Noname manuscript No.
(will be inserted by the editor)

Penalized estimation in additive varying coefficient models using grouped regularization

A. Antoniadis · I. Gijbels · S. Lambert-Lacroix

Received: date / Accepted: date

Abstract Additive varying coefficient models are a natural extension of multiple linear regression models, allowing the regression coefficients to be functions of other variables. Therefore these models are more flexible to model more complex dependencies in data structures. In this paper we consider the problem of selecting in an automatic way the significant variables among a large set of variables, when the interest is on a given response variable. In recent years several grouped regularization methods have been proposed and in this paper we present these under one unified framework in this varying coefficient model context. For each of the discussed grouped regularization methods we investigate the optimization problem to be solved, possible algorithms for doing so, and the variable and estimation consistency of the methods. We investigate the finite-sample performance of these methods, in a comparative study, and illustrate them on real data examples.

Keywords Grouped lasso regularization · Multiple linear regression models · Variables selection · Varying coefficient models

Laboratoire Jean Kuntzmann, Department de Statistique, Université Joseph Fourier, Tour IRMA, B.P.53, 38041 Grenoble CEDEX 9, France.

Department of Mathematics and Leuven Statistics Research Centre (LStat) Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), Belgium

UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG
UMR 5525, Grenoble, F-38041, France,
E-mail: Sophie.Lambert@imag.fr

1 Introduction

In a classical linear regression model the influence of covariates $X^{(1)}, \cdots, X^{(p)}$ on a response variable Y is modelled via

$$Y = \beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} + \varepsilon,$$

where ε denotes the error term in the regression model. A useful extension of this classical linear regression model is functional (varying) coefficient models, where model parameters (such as β_j, $j = 0, \cdots, p$) may change with the value of other variables (factors). To formalize the functional coefficient, parametric representations such as finite order polynomials or Fourier expansions, or otherwise nonparametric approaches can be employed.

In the varying coefficient model of [21], the regression function depends linearly on some regressors, with coefficients considered as smooth functions of other predictor variables, called tuning variables. A special type of varying coefficient model is called the functional coefficient model by [11] (see also [18]). There, all tuning variables are the same and univariate. Such models have been used for longitudinal data where subjects are often measured repeatedly over a given period of time, so that the measurements within each subject are possibly correlated with each other (see [39,37]).

While many procedures have been developed in the literature for estimating the varying coefficients, the problem of variable selection for such models has rarely been addressed. Recently, [31] have studied the problem of variable selection for partial linear varying coefficient models, where the parametric components are identified via the Smoothed Clipped Absolute Deviation (SCAD) procedure of [15] but the varying coefficients are selected via the generalized likelihood ratio test of [17]. Their approach can be viewed as a combination of shrinkage and hypotheses testing methods. In [2] the authors use an extension of the nonnegative garrote selection method to select variables in a varying coefficient model. That paper also discusses a selection method that is equivalent to a grouped LASSO regularization method, discussed in our review of methods.

In this paper we present in a unifying framework several regularized estimation procedures for variable selection in nonparametric varying coefficient models using basis function approximations and grouped type of penalties. We focus on a varying coefficient model used in the context of longitudinal data. Such data arise in many scientific studies, where measurements possibly change over time t, leading to a response variable $Y(t)$ and covariates $X^{(1)}(t), \cdots, X^{(p)}(t)$. It is then of interest to study the association between the covariates and the responses and to examine how the association varies with time. A simple and useful model for studying the association between $Y(t)$ and the covariates $\left(X^{(1)}(t), \cdots, X^{(p)}(t)\right)$ is then the linear model

$$Y(t) = \beta_0(t) + \beta_1(t) X^{(1)}(t) + \cdots + \beta_p(t) X^{(p)}(t) + \varepsilon(t), \tag{1.1}$$

