



**HAL**  
open science

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

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

### ► To cite this version:

Nathalie Vialaneix, Noslen Hernández, Alain Paris, Céline Domange, Nathalie Priymenko, et al.. On combining wavelets expansion and sparse linear models for Regression on metabolomic data and biomarker selection. *Communications in Statistics - Simulation and Computation*, 2016, 45 (1), pp.282-298. 10.1080/03610918.2013.862273 . hal-01270963

**HAL Id: hal-01270963**

**<https://hal.science/hal-01270963>**

Submitted on 8 Feb 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Nathalie Villa-Vialaneix<sup>1,2\*</sup>, Noslen Hernández<sup>3</sup>, Alain Paris<sup>4</sup>,  
Céline Domange<sup>5,6</sup>, Nathalie Priymenko<sup>7</sup>, Philippe Besse<sup>8</sup>

<sup>1</sup> SAMM, Université Paris 1, 90 rue de Tolbiac, F-75013 Paris - France

<sup>2</sup> Université de Perpignan Via Domitia, IUT, Dpt STID, F-66860 Perpignan - France

<sup>3</sup> Advanced Technologies Application Center, CENATAV, Havana - Cuba

<sup>4</sup> INRA, Unité Mét@risk, AgroParisTech, 16 rue Claude Bernard, F-75005 Paris - France

<sup>5</sup> AgroParisTech, UMR 0791 Modélisation Systémique Appliquée aux Ruminants,  
F-75005 Paris - France

<sup>6</sup> INRA, UMR 0791 Modélisation Systémique Appliquée aux Ruminants, 16 rue Claude  
Bernard, F-75005 Paris - France

<sup>7</sup> ENVT, INRA, UMR 1089, Université de Toulouse, F-31076 Toulouse - France

<sup>8</sup> 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

23 data, a commonly used basis for projecting the data is the Wavelet Trans-  
24 form (WT) [Mallat, 1999]. Wavelet expansion is frequently performed  
25 to correct the baseline and to de-noise the data by removing the small-  
26 est details with a thresholding method. Then, in a second phase, a re-  
27 gression or a classification method is applied on the thresholded signal  
28 [Xia et al., 2007, Alexandrov et al., 2009]. On the other hand, selection  
29 methods select a small number of variables among the original ones to ensure  
30 an easy interpretation, often at the cost of deteriorated prediction perfor-  
31 mances: as an example, [Wongravee et al., 2009] used a bootstrap approach  
32 and PLS-DA to select variables in a large metabolomic dataset prior a clas-  
33 sification. Finally, projection and variable selection are sometimes combined  
34 as in [Alsberg et al., 1998a, Kim et al., 2008].

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

43 A one phase approach is tested that performs a sparse or a regularized  
44 regression method on the wavelet coefficients resulting from the wavelet rep-  
45 resentation of the spectra. Contrary to thresholding methods, where the

46 coefficients selection is not directly related to the prediction of the target  
47 variable, the introduced approach automatically selects the most relevant  
48 wavelet coefficients in relation to the target variable. The relevance of the  
49 proposal is assessed through a case study. The purpose is to recover the drug  
50 dose ingested by a mouse from its metabolomic profile, in order to prevent a  
51 possible illness. A comparison study is performed on this real world problem,  
52 that leads to several conclusions: first, as was expected, wavelet transform is  
53 well adapted to the representation of metabolomic data and leads to better  
54 predictive performances. Then, variable selection by a blind thresholding  
55 of the wavelet coefficients deteriorates the predictions contrary to a variable  
56 selection performed by means of a sparse approach. This last method leads  
57 to the most accurate prediction performances.

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

## 66 2 Case study and material

### 67 2.1 Problem description

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

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

85 The issue of interest is to predict the total dose of HR ingested, which is  
86 the daily HR dose multiplied by the number of days of ingestion, from the

87 metabolomic data. This problem can be written as a regression problem:

$$y_i = \Phi(X_i) + \epsilon_i \quad (1)$$

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

- 91 • the first motivation is to know if the metabolomic profile alone is enough  
92 to predict the drug dose ingested by an animal, which can be useful to  
93 prevent an illness;
- 94 • conversely, the second motivation is to understand if the influence of  
95 the HR dose ingestion is strong enough not to be seen as an artifact: if  
96  $y_i$  can be accurately estimated from  $X_i$  then this is a strong indication  
97 that the HR dose and more precisely, its cumulative effect, is really  
98 disrupting the mouse metabolomic profile;
- 99 • finally the last motivation is to use the estimated regression function  
100 to corroborate a set of relevant metabolites influenced by the HR in-  
101 gestion. The chosen approach is to extract the explanatory variables  
102 (i.e., the part of the metabolomic profiles) with the strongest predictive  
103 power, from the estimated regression function.



## 104 **2.2 Data pre-processing**

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

110 [Figure 1 about here.]

