Phenotypic prediction based on metabolomic data: LASSO vs BOLASSO, primary data vs wavelet transformation

F. Rohart[*†], N. Villa-Vialaneix[‡], A. Paris[§], C. Canlet[¶], J. Molina[‡], D. Milan[*],
B. Laurent[†] and M. SanCristobal[*]

Introduction

Understanding the relations between various omics data (such as metabolomics or genomics data) and phenotypes of interest is one of the current major challenges in biology. This question can be addressed by trying to predict the phenotype value from the omic from joint observations of the omic and of the phenotype. In this paper, we focus on the prediction of a phenotype from metabolomic data. Metabolomic data usually are high dimensional data. The number of observations is often reduced, model selection methods are a way both to obtain a relevant solution to the prediction problem but also to select the most important metabolites related to the phenotype under study. In this paper, the number of observations and of metabolomic discretized variables are of the same order; model selection is nevertheless useful to solve the prediction problem.

During the past years, model selection has known a growing interest in the statistical community: the first - and also probably the mostly used - selection method has been introduced by Tibshirani (1996) under the name of LASSO. Several variants of this original approach have then been proposed such as, recently, a bootstraped LASSO, named BOLASSO, introduced by Bach (2009).

The aim of this paper is to combine a wavelet representation of the metabolome spectra (see Mallat (1999) and Antonini et al. (1992) for a complete introduction to wavelets) with the BOLASSO approach. We compare this methodology to more classical methods using either the original spectra as predictors (instead of the wavelet representation) or the original LASSO to select the model.

Material and methods

The purpose is to predict a given phenotype, real-valued, from metabolomic data. As shown in Figure 1, metabolomic data are spectra observed on a discrete sampling grid of size q. To

[*] UMR 444 Laboratoire de Génétique Cellulaire, INRA Toulouse, 31320 Castanet Tolosan cedex, France

[†] INSA, Departement de Génie Mathématique, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France

[‡] Institut de Mathématiques de Toulouse, Université de Toulouse et IUT STID de Carcassonne, Université de Perpignan, France

[§] UR1204 Méthodologies d'Analyses de Risque Alimentaire, AgroParisTech, 16 rue Claude Bernard, 75231 Paris, France

[¶] UMR 1089 Xénobiotiques, INRA Toulouse, BP 93173, 31027 Toulouse cedex 3

learn how the phenotype can be predicted from the metabolomic spectra, n i.i.d. observations of these variables are available. In the following, Y will denote the vector of the n observations of the phenotype to predict and $X = (X^1, \dots, X^q)$ will refer to the matrix of the n discrete observations of the metabolomic data on the sampling grid (hence, each X^i is a vector in $\mathbb{R}^n$). In this paper, we focus on a linear relation between X and Y:

$$Y = X\beta + \epsilon, \tag{1}$$

where $\beta \in \mathbb{R}^q$ are the parameters to estimate and ϵ are i.i.d. Gaussian random variables with variance σ^2.

Wavelet Analysis

The metabolomic data consisted in 508 spectra, each of dimension 375. One of these spectra is plotted in Figure 1.

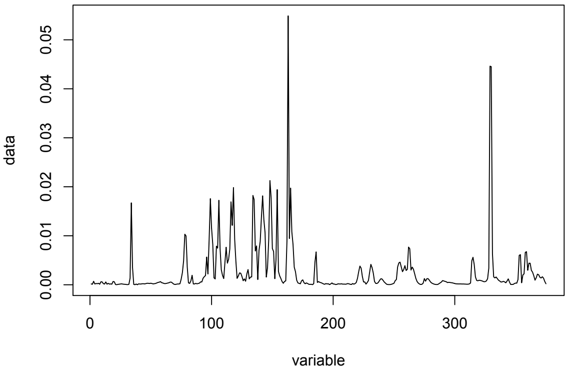

Figure 1: Spectrum of one individual

Following the idea of Villa-Vialaneix and Hernandez-Gonzalez (2009), each spectrum has been decomposed onto a Haar basis. The corresponding wavelet coefficients have been thresholded with a soft-thresholding method (see Mallat (1999) for details). In the following, $\tilde{X}$ will denote the $n \times p$ matrix of thresholded wavelet coefficients where, in general, the number of wavelets p is smaller than the original dimension of the data, q. In the data set described in the next section, p is equal to 367. Finally, note that this preprocessing leads to a change in the original model (1); the new linear model is written as:

$$Y = \tilde{X}\tilde{\beta} + \tilde{\epsilon}. \tag{2}$$

LASSO and BOLASSO

Basically, the LASSO is a penalized least squares approach used to solve ill-posed or badly conditioned linear regressions. The parameter β of (1) is estimated by:

$$\hat{\beta}_{lasso} = \arg\min_{\beta \in \mathbb{R}^q} \|Y - X\beta\|_2^2 + \mu\|\beta\|_1 \tag{3}$$

where $\mu \geq 0$ is the regularization parameter, and $\|.\|_k$ the L^k norm (in a straightforward way, the same method can be applied to find the parameters $\tilde{\beta}$ in Equation (2)). A complete description of the LASSO method for linear regression can be found in Tibshirani (1996).

