Classification of signals using wavelets and principal components reduction, with application to auditory brain activity.

Irène Gannaz

CLASSIFICATION OF SIGNALS USING WAVELETS AND PRINCIPAL COMPONENTS REDUCTION, WITH APPLICATION TO AUDITORY BRAIN ACTIVITY.

Irène Gannaz

Université de Lyon
CNRS UMR 5208
INSA de Lyon
Institut Camille Jordan
20, avenue Albert Einstein
69621 Villeurbanne Cedex
France.
E-mail: irene.gannaz@insa-lyon.fr

Abstract

The paper deals with a generalized linear model with functional data using a wavelet representation of the signals. A reduction of dimension is first obtained through a principal component analysis. The discriminative function is then given by a loglikelihood maximization, with a LASSO penalization, in order to ensure the sparsity of the wavelet representation. In order to have a data-driven procedure, we explore different cross-validation schemes. A simulation study is presented, showing our estimator that is competitive with those described in Reiss and Ogden (2013). We apply this model to a classification of functional EEG data, to study the capacity of discrimination of nearby sounds.

Keywords: Generalized functional linear model; Curves classification; Wavelets; Principal component reduction; LASSO penalization; Cross-validation.

1 Introduction

This paper is motivated by a study where the goal is to find out whether the EEG signals measured by external electrodes contain information relative to the cerebral activity in response to auditory stimuli. We are thus interested in discrimination and classification of curve data. More precisely we wish to predict a categorical variable Y with respect to an explanatory functional variable $(X(t))_{t\in[0,1]}$.

We refer to the monograph of Ramsey and Silverman (2005) for an overview of a wide variety of statistical problems dealing with functional data. In this paper we focus on a functional generalized linear regression model, which can be seen as a generalized linear model with functional predictors. Because of our motivation in real data application, we are in particular interested in the functional logistic linear regression, which is a particular case of the generalized framework.

Due partially to the large scope of applications, there is an amount of literature on functional linear regression. Two main approaches were investigated. On the first hand, the functional principal component regression based inference was explored by many authors. Among others, we can cite Cardot et al (1999), Cardot et al (2003a), Cai and Hall (2006), Hall and Horowitz (2007) or Delaigle et al (2009) in a Gaussian model. In a logistic or generalized framework, it was explored for example by Müller and Stadtmüller (2005) Aguilera et al (2006) or Escabias et al (2007). On the second hand, many papers choose to estimate the discriminative function of the model through the projection of the explanatory curves on a functional base, applying a roughness penalty in order to reduce the dimension of the problem. Different bases were used to decompose functions in literature. Among others, we can cite the use of splines by James (2002) with an EM algorithm. Marx and Eilers (1999), Cardot et al (2003b), Marx and Eilers (2005), Cardot and Sarda (2005) consider a splines-based approach with different roughness penalties.

Recently, the use of wavelets-basis has been developed by Brown et al (2001) within a bayesian framework, or by Amato et al (2006). More recently, Zhao et al (2012) explored a wavelets-based procedure with a LASSO penalty in functional linear regression. The authors establish that, provided an appropriate choice of the smoothing parameter associated with the penalty, the estimator is consistent. They also present simulated and real data applications.

A combination of a principal component reduction and a penalized-decomposition based approach was also suggested by Reiss and Ogden (2007) with splines basis and in Reiss et al (2013) with wavelet bases. The estimators defined in this paper are quite similar to those of Reiss et al (2013) and a comparison is yield in the simulation section.

The contribution of this paper is to propose an alternative method of Reiss et al (2013) in a generalized regression framework. The asymptotic behaviour of the estimation procedure is not studied here and may be studied in a further research. This paper focuses on the estimation procedure and the real data application on cerebral responses to auditory stimuli. Due to the high frequency of the signals measured in this application, the wavelet decomposition indeed seems more adapted than others basis like splines. Moreover we are expected that the features allowing the discrimination of the signals are localized in time and in frequency, and thus the time and frequency localization property of the wavelets basis will be useful to capture them. Like in Reiss and Ogden (2007) and Reiss et al (2013) we also choose to combine this procedure with a principal component reduction on the wavelets coefficients of the curve predictors, to ensure an important reduction of the dimension of the problem.

The paper is structured as follows: the first section describes the functional generalized linear regression model, while the second section details the estimation scheme proposed. The last section presents an algorithmic construction of the estimators and their applications on simulated and real data.

2 The functional generalized linear regression

We observe independent data $(Y_i, \{X_i(t),\, t \in [0,1]\})$, $i = 1, \ldots, n$, where the predictor variables $(\{X_i(t),\, t \in [0,1]\})_{i=1,\ldots,n}$ are curves. The functional predictors $X_i(.)$ are supposed to belong to the separable Hilbert space $L^2([0,1])$ whose usual inner product is defined as

$$\forall \alpha, \beta \in \mathbb{L}^2([0,1]), \quad <\alpha, \beta> = \int_{[0,1]} \alpha(t)\beta(t)\, dt.$$

We consider a generalized regression problem, where the response values Y_i are drawn independently from a one-parameter exponential family of distributions, with a probabilistic density of the form:

$$\exp\left(\frac{y\eta(X_i) - b(\eta(X_i))}{\phi} + c(y, \phi)\right). \tag{1}$$

In this expression, $b(\cdot)$ and $c(\cdot)$ are known functions, which determine the specific form of the distribution. The parameter ϕ is a dispersion parameter and is also supposed to be known in what follows. The unknown function $\eta(\cdot)$ is the natural parameter of the exponential family, which carries information from the explanatory variables. It is linked with the quantity $\mathbb{E}(Y_i|X_i)$ through the relation $g(\mathbb{E}(Y_i|X_i)) = \eta(X_i)$. The function $g = \dot{b}^{-1}$ is called link function.

In the functional generalized linear model, $\eta(\cdot)$ is given by a functional linear relation:

$$\eta(X_i) = \int_{[0,1]} X_i(t)\beta(t)\, dt. \tag{2}$$

The function $\beta(\cdot)$ captures the distinctions between the curves $(X_i(\cdot))_{i=1,\ldots,n}$. Intuitively if $|\beta(x_0)|$ is large this means the value taken by the signal at the point x_0 is discriminant. Yet, this interpretation is sometimes not easy in some applications. We refer to McCullagh and Nelder (1989) and Fahrmeir and Tutz (1994) for presentations of the generalized linear models and to the aforementioned papers for generalized regression with functional predictors. Note that we are motivated by curves predictors but an extension to image predictors, as it was done in Reiss and Ogden (2010) and Reiss et al (2013), is possible.

In applications, the curves $X_i(t)$ are discretized at points $t_1, \ldots, t_d$ equidistant on the interval $[0,1]$. This hypothesis excludes applications with random observations of curves but it holds in many applications where t represents time and the signals are observed at equispaced times. In the discretized model, equation (2) can be written on a matricial form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

with the notation $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T = (\eta(X_1), \ldots, \eta(X_n))^T$, $\boldsymbol{\beta} = (\beta(t_1), \ldots, \beta(t_d))^T$ and X the $n \times d$ matrix of general term $(X_i(t_j))_{i=1,\ldots,n,\, j=1,\ldots,d}$. The exponent T denotes the transpose operator.

In the specific logistic framework, which mainly interests us for applications, the response Y_i is drawn from a Binomial distribution, *i.e.* we have $m\, Y_i \sim \mathcal{B}(m, \mu_i)$. The number of classes, $m+1$, is considered known, which corresponds to a supervised classification. The relation $g(\mathbb{E}(Y_i|X_i)) = \eta(X_i)$ is obtained with g the logistic link, $g(u) = \frac{\exp(u)}{1+\exp(u)}$. The data is classified as follows: the label k is given to the ith curve $X_i(\cdot)$ if k is the closest integer of $m\, g(X_i)$. In the case $m = 1$, the label associated with the curve $X_i(\cdot)$ is consequently $\mathbb{1}_{\{g(X_i)>0.5\}}$.

3 Estimation procedure

Our estimation procedure is closed to the one investigated by Zhao et al (2012). We first apply a wavelet transform on the data and apply a LASSO penalization to reduce the dimension of the regression problem. Then, as suggested by Reiss and Ogden (2007), we add a constraint on the wavelets coefficients of the regression function, imposing that they belong to the space generated by the first components of the wavelet representation of the explanatory curves. A cross validation step is necessary to calibrate the parameters of the procedure. A brief presentation of different cross-validation criterions is thus done.