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

125 In the remaining, the obtained wavelet coefficients are denoted by

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

## 129 **3 Methodological proposal**

### 130 **3.1 State-of-the-art on using DWT in regression prob-** 131 **lems**

132 Wavelet transforms are often applied to signals as a pre-processing step be-  
133 fore the statistical analysis [Davis et al., 2007, Xia et al., 2007]. A threshold-  
134 ing approach on the discrete wavelet transform is then generally performed  
135 in order to remove the smallest (and most irrelevant) detailed coefficients  
136 from the spectra representation. Standard thresholding strategies are the  
137 so-called “hard thresholding” that simply removes the smallest coefficients  
138 and leaves the others unchanged and the “soft thresholding” that removes  
139 the coefficients smaller than a given threshold and reduces the others from  
140 the value of this threshold. Of course, the choice of the threshold is very  
141 important and several solutions have been proposed: for instance, the SURE  
142 and Universal policies are calculated from an estimation of the level of noise  
143 and justified by asymptotic properties (see [Donoho and Johnstone, 1994,  
144 Donoho, 1995, Donoho and Johnstone, 1995, Donoho et al., 1995]). Also,  
145 [Nason, 1996] suggests to use a cross-validation criterion to choose the thresh-

146 old and [Johnstone and Silverman, 1997] to rely on a different threshold for  
147 each level. More recently, [González et al., 2013] shows that keeping solely  
148 the finest details coefficients at the lowest decomposition level produces a  
149 representation of the data having the ability to correct a putative baseline  
150 default.

151 A natural approach to predict a phenotype from metabolomic profiles  
152 expressed in the wavelet domain would then be to perform a threshold-  
153 ing prior to the application of a well chosen regression method (see, e.g.,  
154 [Xia et al., 2007]). But this methodology does not link the wavelet coef-  
155 ficients selection to the prediction purpose. An alternative solution is to  
156 perform a variable selection method, that takes into account the target vari-  
157 able, before learning the regression or the classification function. In this  
158 direction, [Alexandrov et al., 2009] uses a multiple testing approach with  
159 a Benjamini & Hochberg adjustment to select the relevant wavelet coef-  
160 ficients in relation to a target factor variable before building a classifica-  
161 tion model (based on SVM) to predict it. [Saito et al., 2002] proposes to  
162 select the wavelet coefficients that maximize the Kullback-Leibler diver-  
163 gence between estimated densities obtained for the various levels of a fac-  
164 tor target variable before learning a classification function on the basis of  
165 the selected coefficients. Also, [Jouan-Rimbaud et al., 1997] uses a “Rele-  
166 vant Component Extraction” that thresholds the less informative wavelet  
167 coefficients from a PLS between the spectra and a target variable of inter-  
168 est. These latter approaches explicitly focused on wavelet coefficients se-

169 lection but any feature selection method is expendable for such a task (see  
170 [Liu and Motoda, 1998, Guyon and Elisseeff, 2003] for reviews about feature  
171 selection). Feature selection algorithms can be time consuming and it has  
172 also been pointed out in [Raudys, 2006] that they can lead to feature *over*-  
173 selection that hinders the prediction performances.

174 Another approach is to simultaneously select the variables and opti-  
175 mize the prediction error: [Alsberg et al., 1998b] select the wavelet co-  
176 efficients that minimize the cross validation error of a PLS regression.  
177 Model selection methods penalize the prediction error with a quantity de-  
178 pending on the number of variables involved in the regression (see, i.e.,  
179 [Biau et al., 2005, Rossi and Villa, 2006] for examples in a similar framework  
180 where the signal is projected onto an orthogonal basis for classification pur-  
181 pose where the data are functions). However, model selection requires the  
182 definition of a relevant penalty term that can be hard to choose effectively,  
183 as pointed out in [Fromont and Tuleau, 2006].

## 184 **3.2 A sparse one-phase approach**

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

190 Our proposal is to use penalized regression methods to simultaneously

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

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

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

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

- 202 • if  $p(\beta) = \|\beta\|_{L^1} = \sum_{j=1}^D |\beta_j|$ , the linear regression is a sparse linear  
 203 regression also named LASSO [Tibshirani, 1996]. It selects wavelet  
 204 coefficients, in the set of  $D$  original coefficients, in a optimal way for  
 205 prediction purpose;
- 206 • if  $p(\beta) = \|\beta\|_{\mathbb{R}^D}^2$ , the linear regression is a ridge regression which tends  
 207 to produce  $\beta$  with small norms but does not perform a selection of the  
 208 wavelet coefficients;
- 209 • if  $p(\beta) = (1 - \alpha)\|\beta\|_{\mathbb{R}^D}^2 + \alpha\|\beta\|_{L^1}$ ,  $\alpha \in ]0, 1[$  the linear regression is

210 the so-called “elasticnet” method [Zou and Hastie, 2005], proposed in  
211 an attempt to use the advantages of the two previous penalties. As  
212 LASSO, it selects a reduced number of wavelet coefficients involved in  
213 the regression function but this number is usually larger than the one  
214 obtained when using the LASSO method.

