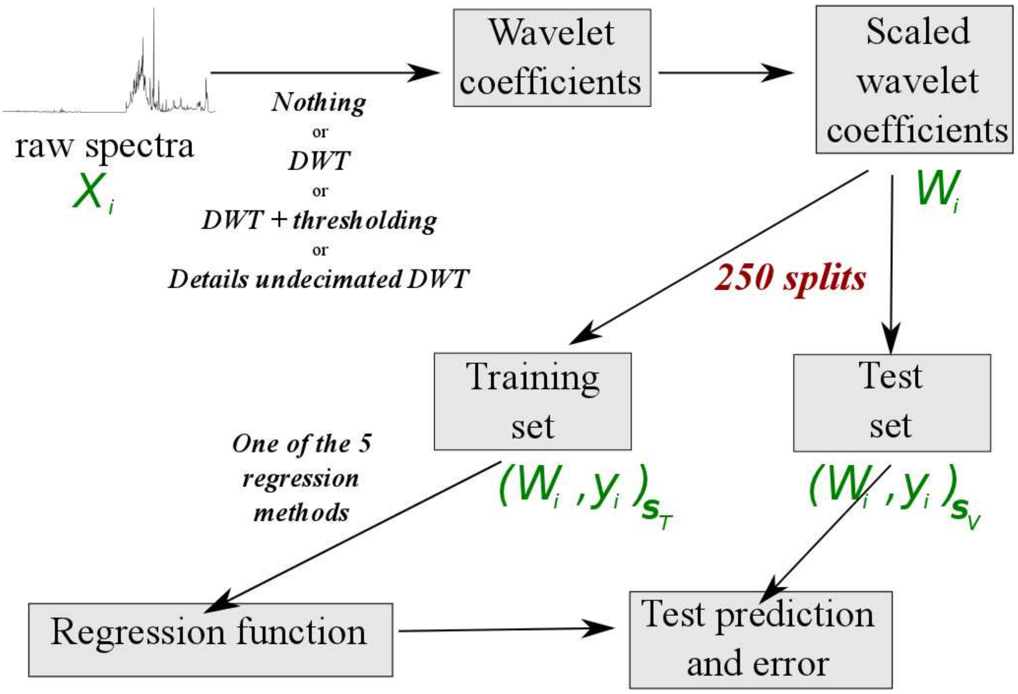

Figure 2: Illustration of the methodology used to compare various combinations of wavelet transforms and regression methods

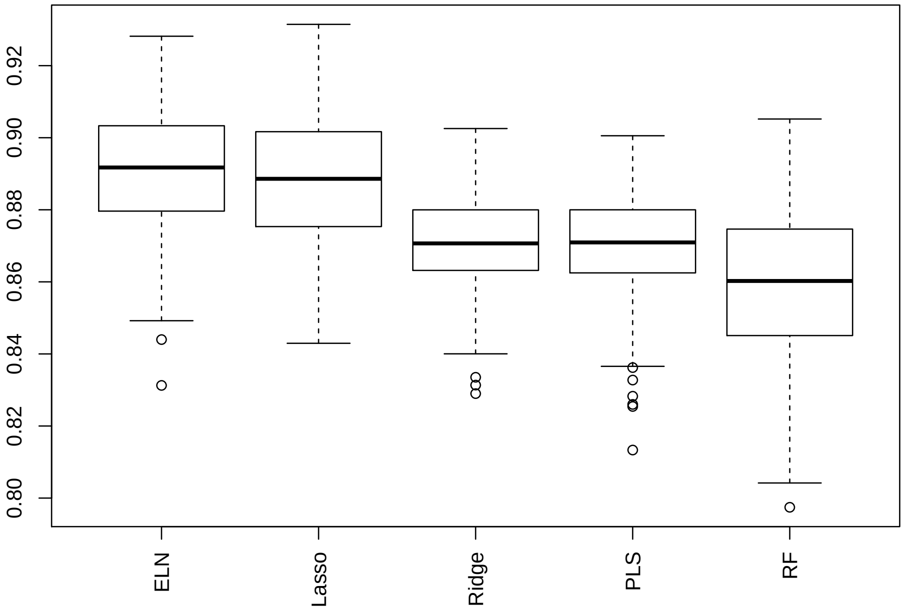

Figure 3: Boxplots of the $R^2$ of the mean square errors over the 250 test sets for the prediction of the total dose of HR ingested with various learning methods and a full representation with D4 wavelets.