3.1 Projection into the wavelet domain

We consider a wavelet transform of the explanatory curves. Thanks to their time and frequency localization, wavelets allow to capture local singularities and to consider more spatially inhomogeneous functions than kernel or splines. They seem well adapted for dimension reduction purposes and to discriminate signals with localized features. For the sake of concision, we present very briefly the wavelets theory and we refer to Daubechies (1992), Meyer (1992) or Mallat (1999) for more details. For their use in statistics, see Härdle et al (1998), Abramovich et al (2000).

Throughout the paper we assume that we are working within an R-regular ($R \geq 0$) multiresolution analysis of $\left(L^2[0,1], \langle \cdot, \cdot \rangle\right)$, associated with an orthonormal basis generated by dilatations and translations of a compactly supported scaling function, $\varphi(t)$, and a compactly supported mother wavelet, $\psi(t)$. For simplicity reasons, we will consider periodic wavelet bases on $[0,1]$.

For any $j \geq 0$ and $k = 0, 1, \ldots, 2^j - 1$, let us define $\varphi_{j,k}(t) = 2^{j/2}\varphi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$. Then for any given resolution level $j_0 \geq 0$ the family

$$\left\{ \varphi_{j_0,k},\, k = 0, 1, \ldots, 2^{j_0} - 1;\ \psi_{j,k},\, j \geq j_0;\, k = 0, 1, \ldots, 2^j - 1 \right\}$$

is an orthonormal basis of $L^2[0,1]$.

As we consider dicretized versions of curves, we introduce $\mathcal{W}$ the discrete projection operator on this basis. We suppose each curve $X_i(\cdot)$ is observed at equidistant points $t_1, \ldots, t_d$ on the interval [0,1]. We moreover suppose that exists $J \in \mathbb{N}$ such that $d = 2^J$. In real data application it is often sufficient to cut off the end of the signal when it is of no interest.

Every signal $X_i(\cdot)$ is decomposed by $\mathbf{X}_i = \mathcal{W}^T \boldsymbol{\theta}_i$, for $i = 1, \ldots, n$. We introduce $\boldsymbol{\omega}$ the vector of wavelets coefficients of the function $\beta(\cdot)$. In the matricial form, $\boldsymbol{\beta} = \mathcal{W}^T \boldsymbol{\omega}$, and thus $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\Theta}\boldsymbol{\omega}$. The functional regression model is expressed like a regression on the wavelet coefficients. To reduce the dimension of the resulting problem but also to respect the functional nature of the regressor $\beta(\cdot)$, we impose the sparsity of the wavelet coefficients vector $\boldsymbol{\omega}$. This is usually done in literature thanks to a ℓ^1-penalization. It is widely used in wavelets-based estimation and leads to a soft-thresholding of coefficients. We refer for example to Donoho and Johnstone (1994), Donoho and Johnstone (1995) or Donoho and Johnstone (1998) and to Gannaz (2013) in a generalized framework.

This approach has been proposed by Zhao et al (2012) in functional linear models with real responses. The authors establish the asymptotic consistency of the wavelet estimator obtained by the maximisation of a loglikelihood criterion with a ℓ^1-penalty on the wavelet coefficients.

3.2 Principal component reduction

Following Reiss and Ogden (2007), we introduce an additional dimension reduction through a constraint on the wavelet coefficients $\boldsymbol{\omega}$. We impose that they belong to the space generated by the first components of the wavelet coefficients matrix of the curve predictors.

Let $a_1, a_2, \ldots, a_d$ be the eigenvalues of $\boldsymbol{\Theta}$. We suppose the eigenvalues are sorted in the descending order $a_1 \geq a_2 \geq, \cdots \geq a_d$. We introduce the matrix V_q of size $d \times q$ such as the ith column of the matrix V_q is the eigenvector associated with the eigenvalue a_i, for $i = 1, \ldots, q$. We then search $\boldsymbol{\omega}$ such that there exists $\boldsymbol{\gamma} \in \mathbb{R}^q$ verifying $\boldsymbol{\omega} = V_q \boldsymbol{\gamma}$. Note that an extension to Partial Least Squares reduction can easily be done, similarly to Reiss and Ogden (2007).

3.3 Definition of the estimators

Actually, the estimators are defined as follows:

$$\widetilde{\boldsymbol{\omega}}_n(q,\lambda) = \underset{\boldsymbol{\omega}}{argmin} \; -\sum_{i=1}^{n} \mathcal{L}\left(Y_i, \theta_i^T \boldsymbol{\omega}\right) + \lambda \sum_{\ell=2^{j_0}+1}^{d} |\omega_\ell| \tag{3}$$

$$w.r.t. \;\; \exists \boldsymbol{\gamma} \in \mathbb{R}^q, \; \boldsymbol{\omega} = V_q \boldsymbol{\gamma} \tag{4}$$

where $\mathcal{L}$ is the loglikelihood. The estimation of the canonical parameter $\boldsymbol{\eta}$ is given by $\widetilde{\boldsymbol{\eta}}_n(q,\lambda) = \boldsymbol{\Theta}\, \widetilde{\boldsymbol{\omega}}(q,\lambda)$. Note that only the wavelets coefficients are include in the LASSO penalty.

In order to have a data-driven procedure, one needs to choose the smoothing parameter λ and the number of principal components q. We propose different cross-validation procedures. First, a usual minimization of the mean squared error using multiple folds validation can determine λ and q. Parameters also can be chosen by the minimization of a Generalized Cross Validation criterion initially proposed by Craven and Wahba (1979). It has also been applied by O'Sullivan et al (1986) in generalized functional regression models and was used for an automatic choice of the wavelet thresholds *e.g.* in Jansen et al (1997). For a given estimator $\widetilde{\boldsymbol{\omega}}_n(q,\lambda)$, let $\widetilde{\mu}_n(q,\lambda) = g^{-1}(\widetilde{\boldsymbol{\eta}}_n(q,\lambda))$ be the resulting estimation of $(\mathbb{E}[Y|X_i])_{i=1,\ldots,n}$. If $df_n(q,\lambda)$ is the number of degree of freedom of the model, the GCV score is equal to

$$GCV(q,\lambda) = -\frac{\sum_{i=1}^{n} \mathcal{L}(y_i, \widetilde{\mu}_i(q,\lambda))}{df_n(q,\lambda)^2}.$$

If $\widetilde{\mu}_n(q,\lambda) = \mathcal{R}_n(q,\lambda) y$, with $\mathcal{R}_n(q,\lambda)$ operator depending on the number of components q and of the degree of smoothness λ of the procedure, than the degrees of freedom $df_n(q,\lambda)$ in the denominator is $df_n(q,\lambda) = n - trace(\mathcal{R}_n(q,\lambda))$. As it can be seen in the algorithm section 4.1,

we have

$$df_n(q,\lambda) = trace\left(\mathbf{\Theta}\,\mathrm{diag}\left(\mathbb{1}_{\widetilde{\boldsymbol{\omega}}_n(q,\lambda)\neq 0}\right) V_q \left(V_q^T\mathbf{\Theta}^T\widetilde{W}^{-1}\mathbf{\Theta}V_q\right)^{-1} V_q^T\mathbf{\Theta}^T\widetilde{W}^{-1}\right),$$

with $\widetilde{W}$ corresponding to the variance matrix. With a Gaussian distribution, $\widetilde{W}$ is the identity matrix and in the logistic model, $\widetilde{W} = \mathrm{diag}\left(\exp(\widetilde{\eta}_i)(1+\exp(\widetilde{\eta}_i))^{-2}, i = 1,\ldots,n\right)$.

The third method for choosing (q,λ) is an Akaike Information Criterion,

$$AIC(q,\lambda) = -2\sum_{i=1}^{n}\mathcal{L}(y_i,\widetilde{\mu}_i(q,\lambda)) + 2\left(n - df(q,\lambda)\right)/n,$$

or its corrected version

$$AIC_{corr}(q,\lambda) = AIC(q,\lambda) + 2\,\frac{(n-df_n(q,\lambda))(n-df_n(q,\lambda)+1)}{df_n(q,\lambda)-1}.$$

We refer to Burnham and Anderson (2002) for an overview on AIC and its derivatives. According to Wood (2006), the AIC criterion is better justified than the GCV criterion, and Burnham and Anderson (2002) advise the use of the corrected AIC «unless the sample size is large with respect to the number of estimated parameters ». The four cross validation procedures (5-folds cross validation, GCV, AIC and corrected AIC) have been implemented in order to better compare their efficiency.