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

### 224 **3.3 Comparison methodology**

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

- 228 • the possible wavelet approximations applied to the pre-processed data  
229 (as described in Section 2.2) are raw spectra (no wavelet approxima-  
230 tion), wavelet coefficients (Haar or D4 bases), thresholded wavelet co-

231       efficients (D4), undecimated wavelet detailed coefficients (D4).  
232       “*thresholded wavelet coefficients*” correspond to the wavelet coefficients  
233       that remain positive after a soft threshold with SURE policy and “*un-*  
234       *decimated wavelet detailed coefficients*” correspond to the union of the  
235       finest details coefficients of the original spectra with the finest de-  
236       tails coefficients of the shifted spectra (obtained using the approach  
237       of [Beylkin, 1992, González et al., 2013]). When using the full wavelet  
238       decomposition or the “undecimated wavelet” approach, the dimension-  
239       ality of the original problem,  $D = 751$  is left unchanged whereas the  
240       “thresholded wavelet” approach leads to a dimensionality reduction  
241       ( $D = 71$  for D4 DWT), which is a standard way to handle large dimen-  
242       sion regression tasks.

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

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

- 250       • any wavelet approximation is combined with the elasticnet regression.  
251       Our proposal is to use the full wavelet decomposition (without thresh-  
252       olding) with a sparse regression method. To enlighten the uselessness of

253 the thresholding when using a sparse regression method, thresholding  
254 is also combined with elasticnet in the comparison;

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

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

258 [Table 1 about here.]

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

261 **Wavelet transform** First, the data are or are not preprocessed by a DWT.

262 The obtained coefficients are also scaled (each coefficient is centered to  
263 a zero mean and scaled to a standard deviation equal to 1).

264 **Split** The observations (i.e., the pairs  $(W_i, y_i)_i$ ) are randomly split into a  
265 training set  $\mathcal{S}_T$  and a test set  $\mathcal{S}_V$  with balanced sizes (approximately  
266 200 observations each) taking into account the proportion of observa-  
267 tions in the groups defined by sex, dose (including the control group  
268 to train the regression function so that it can predict when the animal  
269 is not affected by HR ingestion) and day of measure. To estimate the  
270 methods variability, this step is repeated 250 times giving 250 training  
271 sets and the corresponding test sets.

272 **Train** The regression method is then applied to each training set. Several  
273 methods involve hyper-parameters that have to be tuned: for random



274 forest, the hyper-parameters are the number of trees, the number of  
275 variables selected for a given split, ... They are set to the default val-  
276 ues, coming from useful heuristics; the stabilization of the out-of-bag  
277 error is achieved using that strategy.

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

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

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

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

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

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

294 transform and regression algorithm.

295 [Figure 2 about here.]

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

## 302 4 Results and discussion

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

## 315 4.1 Numerical performances comparison

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

318 [Table 2 about here.]

319 In addition, the boxplot of the  $R^2$  over the 250 test sets<sup>1</sup> are given in  
320 Figure 3 for the case where the data are expanded on the D4 basis and  
321 where all wavelet coefficients are kept.

322 [Figure 3 about here.]

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

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

---

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

### 335 4.1.1 Comparison of the wavelet transforms

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

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

349 To assess the significance of these conclusions, paired t-test were com-  
350 puted to compare the RMSE of the various wavelet transforms: the differ-  
351 ences between the use of Haar or D4 wavelets are not significant (at level  
352 1%) but the differences between the use of all D4 wavelet coefficients and the  
353 use of either the raw spectra, the D4 undecimated wavelet approach or the  
354 D4 thresholded coefficients are all significant. Note that, even if the differ-  
355 ences between the averaged RMSE seem to be small, they are calculated over  
356 250 replica which is a large enough number to provide confidence in these  
357 conclusions.

#### 358 4.1.2 Comparison of the regression methods

359 Comparing the regression methods, those that are (at least partially) based  
360 on a sparse regularization, such as elasticnet and LASSO, obtain the best  
361 results. Ridge regression is not as accurate as the methods based on a sparse  
362 regularization but its variability is lower. Actually, combining a ridge and a  
363 sparse penalty in the elasticnet seems to slightly decrease the variability of  
364 the elasticnet results compared to those of the LASSO (except for two outlier  
365 samples). Moreover, the influence of the mixing parameter  $\alpha$  is not really  
366 strong: test errors for elasticnet with  $\alpha = 0.1$ ,  $0.25$  or  $0.75$  are not shown in  
367 the paper but would have mostly lead to the same conclusion:  $\alpha = 0.1$  or  
368  $0.25$  has slightly deteriorated (but comparable) test errors, whereas  $\alpha = 0.75$   
369 has test errors closer to the LASSO.

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

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

381 to handle regression problems where the predictors are metabolomic data.

### 382 **4.1.3 Number of selected wavelet coefficients**

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

390 [Figure 4 about here.]

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

## 397 **4.2 Important biomarkers extraction**

398 The relevance of the application of elasticnet on all the wavelets coefficients  
399 is assessed by using the learned regression function, obtained in the previous  
400 section, in order to extract the most important features related to the total

401 dose of HR ingested. A natural approach would be to directly analyze the  
402 variables selected by the sparse regression but, because of the wavelet trans-  
403 form preprocessing, these are not directly linked to the spectra locations that  
404 are of interest.