where $\varepsilon(t)$ is a zero-mean correlated stochastic process that cannot be explained by the covariates. Such a model has been considered in [18] as a functional linear model for longitudinal data. Model (1.1) is also a specific model within a class of functional linear models introduced by [38] in a somewhat different context. For the varying coefficient models, smoothing spline and kernel methods are proposed in [21]. In [22], smoothing spline and kernel methods were studied whereas in [10] the smoothing spline method was considered for functional analysis-of-variance (ANOVA) models which are special cases of functional linear models. Although the spline method has better performance than the kernel method due to its introduction of multiple smoothing parameters [22], its computation is very intensive even for a longitudinal data set of moderate size, not to mention the difficulty of selecting the multiple smoothing parameters which involves high dimensional optimization problems. For some longitudinal data sets with special structure [18] proposed two-step procedures that overcome the inflexibility of traditional spline and kernel methods.

Model (1.1) is also the same as the one used by [24] but where a global smoothing procedure is developed for estimating the parameters using a basis function approximation for the varying coefficients functions in a repeated measurements longitudinal data model. It is also the model studied by [47] where the varying coefficients functions are estimated by some locally kernel weighted least squares procedures. Model (1.1) further includes many other useful models proposed in the literature, as will be discussed in the next section.

In this paper we study the variable selection problem in the context of model (1.1). We use the method of basis expansion to estimate the smooth functions $\beta_j(\cdot)$ and discuss various grouped regularization methods for variable selection, including grouped LASSO regularization, grouped SCAD regularization, grouped Bridge regularization and grouped COSSO regularization. Using results of [33], we show that the grouped LASSO regularization method is variable selection consistent and also estimation consistent, but not simultaneously, in asymptotic sense, even when the dimensionality p increases much faster than the sample size. For the other methods we also briefly discuss available asymptotic results on variable selection and estimation consistency available in the literature. For each grouped regularization selection method we comment on the available algorithms for solving the specific optimization problem.

The paper is organized as follows. In Section 2 we introduce the modeling framework with the necessary notations. In Sections 3–6 we discuss four grouped regularization methods, of which the finite-sample performances are investigated via a simulation study in Section 7. In the same section the use of the grouped regularization techniques on some real data is illustrated.

2 Model formulation and set up

We consider a varying coefficient model

$$Y(t) = \mathbf{X}(t)\boldsymbol{\beta}(t) + \varepsilon(t), \tag{2.1}$$

where $\mathbf{X}(t) = (1, X^{(1)}(t), \cdots, X^{(p)}(t))$, of dimension $1 \times (p+1)$, is the vector of time-dependent covariates, and $\boldsymbol{\beta}(t) = (\beta_0(t), \beta_1(t), \beta_2(t), \cdots, \beta_p(t))^T$ is a vector of time-varying coefficients, with $\mathbf{A}^T$ denoting the transposed of a vector or matrix $\mathbf{A}$. The first elements in these vectors ensure the inclusion of an intercept parameter function in the model.

For the error term we assume that for all t and s,

$$E\left(\varepsilon(t)\right) = 0, \qquad \text{and} \qquad \text{Cov}\left(\varepsilon(t), \varepsilon(s)\right) = \sigma^2 \delta_{st},$$

with δ_{st} the Kronecker delta, defined as the function of (s,t) that is 1 if $s = t$ and 0 otherwise.

In the context of longitudinal data, we have for each individual/subject under study (for $i = 1, \cdots, n$), observations at discrete time point $t_{i1}, \cdots, t_{iN_i}$, denoted by

$$\left((\mathbf{X}_i(t_{i1}), Y_i(t_{i1})), \cdots, (\mathbf{X}_i(t_{iN_i}), Y_i(t_{iN_i}))\right),$$

with $\mathbf{X}_i(t_{ij}) = \left(1, X_i^{(1)}(t_{ij}), \cdots, X_i^{(p)}(t_{ij})\right)$ the observed covariate values for individual i at time point t_{ij}. So, n denotes the number of subjects/individuals, N_i is the number of observations at discrete time points for individual/subject i, and p is the number of covariates.