Let $(\widehat{q},\widehat{\lambda})$ be the argument minimizing one of the criterions described in this section. The estimator of the regression coefficients $\boldsymbol{\omega}$ are then given by $\widehat{\boldsymbol{\omega}}_n = \widetilde{\boldsymbol{\omega}}_n(\widehat{q},\widehat{\lambda})$. And the function $\beta(\cdot)$ is estimated by $\widehat{\boldsymbol{\beta}}_n = \mathcal{W}\widehat{\boldsymbol{\omega}}_n$.

4 Application

This section deals with the implementation of the procedure. We first describe an algorithm to compute the estimator and we then study its behaviour on simulations and on a real data example. All the calculations were carried out in R version 2.14.1 (R Core Team, 2013) on a Unix environment. For the discrete wavelet transform, we used the *wavelets* package 0.2-6 developed by Aldrich (2010).

4.1 Algorithm

The optimisation problem (3) is solved by Iterative Reweighted Least Squares (see page 40 of McCullagh and Nelder (1989)). Each step of the algorithm must take into account the ℓ^1-penalization on the wavelets coefficients. We apply the algorithm presented in Section 2.1. of Friedman et al (2010), more precisely their naïve adaptation of an algorithm proposed by Van der Kooij (2007).

Actually, the kth iteration of the algorithm becomes:

Let $\boldsymbol{\eta}^{(k)} = \boldsymbol{\Theta}\widetilde{\boldsymbol{\omega}}^{(k-1)}$ and $\mu^{(k)} = g(\boldsymbol{\eta}^{(k)})$.

We introduce $\mathbf{Y}^{(k)} = (\mathbf{Y} - \mu^{(k)}) \left.\frac{d\eta}{d\mu}\right|_{\mu=\mu^{(k)}}$ and $W^{(k)} = \text{diag}\left(\left.\frac{d\eta}{d\mu}\right|_{\mu=\mu^{(k)}}\right)$.

The threshold levels are $\lambda^{[k]} = \lambda V_q (V_q^T \boldsymbol{\Theta}^T W^{(k)-1} \boldsymbol{\Theta} V_q)^{-1} V_q^T \mathbb{1}_{d\times 1}$.

For $j = 1, \ldots, d$.

$$\boldsymbol{\omega}^{(k)} = V_q \left(V_q^T \boldsymbol{\Theta}^T W^{(k)-1} \boldsymbol{\Theta} V_q\right)^{-1} V_q^T \boldsymbol{\Theta}^T W^{(k)-1} \left(\boldsymbol{\Theta}_{.j} \gamma_j^{(k)} + \mathbf{Y}^{(k)}\right)$$
$$\widetilde{\omega}_j^{(k)} = sign(\omega_j^{(k)}) \left(|\omega_j^{(k)}| - \lambda_j^{(k)}\right)_+$$

The dependence with the parameters (q, λ) is omitted in the algorithm to simplify the notations.

Zhao et al (2012) suggest to implement the LASSO penalization either by homotopy (Osborne et al, 2000) or by least angle regression (Efron et al, 2004). The algorithm of Friedman et al (2010) seems more adapted in the generalized framework, where an iterative resolution of the maximum likelihood is already needed. Moreover, its computation is quite easy in this framework.

4.2 A simulation study

We compare our procedure with four estimators of the *refund* package (Crainiceanu et al, 2012) corresponding to the estimators studied in Reiss et al (2013). Actually, we will compare the behaviour of seven procedures:

SPCR A splines-based estimation, with a principal component reduction and a penalization, initially described in Reiss and Ogden (2007). Parameters are given by a 5-folds cross validation.

WCR A wavelets-based estimation, with the sparse principal component reduction described in Johnstone and Lu (2009), associated with a 5-folds cross validation.

WNET A wavelets-based estimation with a ℓ^1-penalty. Its asymptotic behaviour and its implementation are studied in Zhao et al (2012). The smoothness parameter is given by a 5-folds cross validation. It is available for elastic-net penalties in the *refund* package but only the ℓ^1-penalty has been considered to have more comparable procedures.

WPCR The wavelets-based estimation described previously in the manuscript, with a principal component reduction and a ℓ^1-penalization. A 5-folds cross validation is used to choose the number of components and the smoothing level.

G-WPCR The same as WPCR but with parameters given by a GCV minimization.

A-WPCR The same as WPCR but with parameters obtained by an AIC procedure.

Ac-WPCR The same as WPCR but with parameters obtained by a corrected AIC procedure.

Reiss et al (2013) already compared the first four estimators cited here but in a quite different context since the regressors were images rather than curves. It seems interesting to include four of them in our study. Note that Reiss et al (2013) also propose estimators based on a Partial Least Squares reduction step instead of a Principal Components reduction, but they are not studied here for the sake of concision.

Note that the conceptions of estimations WCR and WPCR are quite different: the first one is looking for sparsity in the reduction of dimension on the explanatory curves whereas in WPCR we are imposing sparsity to the discriminative function in addition with a reduction of the dimension of the curves predictors. Nevertheless, D'Aspremont et al (2008) argue that there exists connexions between the LASSO variable selection and ℓ^0-penalized extremal eigenvalues decomposition (see references therein) and it would be interesting to investigate whether there is a link between our procedure and a sparse component reduction.

The 5-folds cross validation like the GCV do not need any knowledge of the dispersion parameter ϕ. But it is necessary to apply the AIC and its corrected version. In the Bernoulli model, as we consider supervised classification, $\phi = 1/m$ is given. We will suppose in the Gaussian model that it is known in the simulations. In a real data application one can use the others cross-validation procedures to estimate ϕ and then plug it into the AIC.

We consider the same explanatory curves $X_i(\cdot)$ than Reiss and Ogden (2010), drawn from a Gaussian process on $[0;1]$. More precisely, $X_i(\cdot)$ are independent zero-mean Gaussian processes with $X_i(0) = X_i(1) = 0$ and such that $Cov(X_i(s), X_i(t)) = s(1-t)$ for $s < t$. The target functions β are classical signals introduced by Donoho and Johnstone (1994): *Heavisine*, *Doppler* and *Blocks* and *Bumps*. All of them are given in Figure 1. The signals were discretized in an equidistant grid of 256 points.

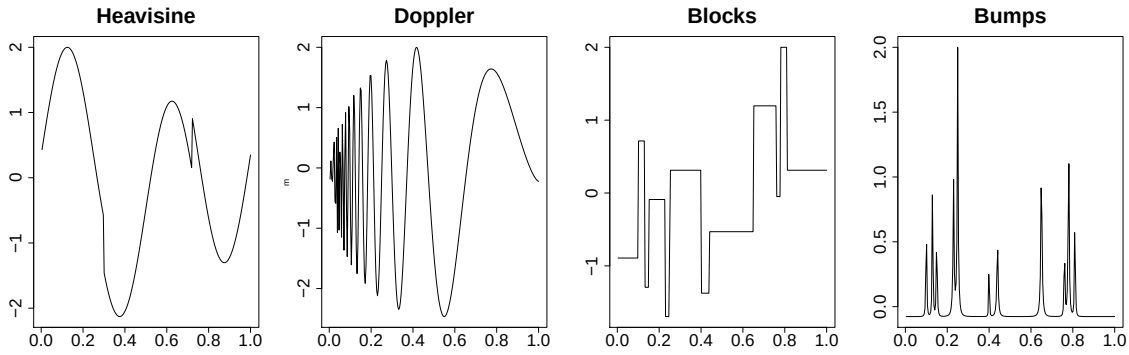

Figure 1: Target functions $\beta(\cdot)$.

In the Gaussian model, following Cardot et al (2003b), we define the signal to noise ratio of the model by the ratio between the standard deviation of $\mathbf{X}_i\boldsymbol{\beta}$ and $\sqrt{\phi}$. We simulate data with a small noise level $\phi = 0.02^2\, Var(\mathbf{X}_i\boldsymbol{\beta})$ and data with a high noise level $\phi = 0.18^2\, Var(\mathbf{X}_i\boldsymbol{\beta})$, taking the values proposed by Cardot et al (1999). This signal-to-noise ratio does not take into account the fact that we do not wish to estimate $\mathbf{X}_i\boldsymbol{\beta}$ but the function $\boldsymbol{\beta}$. Concerning the Binomial model, comparisons were only done on the case of two classes, because estimators of Reiss et al (2013) have not been implemented in the *refund* R-package for more classes. A multiplicative factor was also

applied to curves $X_i(\cdot)$ in order that $\int X_i(t)\beta(t)dt$ belongs to the interval $[-4,4]$ (the multiplicative factor being respectively 0.1, 0.1, 0.17 and 0.33 for *Heavisine*, *Doppler*, *Blocks* and *Bumps*). Indeed, this condition is necessary so that the Binomial model is consistent, see for example on p. 28 of Fahrmeir and Tutz (1994).