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

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

#### 422 4.2.1 A measure of the importance of the variables

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

433 We propose to use a similar approach to describe the way a wrong value for  
434 a given variable (here a given value in the spectrum) propagates through the  
435 wavelet transform and the regression function and affects the accuracy of the  
436 final prediction of the total dose of HR ingested by the mouse. This  
437 analysis is focused on the best regression approach, i.e., the use of all wavelet  
438 coefficients coming from a D4 basis expansion combined with elasticnet. As in  
439 the approach proposed in [Breiman, 2001], the importance is calculated from  
440 observations that are not used during the training process. More precisely,  
441 the 250 test samples described in Section 3.3 are used to calculate importance  
442 measures: the “importance” of a variable is the mean rate (over the test  
443 sets) of MSE increase after a random permutation of its values among the  
444 individuals (the other variables remaining with their true values). The idea



445 is to assess the prediction power of a variable by means of the prediction  
 446 accuracy disruption when this variable is given false values. The process is  
 447 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

---

**Algorithm 1** Variables importance calculation

---

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

448

449 quality of the prediction of the total dose of HR ingested. It thus gives an  
 450 assessment to the most relevant features in metabolomic spectra (i.e., the  
 451 features that contribute the most to an accurate prediction of the HR dose),  
 452 despite the series of transformations done.

453 **4.2.2 Results of the biomarkers extraction and comments**

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

456

[Figure 5 about here.]

457 One variable is clearly much more important than all the other ones because  
458 random permutations of its values cause an increase of almost 80% in MSE.  
459 Three other variables seem to be important (with importance greater than  
460 20%) and another group of 5 variables are also important to a lesser extent  
461 (between the yellow line and the orange line in Figure 5).

462 The list of the “most important” spectra locations and the names of the  
463 associated metabolites (when it is known) are given in Table 3. Moreover,  
464 the location of these metabolites in a  $^1\text{H}$  NMR spectrum is shown in Figure 6.

465

[Table 3 about here.]

466

[Figure 6 about here.]

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

476 Also, some differences arise when comparing this list with the list  
477 of biomarkers identified in [Domange et al., 2008]. Part of these dif-

478 ferences may be explained by the fact that the dependent variable in  
479 [Domange et al., 2008] is the daily HR dose ingested (i.e., a factor variable  
480 with 3 levels) whereas, here, the total ingested dose was used in order to take  
481 into account the cumulative effect of the ingestion. But it is also the positive  
482 counterpart of not using a test approach and thus avoiding the standard false  
483 positive issue that comes with them. As the extracted spectra locations are  
484 directly related to the quality of the prediction, they are more reliable, even  
485 if not so well theoretically justified.

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

## 492 **5 Conclusion**

493 Wavelet transformation is commonly used to deal with spectrometric data in  
494 biology, especially for de-noising purposes. Moreover, this paper shows that,  
495 associated with a convenient learning method, it improves the understanding  
496 of the relation between metabolomic spectrum and a phenomenon of interest  
497 (for instance, metabolic disruptions linked to HR ingestion). It is also shown  
498 that using a de-noising approach, not related to the variable to be predicted,

499 can lead to a dramatic loss of information. More precisely, some important  
500 variables seem to be located in parts of the spectra that could be seen as  
501 “minor” details. It is thus important to combine the wavelet transform and  
502 de-noising with the purpose of the study. Sparse methods, that combine  
503 a regression model and a variable selection seem to be well suited to this  
504 task: they perform a kind of joint thresholding of the wavelet coefficients  
505 that is directly related to the target variable. In particular, elasticnet gave  
506 the best performance in prediction and was also able to provide a relevant  
507 list of biomarkers, linked to the target variable, in our case study.

508 In conclusion, the combination of DWT with elasticnet can be used to ac-  
509 curately predict a numerical variable of interest from the metabolomic profile.  
510 It is also useful to identify and confirm the most important features involved  
511 in the biological process under study thanks to the importance measure in-  
512 troduced in this article.

## 513 **6 Acknowledgements**

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

## 516 **References**

- 517 [Alexandrov et al., 2009] Alexandrov, T., Decker, J., Mertens, B., Deelder,  
518 A., Tollenaar, R., Maass, P., and Thiele, H. (2009). Biomarker discovery  
519 in maldi-tof serum protein profiles using discrete wavelet transformation.  
520 *Bioinformatics*, 25(5):643–649.
- 521 [Alsberg et al., 1998a] Alsberg, B., Kell, D., and Goodacre, R. (1998a). Vari-  
522 able selection in discriminant partial least-squares analysis. *Analytical*  
523 *Chemistry*, 70:4126–4133.
- 524 [Alsberg et al., 1998b] Alsberg, B., Woodward, A., Winson, M., Rowland,  
525 J., and Kell, D. (1998b). Variable selection in wavelet regression models.  
526 *Analytica Chimica Acta*, 368:29–44.
- 527 [Archer and Kimes, 2008] Archer, K. and Kimes, R. (2008). Empirical char-  
528 acterization of random forest variable importance measures. *Computa-*  
529 *tional Statistics and Data Analysis*, 52:2249–2260.
- 530 [Beylkin, 1992] Beylkin, G. (1992). On the representation of operators in  
531 bases of compactly supported wavelets. *SIAM Journal on Numerical Anal-*  
532 *ysis*, 29:1716–1740.
- 533 [Biau et al., 2005] Biau, G., Bunea, F., and Wegkamp, M. (2005). Functional  
534 classification in Hilbert spaces. *IEEE Transactions on Information Theory*,  
535 51:2163–2172.