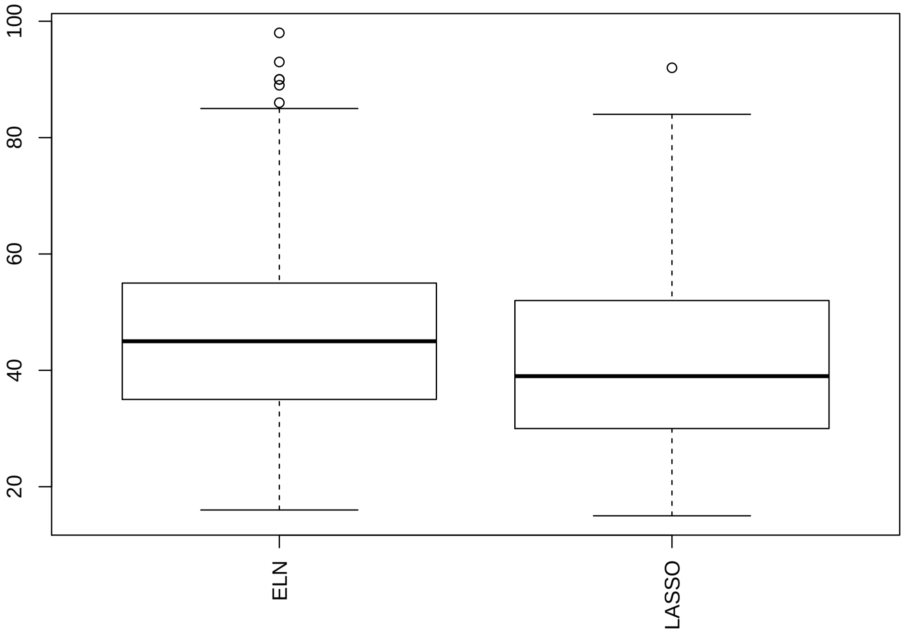

Figure 4: Number of wavelet coefficients selected by elasticnet (ELN) and LASSO over the 250 train sets for D4 wavelet expansion using all the coefficients

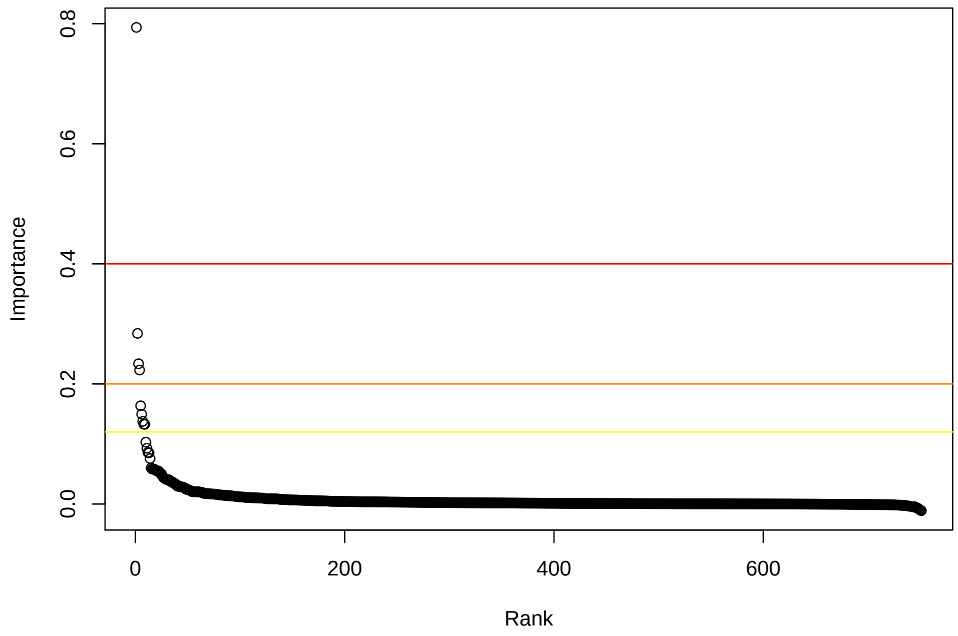

Figure 5: Importance of the 751 spectra locations ranked by decreasing value. The horizontal lines separate increasing degrees of importance from above the red line (very important) to below the yellow line (not important).