First, we do not give here the computation times but our procedure (WPCR, G-WPCR, A-WPCR and Ac-WPCR) is much more longer than the other procedures. The implementation could perhaps be improved because the time efficiency was not aimed in our code, but the recursive algorithm of Friedman et al (2010) chosen is probably more greedy than the Least Angle algorithm.

We carried out the estimation on 100 replications of 100 observations of the model. The parameter q was chosen on a regular grid of step 10 and the parameter λ was seek on by dichotomy. The Least Asymmetric Daubechies wavelet basis with a filter level equal to 16 was used.The quality is given by the Integrated Squared Error: $ISE(\widehat{\beta},\beta) = \frac{1}{d}\sum_{i=1}^{d}\left(\widehat{\beta}(t_i)-\beta(t_i)\right)^2$. Boxplots of the ISE for each simulation frameworks are given respectively in Figure 2 for the Gaussian model with a small noise level, in Figure 4 for the Gaussian model with a small noise level and in Figure 6 for the Bernoulli model.

We also compare the predictive quality of the estimators. We simulate 100 independent replications $(\mathbf{X}_i', y_i')_{\{i=1,\ldots,100\}}$ from the same model and we calculate the prediction obtained by $\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_n$, with $\widehat{\boldsymbol{\beta}}_n$ estimator given on the first 100-simulated observations. The mean prediction error is defined by $\frac{1}{100}\sum_{i=1}^{100}(y_i' - \mathbf{X}_i'\widehat{\boldsymbol{\beta}}_n)^2$ for the Gaussian model and by $\frac{1}{100}\sum_{i=1}^{100}\left|y_i' - \mathbb{1}_{\left\{\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_n>0.5\right\}}\right|$ for the Bernoulli model. Boxplots of the mean prediction error for each simulation frameworks are given respectively in Figure 3 for the Gaussian model with a small noise level, in Figure 5 for the Gaussian model with a small noise level and in Figure 7 for the Bernoulli model.

In the Gaussian framework, Figure 2 shows that, except for the *Heavisine* function, our penalized principal components reduction based estimators performed better than the Splines procedure with a small noise level. In particular, the A-WPCR gives the smallest ISE, with a small dispersion. Note that the insatisfying quality of the WPCR-class estimators for the *Heavisine* function can be improved by imposing the number of components q to be sufficiently small, say under 50 for example. The WNET estimation behaves satisfyingly for the *Heavisine* target but leads to oversmoothed estimators in the others contexts.

The boxplots of the mean prediction error in Figure 3 enhance the bad performance of the WNET procedure. Indeed, even with the *Heavisine* function, it leads to worse prediction quality than the others. Again, the WPCR estimation and its derivatives lead to a good quality of prediction, especially for the AIC choice of the parameters. Splines-based estimation is competitive and gives better predictive values for the *Heavisine* signal, which is coherent with the higher regularity of the target curve.

To better understand the difference between the cross-validation steps, we give in Table 1 the median of the degrees of freedom $df_n(\widehat{q},\widehat{\lambda})$ given by the algorithm in a Gaussian model, with a small noise level. As it can be observed in this Table, the GCV procedure leads to smoother estimators than the others procedures, which explains the better results for the ISE given for

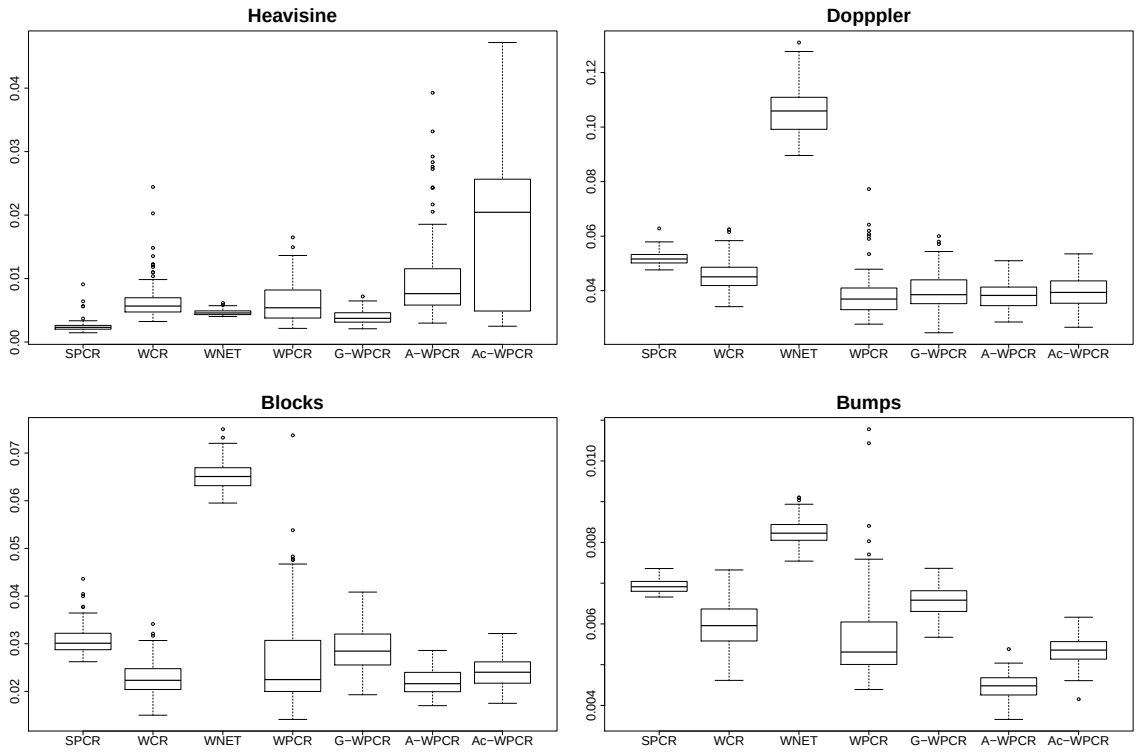

Figure 2: Boxplots of the ISE. Gaussian model, small noise level.

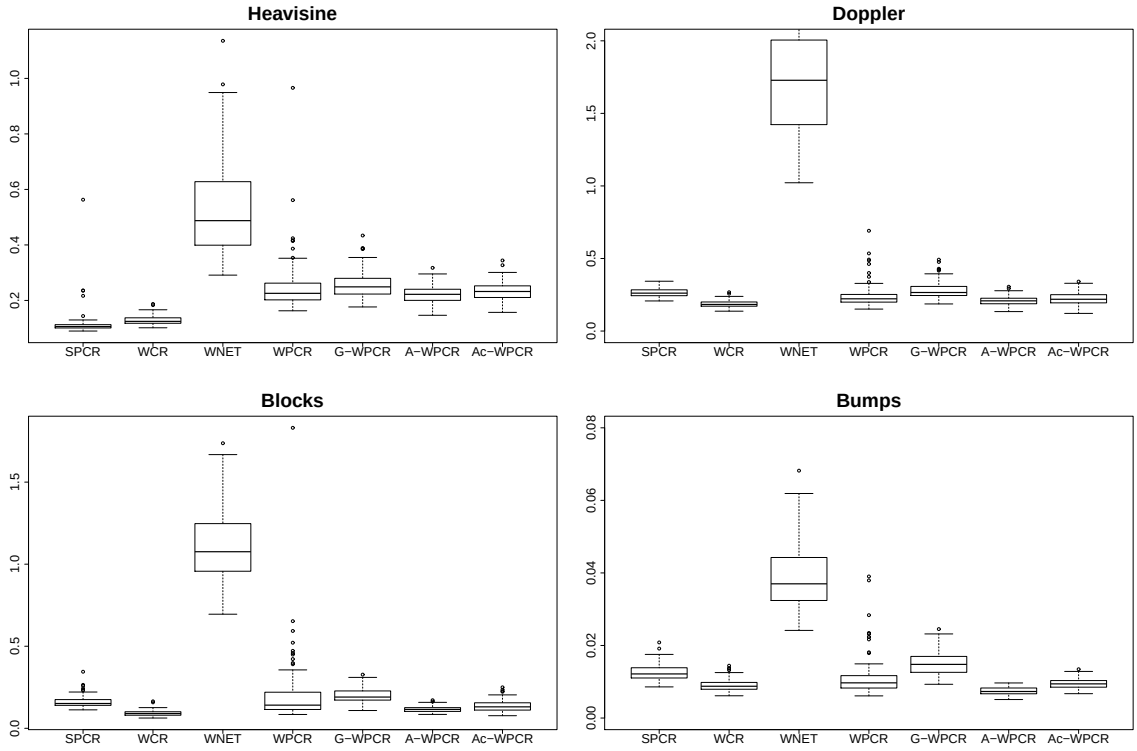

Figure 3: Boxplots of the mean prediction error. Gaussian model, small noise level.

the estimation of the smooth *Heavisine* signal. On the contrary, the AIC and the corrected AIC schemes give undersmoothed estimators and capture well the singularities of the *Blocks* and the *Bumps* functions and the high frequency of the *Doppler* function.