- 536 [Bolstad et al., 2003] Bolstad, B., Irizarry, R., Astrand, M., and Speed, T.  
537 (2003). A comparison of normalization methods for high density oligonu-  
538 cleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–  
539 193.
- 540 [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*,  
541 45(1):5–32.
- 542 [Cox and Cox, 2001] Cox, T. and Cox, M. (2001). *Multidimensional Scaling*.  
543 Chapman and Hall.
- 544 [Davis et al., 2007] Davis, R., Charlton, A., Godward, J., Jones, S., Harri-  
545 son, M., and J.C., W. (2007). Adaptive binning: an improved binning  
546 method for metabolomics data using the undecimated wavelet transform.  
547 *Chemometrics and Intelligent Laboratory Systems*, 85:144–154.
- 548 [Domange et al., 2008] Domange, C., Canlet, C., Traoré, A., Biélicki,  
549 G. Keller, C., Paris, A., and Priymenko, N. (2008). Orthologous metabo-  
550 nomic qualification of a rodent model combined with magnetic resonance  
551 imaging for an integrated evaluation of the toxicity of hypochoeris radi-  
552 cata. *Chemical Research in Toxicology*, 21(11):2082–2096.
- 553 [Domange et al., 2010] Domange, C., Casteignau, A., Pumarola, M., and  
554 Priymenko, N. (2010). Longitudinal study of australian stringhalt cases in  
555 France. *Journal of Animal Physiology and Animal Nutrition*, 94(6):712–  
556 720.

- 557 [Donoho, 1995] Donoho, D. (1995). De-noising by soft-thresholding. *IEEE*  
558 *Transactions on Information Theory*, 41(3):613–627.
- 559 [Donoho and Johnstone, 1994] Donoho, D. and Johnstone, I. (1994). Ideal  
560 spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.
- 561 [Donoho and Johnstone, 1995] Donoho, D. and Johnstone, I. (1995). Adapt-  
562 ing to unknown smoothness via wavelet shrinkage. *Journal of the American*  
563 *Statistical Association*, 90(432):1200–1224.
- 564 [Donoho et al., 1995] Donoho, D., Johnstone, I., Kerkyacharian, G., and Pi-  
565 card, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal*  
566 *Statistical Society. Series B*, 57(2):301–369.
- 567 [Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.  
568 (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.
- 569 [Fromont and Tuleau, 2006] Fromont, M. and Tuleau, C. (2006). Functional  
570 classification with margin conditions. In *Proceedings of the 19th Annual*  
571 *Conference on Learning Theory*, volume 4005 of *Lecture Notes in Com-*  
572 *puter Science*, pages 94–108.
- 573 [Genuer et al., 2010] Genuer, R., Poggi, J., and Tuleau-Malot, C. (2010).  
574 Variable selection using random forests. *Pattern Recognition Letters*,  
575 31:2225–2236.
- 576 [González et al., 2013] González, I., Eveillard, A., Canlet, A., Paris, T.,  
577 Pineau, P., Besse, P., Martin, P., and Déjean, S. (2013). Undeci-

578 mated wavelet transform to improve the classification of samples from  
579 metabolomic data. *JP Journal of Biostatistics*. Forthcoming.

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

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

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

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

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

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



- 599 [Liu and Motoda, 1998] Liu, H. and Motoda, H. (1998). *Feature Selection*  
600 *for Knowledge Discovery and Data Mining*, volume 454 of *The Springer*  
601 *Series in Engineering and Computer Science*. Springer, Nowell, MA, USA.
- 602 [Mallat, 1999] Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Aca-  
603 demic Press.
- 604 [Nason, 1996] Nason, G. (1996). Wavelet shrinkage using cross-validation.  
605 *Journal of the Royal Statistical Society. Series B. Statistical Methodology*,  
606 58:463–479.
- 607 [Nason and Silverman, 1994] Nason, G. and Silverman, B. (1994). The dis-  
608 crete wavelet transform in S. *Journal of Computational and Graphical*  
609 *Statistics*, 3:163–191.
- 610 [R Development Core Team, 2012] R Development Core Team (2012). *R: A*  
611 *Language and Environment for Statistical Computing*. Vienna, Austria.  
612 ISBN 3-900051-07-0.
- 613 [Ramsay and Silverman, 1997] Ramsay, J. and Silverman, B. (1997). *Func-*  
614 *tional Data Analysis*. Springer Verlag, New York.
- 615 [Raudys, 2006] Raudys, S. (2006). *Structural, Syntactic and Statistical*  
616 *Pattern Recognition*, volume 4109 of *Lecture Notes in Computer Sci-*  
617 *ence*, chapter Feature over-selection, pages 622–631. Springer-Verlag,  
618 Berlin/Heidelberg, Germany.