The great interest of this approach comes from the fact that the solution leads to a restricted number of non zero β_i, this number depending on the value of the regularization parameter. Therefore, LASSO is both a shrinkage and a selection method at the same time. However, Bach (2009) showed that LASSO lacks of stability: only small changes in the data bring some variables selected by the LASSO to disappear and some others to appear. Hence, Bach (2009) proposed to combine it with bootstraping to improve its stability: several independent bootstrap samples are generated and the LASSO is performed on each of them. This approach is proved to make the irrelevant variables asymptotically disappear.

A modification is applied to BOLASSO to adapt it to a non-asymptotic framework. An appearance frequency is calculated for each variable X^i by counting the number of times the variable X^i is selected over the bootstrap samples. A high frequency denotes a good predition ability of the variable X^i.

Estimation of the performances

The parameters of each model are estimated first on a part of data set (learning set), thenthe performances are calculated on the other part of the data set (test set). Moreover, in the LASSO method, a single parameter has to be tuned: the regularization parameter, μ. In the BOLASSO method, two parameters have to be tuned: the regularization parameter μ and the appearance frequency threshold. Indeed each variable has a appearance frequency, so a frequency threshold is introduced to select important variables only.

Those parameters have been tuned by cross validation on the learning set. The global procedure (learning with cross validation on the learning set and performances estimation on the test set) was repeated 50 times on several random split on the whole data set. This leads to a collection of performance values that can be displayed through a boxplot in order to evaluate the level of accuracy of each method as well as its variability.

Results and discussion

The chosen phenotype to predict was the “Daily Feed Consumption” and four approaches have been compared to predict this variable: the LASSO and the BOLASSO on the original data (discrete sampling of the spectra) and the LASSO and the BOLASSO on the thresholded wavelet coefficients. The performances were evaluated through the mean squared errors of prediction (MSEP) and the number of selected variables on the 50 test sets (Figure 2). The best prediction is provided by BOLASSO on the wavelet transformed data, using a limited number of predictive variables.

Two conclusions can be driven from this experiment: firstly, BOLASSO improves the accuracy of LASSO and secondly, the wavelet preprocessing leads to better performances than the original data.

This paper focussed on methodological aspects of phenotype prediction based on metabolomic data. A phenotype was chosen for the sake of illustration. Further work is needed to give an

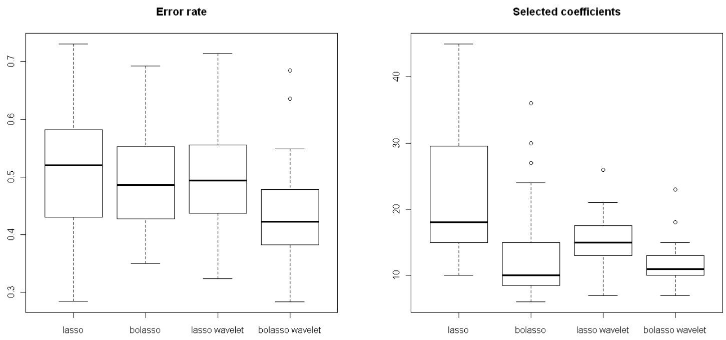

Figure 2: Comparison of LASSO and BOLASSO on original and wavelet transformed data on MSEP (left) and number of selected coefficients (right)

overview on the performance of metabolomic data on the prediction of a large set of quantitative production phenotypes.

Conclusion

Our objective was to predict a phenotype based on metabolomic data. We have shown results about prediction ability of four methods: LASSO or BOLASSO used either on the original data or on thresholded wavelet coefficients. On this data set, the BOLASSO method applied on the wavelet coefficients gave the best results. For prediction purpose in general, a well adapted method of data denoising, coupled with a robust and sparse prediction approach should be recommended.

Acknowledgements

The authors thank the DNA providers (BIOPORC) and French ANR for funding the Délisus project. FR acknowledge financial support from Région Midi-Pyrénées.

References

Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992). *Image coding using wavelet transform*. IEEE Trans. Image Processing., 1, 205-220.

Bach, F. (2009). Model-consistent sparse estimation through the bootstrap. Technical report, hal-00354771, version 1.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, USA.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society, B 58, 267-88.

Villa-Vialaneix, N. and Hernandez-Gonzalez, N. (2009). Using wavelets to explore metabolic data. Technical report, Institut Mathématiques de Toulouse, France.