Function	*Heavisine*	*Doppler*	*Blocks*	*Bumps*
WPCR	60.0	70.0	70.0	77.6
G-WPCR	27.4	37.2	40.0	40.0
A-WPCR	69.3	90.0	98.0	97.0
Ac-WPCR	99.0	99.0	99.0	70.0

Table 1: Median of the degree of freedom for the estimators in the Gaussian model, with a small noise level.

Note that we could easily extend the WPCR procedures from soft-thresholding to hard-thresholding (which corresponds to a ℓ^0-penalization on wavelets coefficients, see *e.g.* Antoniadis and Fan (2001)). As it can be seen in Antoniadis et al (2001), this would probably gives better estimation results for *Blocks* and *Doppler* signals but not for the estimation of *Heavisine*.

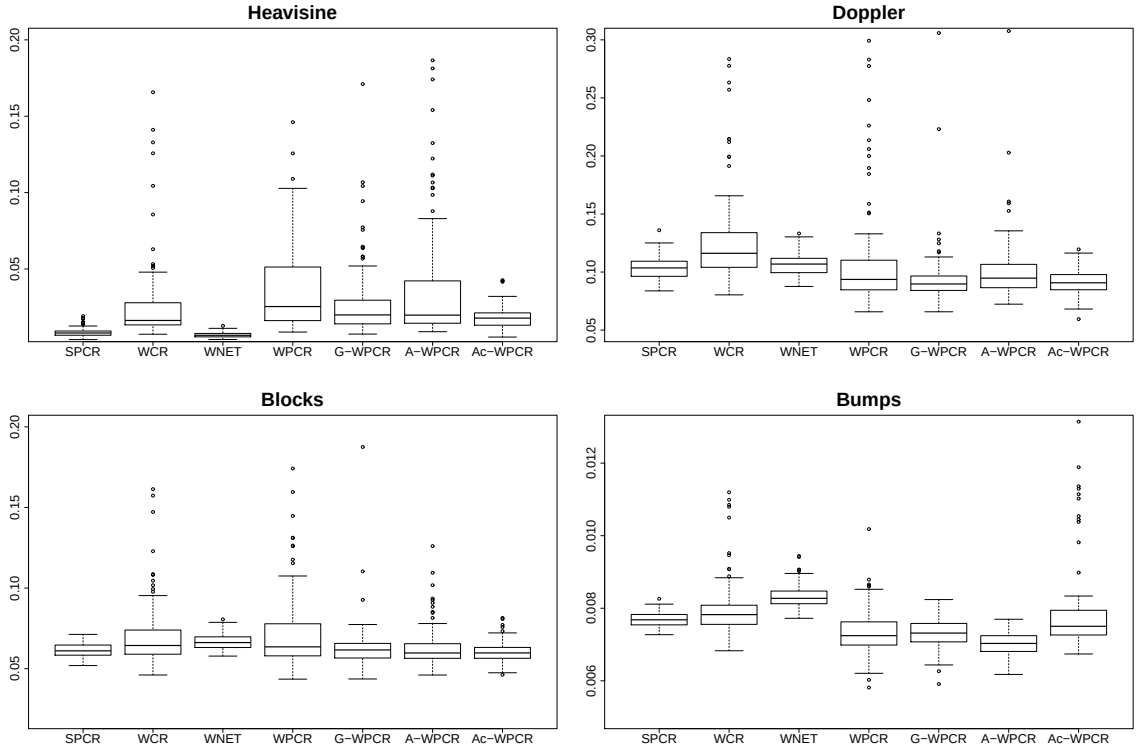

Figure 4: Boxplots of the ISE. Gaussian model, high noise level.

With a higher noise level, the differences of behaviour between the procedures appear less important. As it can be seen in Figure 4, Splines-based and WCR schemes still preform better in the *Heavisine* estimation and WPCR procedures, with all cross-validation steps, are more efficient for the *Bumps* estimation. Yet, for the *Doppler* and the *Blocks* functions, qualities are quite similar. Note that for the *Blocks*model, the structure of the target function is not recognized by any procedures and in this high noise level context procedures give an oversmoothed estimation function.

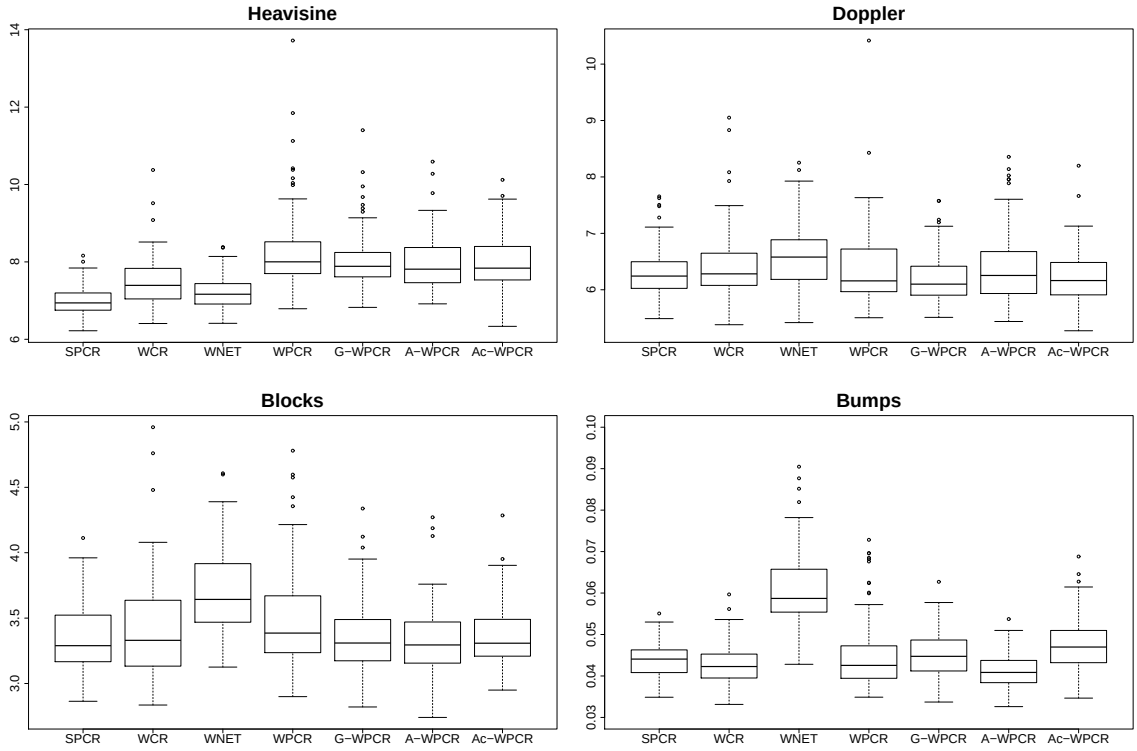

Figure 5: Boxplots of the mean prediction error. Gaussian model, high noise level.

Concerning the prediction quality, Figure 5 confirms the less important difference between the procedures than in a small noise model. WNET gives equivalent or higher prediction errors except for the *Heavisine* function, but all algorithms are competitive.

At least, we consider the Binomial distribution, which corresponds to the classification of the signals $(X_i(\cdot))_{i=1,\ldots,n}$ in two classes. The first remark that can be done in view of Figure 6 is that, on the contrary of the conclusions in the Gaussian framework, the splines-based procedure obtains better quality estimators than the wavelets-based procedures. This shows *a priori* that the validation step used to choose the parameters in the wavelets schemes introduces a bias, more important than in the splines-based estimation. The second remark is the bad performance, according to Figure 6, of the WCR estimator. Nevertheless, Figure 7 shows that its quality of prediction is satisfying.