- 619 [Rohart et al., 2012] Rohart, F., Paris, A., Laurent, B., Canlet, C., Molina,  
620 J., Mercat, M., Tribout, T., Muller, N., Iannuccelli, N., Villa-Vialaneix,  
621 N., Liaubet, L., Milan, D., and San Cristobal, M. (2012). Phenotypic  
622 prediction based on metabolomic data on the growing pig from three main  
623 European breeds. *Journal of Animal Science*, 90(12).
- 624 [Rossi and Villa, 2006] Rossi, F. and Villa, N. (2006). Support vector ma-  
625 chine for functional data classification. *Neurocomputing*, 69(7-9):730–742.
- 626 [Saito et al., 2002] Saito, N., Coifman, R., Geshwind, F., and Warner, F.  
627 (2002). Discriminant feature extraction using empirical probability density  
628 estimation and a local basis library. *Pattern Recognition*, 35:2841–2852.
- 629 [Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection  
630 via the lasso. *Journal of the Royal Statistical Society, series B*, 58(1):267–  
631 288.
- 632 [Wold, 1975] Wold, H. (1975). *Soft modeling by latent variables; the non-*  
633 *linear iterative partial least square approach*. J. Gani, Academic Press,  
634 London.
- 635 [Wongravee et al., 2009] Wongravee, K., Heinrich, N., Holmboe, M., Schae-  
636 fer, M., Reed, R., Trevejo, J., and Brereton, R. (2009). Variable selection  
637 using iterative reformulation of training set models for discrimination of  
638 samples: application to gas chromatography/mass spectrometry of mouse  
639 urinary metabolites. *Analytical Chemistry*, 81(13):5204–5217.

- 640 [Xia et al., 2007] Xia, J., Wu, X., and Yuan, X. (2007). Integration of  
641 wavelet transform with PCA and ANN for metabolomics data-mining.  
642 *Metabolomics*, 3(4):531–537.
- 643 [Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and  
644 variable selection via the elastic net. *Journal of the Royal Statistical Soci-*  
645 *ety, series B*, 67(2):301–320.

646 **List of Figures**

647 1 An example of metabolomic spectra from data discussed in  
648 Section 2 (female mice of the control group at day 0). . . . . 36

649 2 Illustration of the methodology used to compare various com-  
650 binations of wavelet transforms and regression methods . . . . . 37

651 3 Boxplots of the  $R^2$  of the mean square errors over the 250 test  
652 sets for the prediction of the total dose of HR ingested with  
653 various learning methods and a full representation with D4  
654 wavelets. . . . . 38

655 4 Number of wavelet coefficients selected by elasticnet (ELN)  
656 and LASSO over the 250 train sets for D4 wavelet expansion  
657 using all the coefficients . . . . . 39

658 5 Importance of the 751 spectra locations ranked by decreasing  
659 value. The horizontal lines separate increasing degrees of im-  
660 portance from above the red line (very important) to below  
661 the yellow line (not important). . . . . 40

662 6 “Most important” metabolites locations on the  $^1\text{H}$  NMR spec-  
663 tra for the prediction of the total dose of HR ingested by the  
664 mouse. The colors correspond to those of Figure 5. . . . . 41

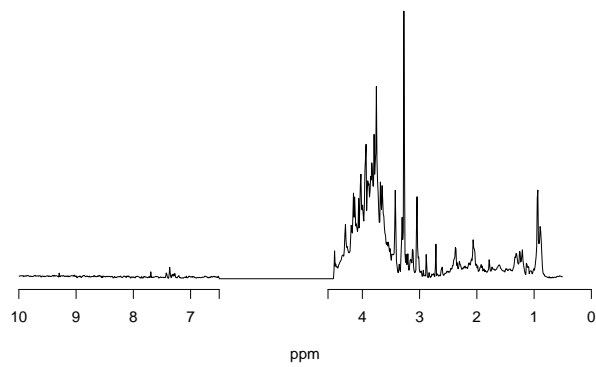


Figure 1: An example of metabolomic spectra from data discussed in Section 2 (female mice of the control group at day 0).