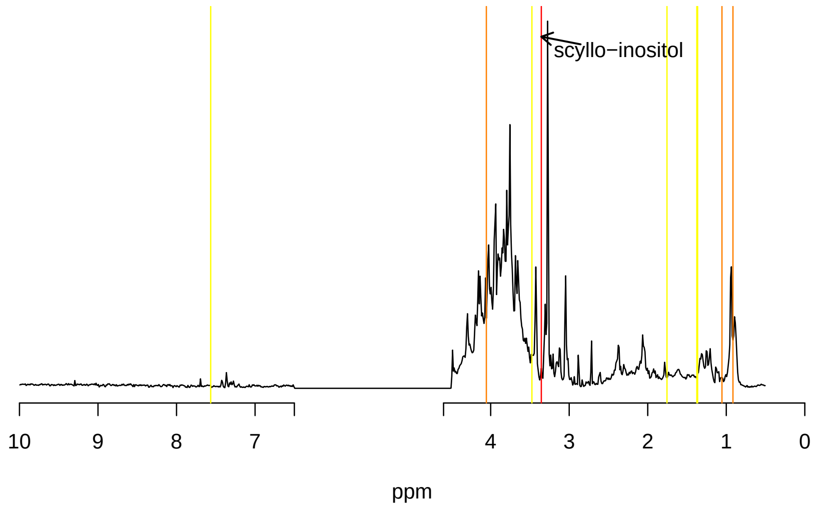

Figure 6: “Most important” metabolites locations on the $^1$H NMR spectra for the prediction of the total dose of HR ingested by the mouse. The colors correspond to those of Figure 5.

# List of Tables

| **DWT** | **Wavelet basis** | **Regression method** |
|---|---|---|
| raw spectra | $\bowtie$ | ELN (elasticnet) |
| full wavelets | Haar | ELN |
| full wavelets | D4 | ELN |
| undecimated wavelets | D4 | ELN |
| thresholded wavelets | D4 | ELN |
| full wavelets | D4 | LASSO |
| full wavelets | D4 | Ridge |
| full wavelets | D4 | PLS |
| full wavelets | D4 | RF |

Table 1: Approaches (wavelet transform and pre-processing combined with a regression method) compared to predict the total HR ingestion from the metabolomic profiles.

| Wavelet transform | Regression method | average RMSE | sd RMSE |
|---|---|---|---|
| Raw spectra | ELN | 16.3 | 1.0 |
| full wavelets (D4) | ELN | **14.3** | 1.1 |
| undecimated wavelets (D4) | ELN | 15.4 | 0.9 |
| thresholded wavelets (D4) | ELN | 42.9 | 52.3 |
| full wavelets (Haar) | ELN | 14.5 | 1.0 |
| full wavelets (D4) | LASSO | 14.5 | 1.1 |
| full wavelets (D4) | Ridge | 15.6 | 0.7 |
| full wavelets (D4) | PLS | 15.6 | 0.9 |
| full wavelets (D4) | RF | 16.2 | 1.2 |

Table 2: Means and standard deviations of root mean squared errors for the prediction of the total dose of HR ingested with various combinations of wavelet transforms and regression methods. “ELN” means “elasticnet”; “Ridge” means “ridge regression”; “RF” means “random forest”; “D4” means “Daubechies 4 wavelet basis” and “Haar” means “Haar wavelet basis”. Bold capitals are used to emphasize to the best method among all experiments.

| ppm | Importance | Metabolites | Change with HR |
|---|---|---|---|
| 3.35 | 79.4% | *scyllo*-inositol | ↗ |
| 4.05 | 28.4% | creatinine | ↘ |
| 1.05 | 23.4% | valine | ↗ |
| 0.91 | 22.3% | unassigned | ↘ |
| 1.37 | 16.4% | unassigned | ↗ |
| 1.36 | 15.0% | unassigned | ↗ |
| 7.56 | 13.8% | hippurate | ↗ |
| 3.47 | 13.3% | unassigned | ↘ |
| 1.75 | 13.3% | unassigned | ↗ |

Table 3: Summary of the "most important" peaks (and, if known, metabolites) for the prediction of the total dose of HR ingested by the mouse.