Concerning the penalized wavelets components reduction estimation developed here, one can see in Figure 6 that if AIC was the best-performing validation step with the Gaussian distribution, its corrected version is preferable with a Binomial distribution. Comparing the quality of prediction, one can see that GCV, AIC and corrected AIC give equivalent results, which are globally the same quality as splines-based estimators. Finally, Figure 7 puts in evidence that the 5-folds cross-validation step leads to a bad quality of prediction. Actually, this is due to the fact that the algorithm WPCR often leads to a zero function.

In conclusion, our procedure is competitive with those described in Reiss et al (2013). The main drawback of our estimation scheme remains its computational time since it is much longer than procedures of the *refund* package (Crainiceanu et al, 2012). The combination of a penalization and

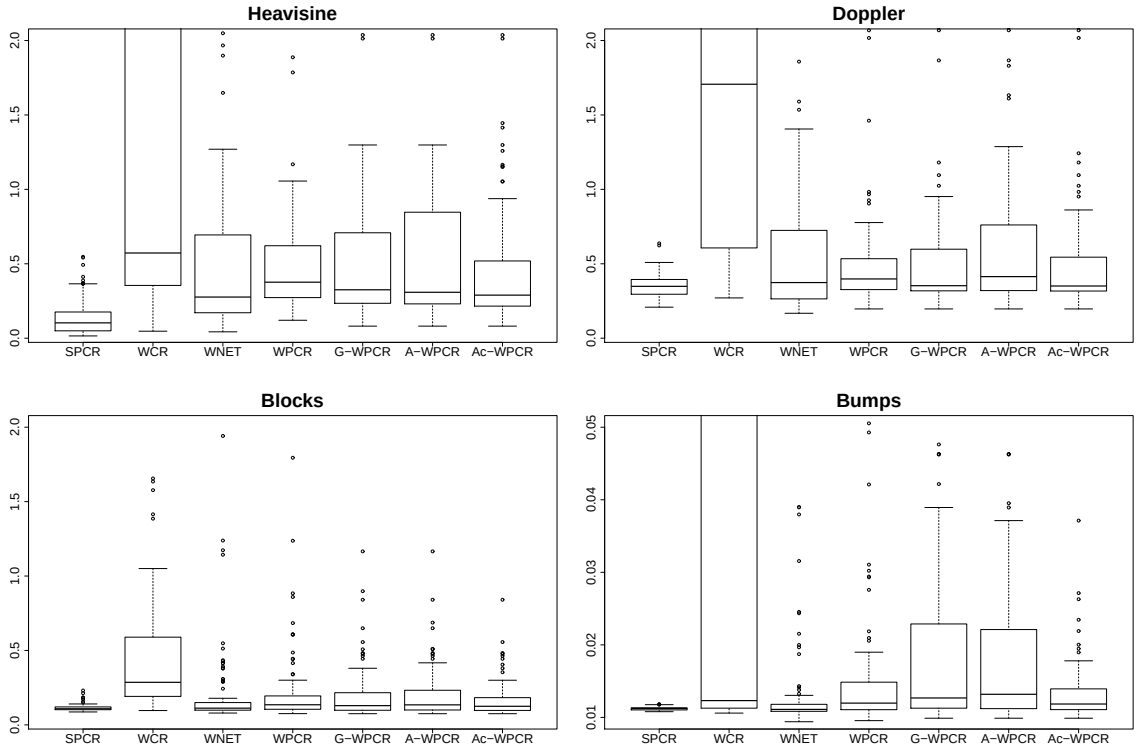

Figure 6: Boxplots of the ISE. Bernoulli model.

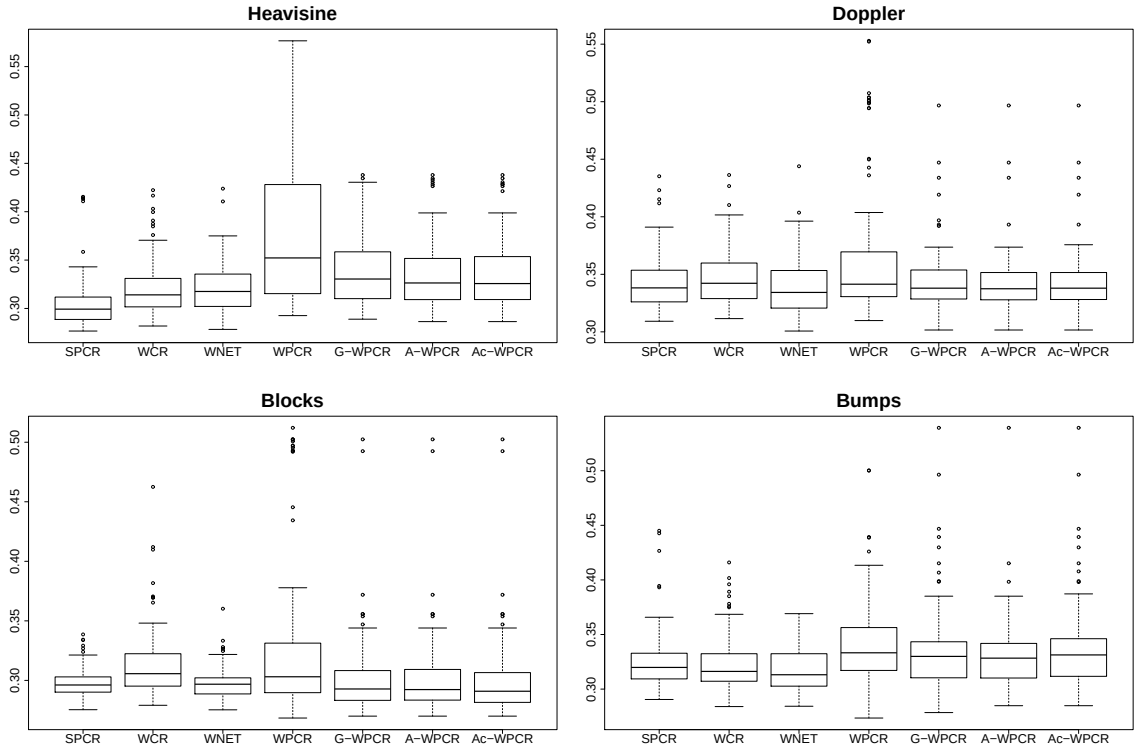

Figure 7: Boxplots of the mean prediction error. Bernoulli model.

of a principal components reduction in a wavelets-based procedure leads to estimators of good quality, among others when comparing to the WNET and the WCR procedures of Reiss et al (2013). We have also established that the quality of the estimation and of the prediction are comparable with the splines-based method developed in Reiss and Ogden (2007) and performs better in a small-noise level Gaussian framework.

4.3 A real data application

This application has been done in a collaboration with Ludovic Bellier, Rafael Laboissière and Fabien Millioz, of the DyCog Team, Lyon Neuroscience Research Center, France. We are interested in the human perception of speech, more precisely in the capacity to discriminate bilabial plosives, here */b/* versus */p/*.

Data was obtained by an electroencephalography (EEG) which recorded the auditory evoked potentials from the scalp in a standard 32-electrodes mounting cap, on a single subject. The subject was placed in a soundproof booth and watched a mute movie while hearing the stimuli, instructed to not pay attention to the heard sounds. Four stimuli were emitted in a random order. Stimuli are presented in Figure 8. They correspond to sounds */ba/*, */pa/* and two intermediates obtained by modifying the offset of the plosives */b/* and */p/*, taking intermediate values. The signals were sampled at 5 kHz and the interstimulus interval was 410 ms. For each of the four stimuli, we dispose of 1000 records, which we will study on a 2^{10} equally spaced grid. For a more detailed description of the experiment, we refer to Bellier et al (2013).

A high-pass filter was applied to the EEG signals to keep only the frequencies corresponding to auditory activity and not to muscular activity. In addition, we consider averages of ten signals in order to get rid of a possible random effect. We then dispose of 100 curves for each stimuli.

We first apply the generalized functional linear model to discriminate the responses to the two pure stimuli */ba/* and */pa/*. For $i = 1, \ldots, 100$, the curve $X_i(\cdot)$ is an average auditory evoked potential measured from the EEG for the ith stimulus heard. The binary label Y_i corresponds to the stimulus, taking the value 0 if the stimulus sound is */ba/* and the value 1 if it is the heard sound */pa/*. We are looking for the estimation of the function $\beta(\cdot)$ in the model (1,2) with a Binomial distribution.