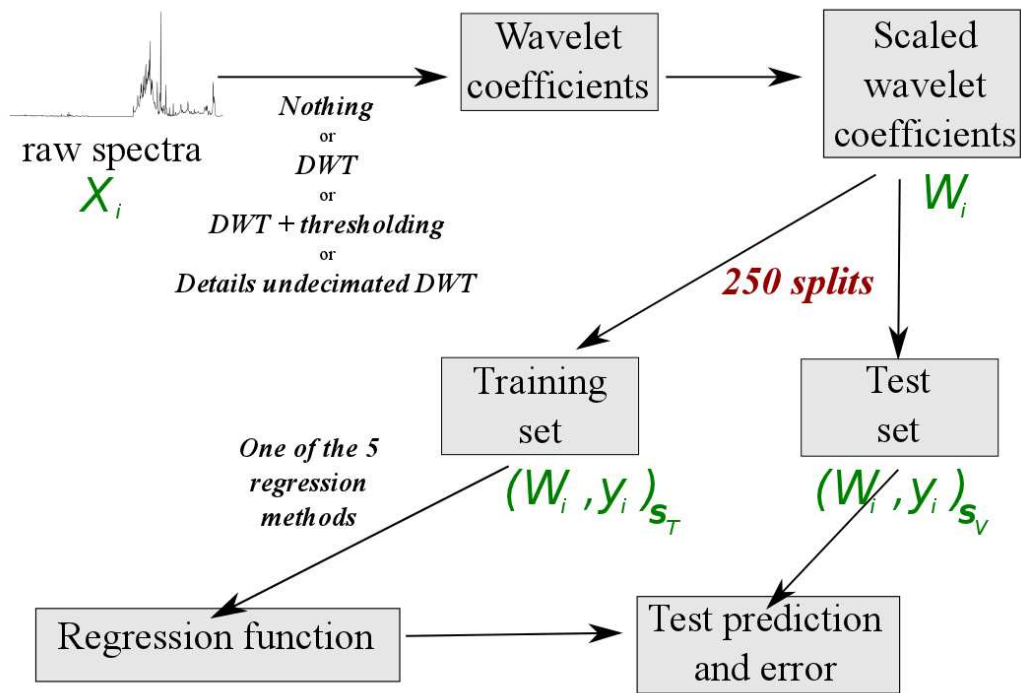


Figure 2: Illustration of the methodology used to compare various combinations of wavelet transforms and regression methods

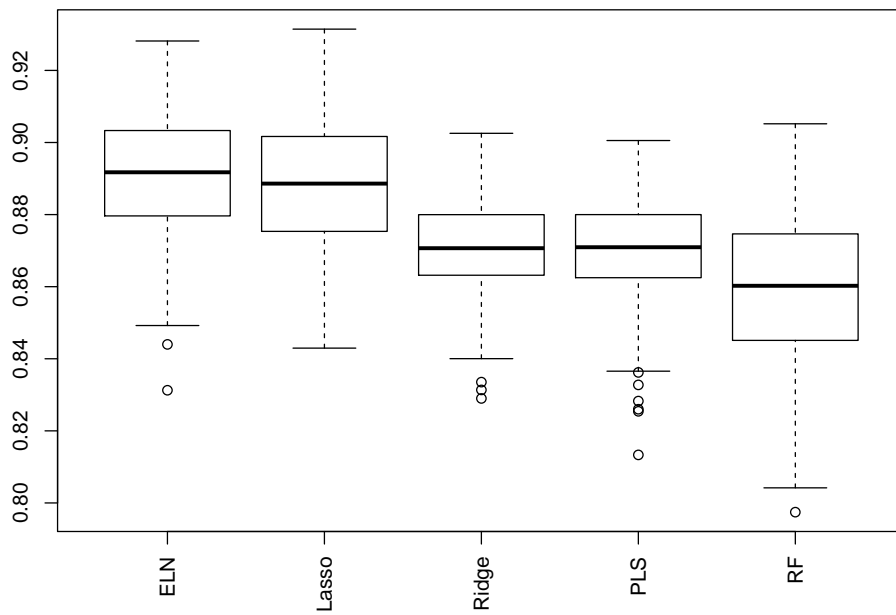


Figure 3: Boxplots of the  $R^2$  of the mean square errors over the 250 test sets for the prediction of the total dose of HR ingested with various learning methods and a full representation with D4 wavelets.

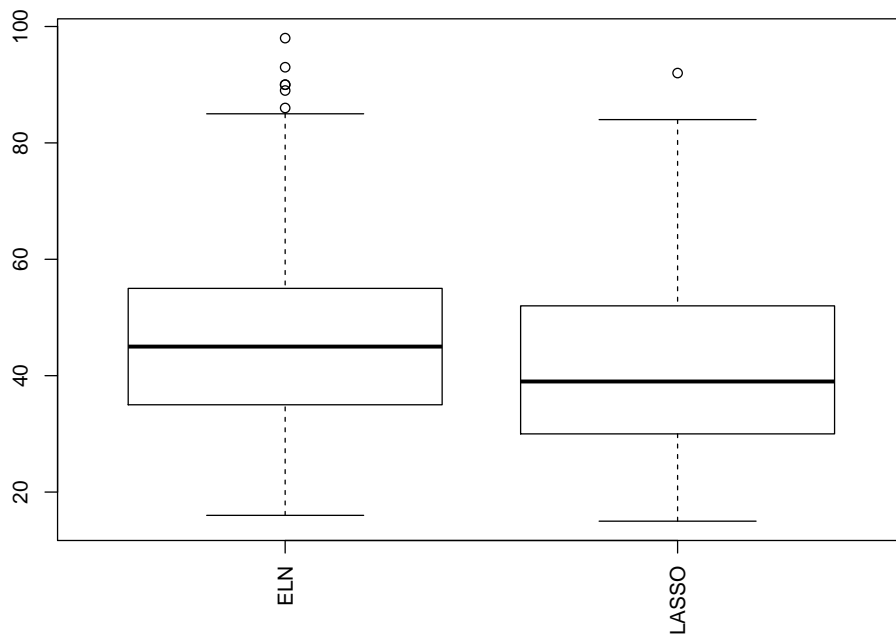


Figure 4: Number of wavelet coefficients selected by elasticnet (ELN) and LASSO over the 250 train sets for D4 wavelet expansion using all the coefficients