The estimators $\widehat{\beta}_n(\cdot)$ obtained by the means of the seven procedures implemented are given in Figure 9. We first notice that the splines-based estimators is far different from the others. Simulation studies have also put in evidence that the WNET procedure leads to more sparsity than schemes with components reduction, which is still observed in Figure 9.

To better understand which part of the curves are discriminative, we give a representation of the discrete wavelet transform of the estimated function $\widehat{\beta}_n(\cdot)$ obtained jointly by G-WFPCR and A-WFPCR in Figure 10. We can see that the most significant coefficients are located at frequencies that effectively contains the difference between the sounds */ba/* and */pa/*. Moreover, the shift between the beginning of the register and the higher coefficients is not surprising as it is well-known that some time is necessary for the well-known by medecine as the time for the sound reaching the auditory cortex.

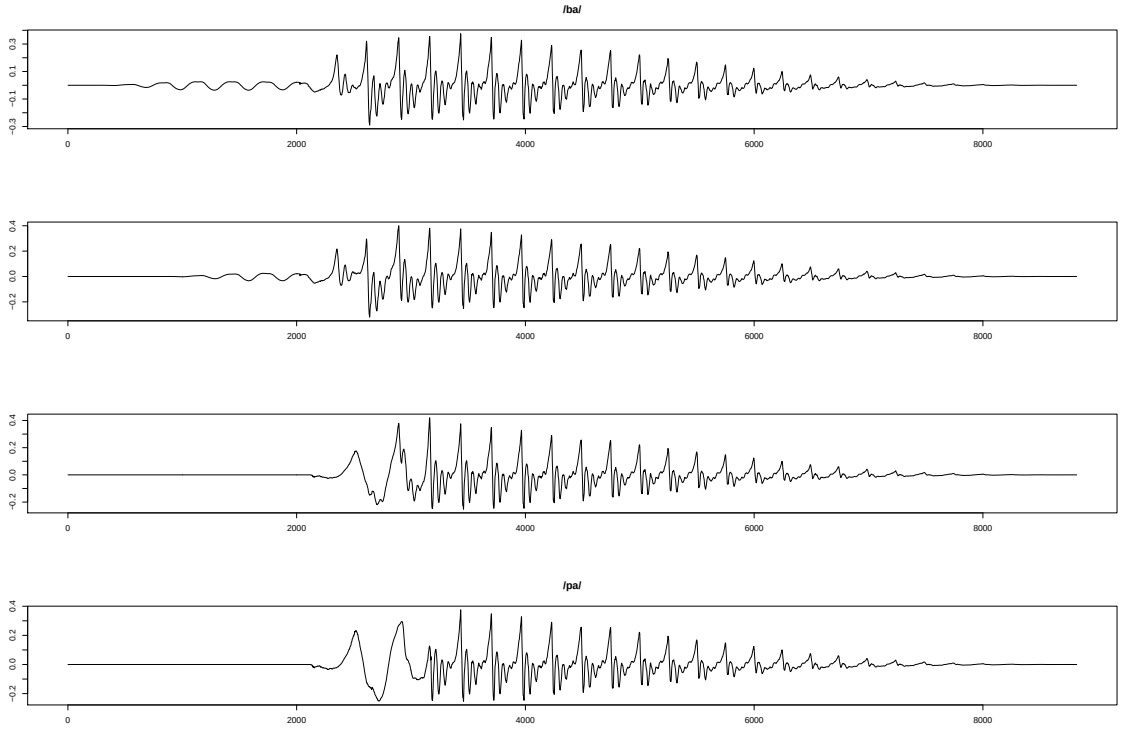

Figure 8: The four stimuli, */ba/*, */pa/*, and two intermediates.

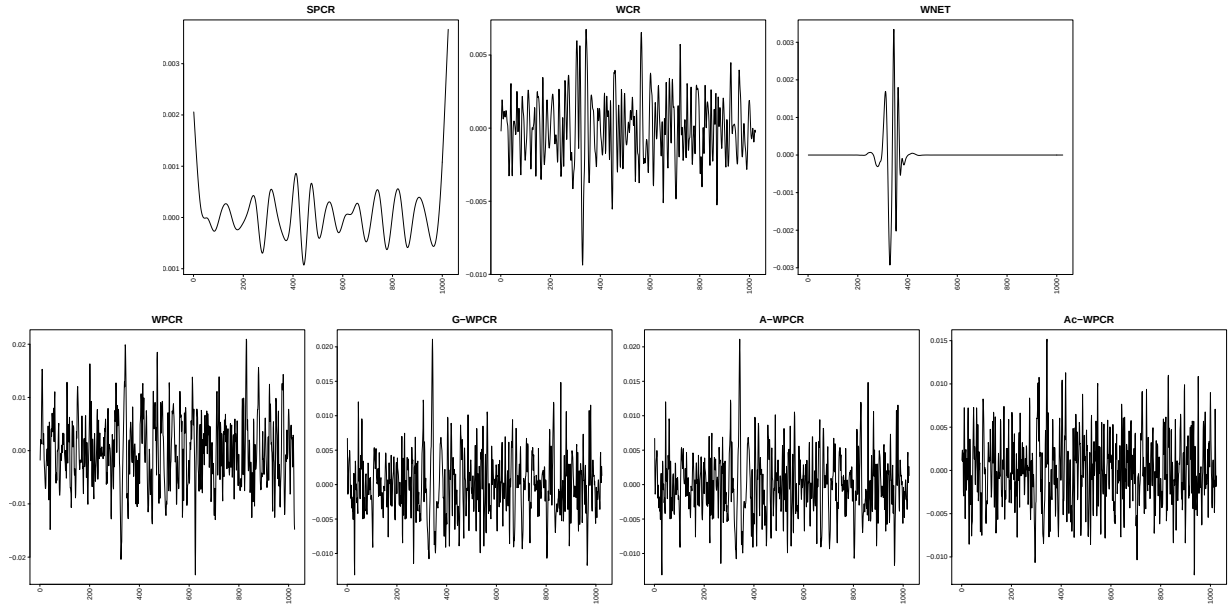

Figure 9: Estimation $\widehat{\beta}_n(\cdot)$ in the discrimination of */ba/* and */pa/* average evoked potentials.

Table 2 gives the predictive label associated with the sound */pa/*. Concerning the quality of classification of the pure sounds */ba/* and */pa/*, Ac-WPCR allows an quasi-exact discrimination, with only one false estimated label. WPCR, G-WPCR and A-WPCR also gives satisfactory results, with

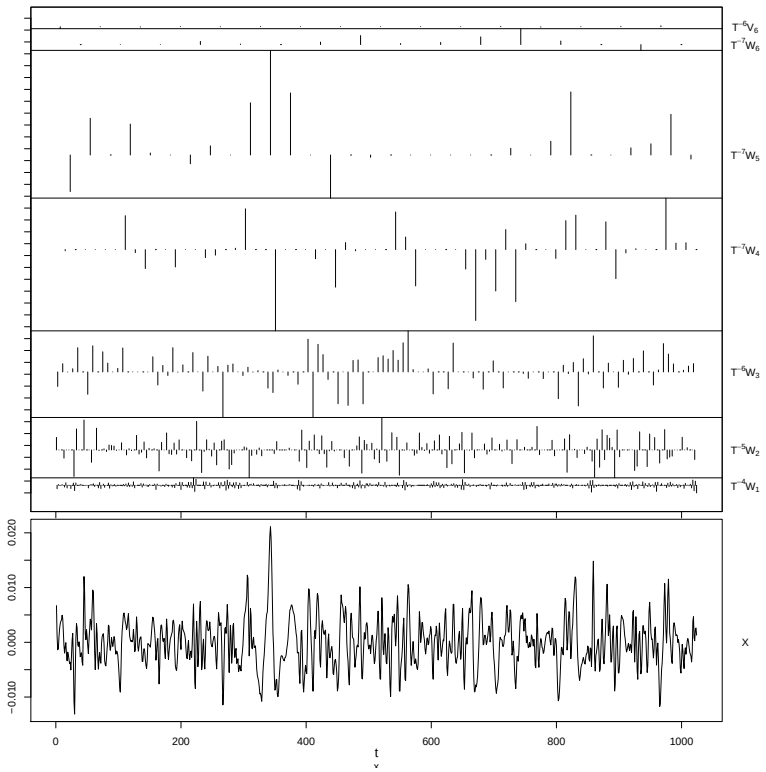

Figure 10: Discrete wavelet transform of $\widehat{\beta}_n(\cdot)$ obtained by G-WFPCR.