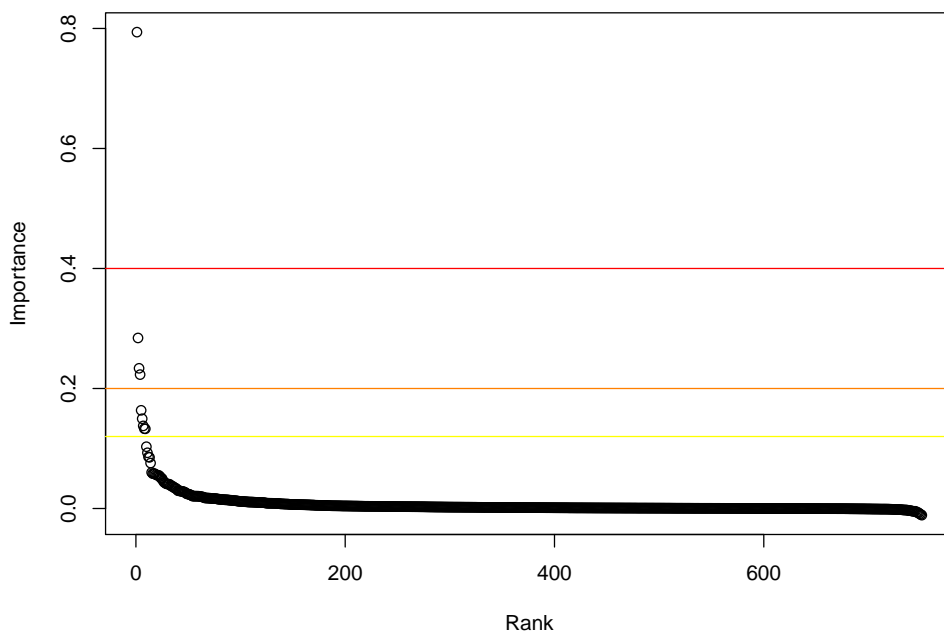


Figure 5: Importance of the 751 spectral locations ranked by decreasing value. The horizontal lines separate increasing degrees of importance from above the red line (very important) to below the yellow line (not important).

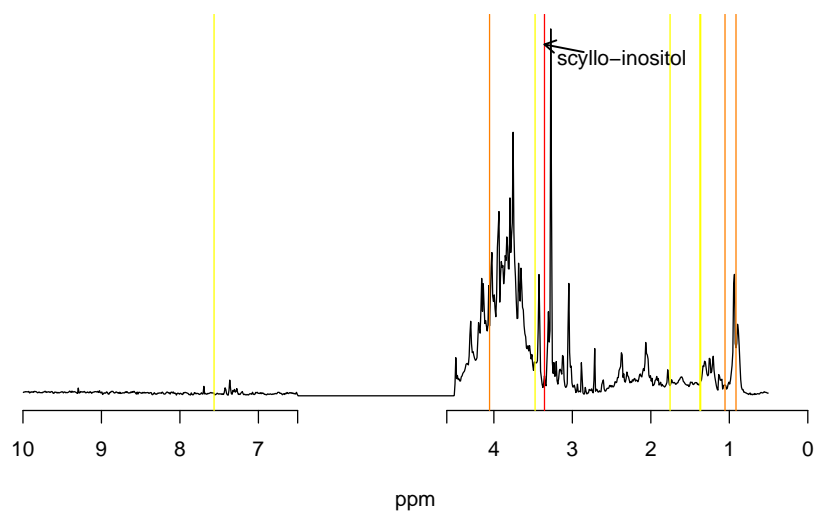


Figure 6: “Most important” metabolites locations on the  $^1\text{H}$  NMR spectra for the prediction of the total dose of HR ingested by the mouse. The colors correspond to those of Figure 5.

665 **List of Tables**

666 1 Approaches (wavelet transform and pre-processing combined  
667 with a regression method) compared to predict the total HR  
668 ingestion from the metabolomic profiles. . . . . 43  
669 2 Means and standard deviations of root mean squared errors  
670 for the prediction of the total dose of HR ingested with vari-  
671 ous combinations of wavelet transforms and regression meth-  
672 ods. “ELN” means “elasticnet”; “Ridge” means “ridge regres-  
673 sion”; “RF” means “random forest”; “D4” means “Daubechies  
674 4 wavelet basis” and “Haar” means “Haar wavelet basis”. Bold  
675 capitals are used to emphasize to the best method among all  
676 experiments. . . . . 44  
677 3 Summary of the “most important” peaks (and, if known,  
678 metabolites) for the prediction of the total dose of HR ingested  
679 by the mouse. . . . . 45

DWT	Wavelet basis	Regression method
raw spectra	$\times$	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	<b>14.3</b>	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%	<i>scyllo</i> -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.