	/ba/	intermediate stimulus 1	intermediate stimulus 2	*/pa/*
SPCR	44%	50%	52%	58%
WCR	20%	23%	54%	77%
WNET	24%	35%	59%	76%
WPCR	3%	44%	52%	97%
G-WPCR	1%	32%	54%	93%
A-WPCR	1%	32%	54%	93%
Ac-WPCR	1%	36%	56%	100%

Table 2: Classification rate in the class corresponding to */pa/* for each of the four stimuli. Each curve is an average of ten EEG signals.

a reasonable misclassification rate. But SPCR, WCR and WNET failed in discriminating the signals.

We then focus on the evoked potentials of the intermediate stimuli. Ideally, we would have liked the first intermediate stimulus to be recognized as a */ba/*, and the second one as a */pa/*. Figure **??** illustrates that we cannot validate this hypothesis. Indeed, if we put label 0 on the first average curves of evoked potential for the first intermediate stimulus and label 1 on the curves for the second intermediate stimulus, the minimum prediction error rate is 39%, reached with G-WPCR and A-WPCR. Yet we do not reject either the hypothesis, since we have established in simulation studies that such rates are effectively observed, even with correctly labelled curves.

Finally, we tried to see if it was possible to discriminate responses to the four stimuli in four classes. It appears that no algorithm gives satisfactory results, all of them leading finally to a two-classes distinction. Consequently, we could say that the fact the two intermediate stimuli are identified by the patient respectively as */ba/* and */pa/* is a plausible hypothesis.

To conclude, the WPCR class estimators are more efficient than alternative methods to discriminate these high frequency signals and thus seem preferable when studying such curves. Simulations also show the good behaviour of the WPCR class estimators within a wide variety of frameworks. Note only that a preference should be given to splines-based procedures when expecting a smooth target function, like *Heavisine*, or else the number of components retained must be forced not to be high.

Conclusion

This paper presents a estimation procedure in generalized functional linear models, based on a wavelet decomposition of the explanatory curves. A principal component reduction is used to ensure a reduction of the dimension of the model, combined with a penalization of the log-likelihood criterion to ensure the sparsity of the estimated discriminative function. The estimators are compared with the four similar schemes presented in Reiss et al (2013), and different cross-validation steps are proposed. The simulation study enhances that estimators are competitive. Yet in the real data application on EEG signals, the estimators developed in this paper seem more efficient. The application consists in discriminating the cerebral response to different sounds, in particular */ba/* and */pa/*. The component-reduction and penalized approach proposed here lead to a good discrimination of the pure sounds, when others methods failed. Nevertheless, the study does not allow to determine precisely whether non-pure sounds give cerebral response comparable to pure ones.

Acknowledgments : *This work was motivated by a collaboration with Ludovic Bellier, Rafael Laboissière and Fabien Millioz, of the DyCog Team, in Lyon Neuroscience Research Center, France.*

References

Abramovich F, Bailey T, Sapatinas T (2000) Wavelet analysis and its statistical applications. The Statistician 49(1):1–29

Aguilera A, Escabias M, Valdemarra M (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics & Data Analysis 50:1905–1924

Aldrich E (2010) Wavelets: A package of funtions for computing wavelet filters, wavelet transforms and multiresolution analyses. R package version 0.2-6

Amato U, Antoniadis A, De Feis I (2006) Dimension reduction in functional regression with applications. Computational Statistics & Data Analysis 50(9):2422–2446

Antoniadis A, Fan J (2001) Regularization of wavelet approximations. Journal of the American Statistical Association 96(455):939–967

Antoniadis A, Bigot J, Sapatinas T (2001) Wavelet estimators in nonparametric regression: a comparative simulation study. Journal of Statistical Software 6(6)

Bellier L, Mazzuca M, Thai-Van H, Caclin A, Laboissière R (2013) Categorization of speech in early auditory evoked responses

Brown P, Fearn T, Vannucci M (2001) Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. Journal of the American Statistical Association 96:398–408

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer-Verlag

Cai T, Hall P (2006) Prediction in functional linear regression. The Annals of Statistics 34(5):2159–2179

Cardot H, Sarda P (2005) Estimation in generalized linear models for functional data via penalized likelihood. Journal of Multivariate Analysis 92:24–41

Cardot H, Ferraty F, Sarda P (1999) Functional linear model. Statistics & Probability Letters 45:11–22

Cardot H, Ferraty F, Mas S, Sarda P (2003a) Testing hypothesis in the functional linear model. Scandinavian Journal of Statistics 30(1):241–255

Cardot H, Ferraty F, Sarda P (2003b) Spline estimators for the functional linear model. Statistica Sinica 13:571–591

Crainiceanu C, Reiss P, Goldsmith J, Huang L, Huo L, Scheipl L (2012) refund: regression with functional data. R package version 0.1-6

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31:377–403

D'Aspremont A, Bach FR, El Ghaoui L (2008) Optimal solutions for sparse principal component analysis. The Journal of Machine Learning Research 9:1269–1294

Daubechies I (1992) Ten lectures on wavelets, vol 61. SIAM press

Delaigle A, Hall P, Apanasovich T (2009) Weigthed least squares methods for prediction in the functional linear model. Electronic Journal of Statistics 3:865–885

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

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

Donoho D, Johnstone I (1998) Minimax estimation via wavelet shrinkage. The Annals of Statistics 26(3):879–921

Efron B, Hastie T, Tibshirani R (2004) Least angle regression. The Annals of Statistics 32:407–499

Escabias M, Aguilera A, Valdemarra M (2007) Functional PLS logit regression model. Computational Statistics & Data Analysis 51:4891–4902

Fahrmeir L, Tutz G (1994) Multivariate statistical modelling based on generalized linear models. Springer

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22

Gannaz I (2013) Wavelet penalized likelihood estimation in generalized functional models. TEST (22):122–158

Hall P, Horowitz J (2007) Methodology and convergence rates for functional linear regression. The Annals of Statistics 37:70–91

Härdle W, Kerkyacharian G, Picard D, Tsybakov A (1998) Wavelets approximation and statistical applications, vol 129. Springer Verlag, Lecture Notes in Statistics

James J (2002) Generalized linear models with functional predictors. Journal of the Royal Statistical Society 64(3):411–432

Jansen M, Malfait M, Bultheel A (1997) Generalized cross validation for wavelet thresholding. Signal Processing 56(1):33–44

Johnstone IM, Lu AY (2009) On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association 104:682–693

Van der Kooij A (2007) Prediction accuracy and stability of regression with optimal sclaing transformations. Tech. rep., Leiden University: Dept. of Data Theory

Mallat S (1999) A wavelet tour on signal processing, 2nd edn. Academic Press

Marx B, Eilers P (1999) Generalized linear regression on sampled signals and curves: a P-spline approach. Technometrics 41:1–13

Marx B, Eilers P (2005) Multidimensional penalized signal regression. Technometrics 47:13–21

McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman and Hall

Meyer Y (1992) Wavelets and operators. Cambridge University Press

Müller H, Stadtmüller U (2005) Generalized functional linear models. The Annals of Statistics 33:774–805

Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. IMA Journal of Numerical Analysis 20:389–404

O'Sullivan F, Yandell B, Raynor W (1986) Automatic smoothing of regression functions in generalized linear models. Journal of the American Statistical Association 81(393):96–103

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, ISBN 3-900051-07-0

Ramsey J, Silverman B (2005) Functional data analysis. Springer, New York

Reiss P, Ogden R (2007) Functional principal component regression and functional partial least squares. Journal of the American Statistical Association 102:984–996

Reiss P, Ogden R (2010) Functional generalized linear models with images as predictors. Biometrics 66(1):61–69

Reiss P, Huo L, Ogden R, Zhao Y, Kelly C (2013) Wavelet-domain methods for scalar-on-image regression, URL `http://works.bepress.com/phil_reiss/29`

Wood SN (2006) Generalized additive models: an introduction with R. Boca Raton, FL:Chapman and Hall/CRC

Zhao Y, Ogden R, Reiss P (2012) Wavelet-based LASSO in functional linear regression. Journal of Computational and Graphical Statistics 21(3):600–617