Observations are from the model (2.1) and hence satisfy

$$Y_i(t_{ij}) = \mathbf{X}_i(t_{ij})\boldsymbol{\beta}(t_{ij}) + \varepsilon_i(t_{ij}), \quad i = 1, \cdots, n,\ j = 1, \cdots, N_i, \tag{2.2}$$

with, for all i, j and all s, t,

$$E\left(\varepsilon_i(t_{ij})\right) = 0 \qquad \text{and} \qquad \text{Cov}\left(\varepsilon_i(t), \varepsilon_j(s)\right) = \sigma^2 \delta_{st} \delta_{ij}.$$

As noted by a referee, such an assumption on the covariance structure of the stochastic process modeling of the longitudinal effects is somewhat restrictive and excludes interesting functional varying coefficient models. However, not only it simplifies our analysis, but note also that such a specification seems unavoidable since the large dimensions of the covariance matrices involved make it infeasible to estimate them in a completely unstructured fashion. We believe that our results can be extended to dependent observations with particular covariance structure and we hope to address this issue in the future.

The main interest in the paper is in the variable selection problem, in particular when $p > n$. For tackling this problem we study procedures of grouped LASSO, grouped SCAD, grouped Bridge and grouped COSSO regularization. The focus is of course on the p univariate functions $\beta_k(\cdot)$, $k = 1, \cdots, p$, since they describe the influence of the covariates $X^{(k)}(\cdot)$ on the response variable $Y(\cdot)$.

In practice, it is more useful to express the model in terms of basis functions. Assume that the functions $\beta_k(\cdot)$, $k = 0, 1, \cdots, p$, belong to a certain space of smooth functions, say that we can write

$$\beta_k(t) = \sum_{\ell=1}^{\infty} \gamma^*_{k\ell} B_\ell(t),$$

where γ^* denotes the true parameter and we further approximate this by

$$\beta_k(t) \approx \sum_{\ell=1}^{L_k} \gamma^*_{k\ell} B^{(k)}_\ell(t), \tag{2.3}$$

where the superscript (k) indicates that the set of approximating basis functions can be different for each univariate function, and where L_k is an integer-valued truncation parameter, possible different for each k. For example, in an approximation with B-splines one could use B-splines of a different degree and/or a different number of knot points for each of the univariate functions. Note that the approximation in (2.3) means that one already has dealt with a (modeling) bias issue. In a modeling setting, we will accept this approximation. Note, however, that when focusing on the asymptotic analysis (when L_k goes to infinity), the rate of convergence obtained for each variable coefficient is the optimal rate for nonparametric regression. Therefore the incurred loss due to this approximation is not important asymptotically.

Hereafter we restrict ourselves to the finite dimensional space of cubic B-splines. For each function β_k we use a cubic B-spline parameterization with a reasonable amount of knots or basis functions. A typical choice would be to use $(L_k - 2) \asymp \min_{i=1,\ldots,n} {N_i}^{1/5}$ interior knots that are placed at the empirical quantiles of $X^{(k)}(\cdot)$, completed by two extra knots placed at the boundaries of the domain of definition of β_k. A truncation parameter $L_k = L_{k_n}$ for each component of the order $\min_{i=1,\ldots,n} {N_i}^{1/5}$ yields a truncation bias that is negligible for twice differentiable functions, i.e. $\left\| \beta_k - \sum_{\ell=1}^{L_k} \gamma^*_{k\ell} B^{(k)}_\ell \right\|^2_{L_2} = O(L_k^{-4})$, see for instance [36].

We now rewrite the (approximate) model in matrix notation. Substituting the approximation (2.3) into the model (1.1) we can write for $i = 1, \ldots, n$, $j = 1, \ldots, N_i$,

$$Y_i(t_{ij}) = \sum_{k=0}^{p} X^{(k)}_i(t_{ij}) (\boldsymbol{B}^{(k)}(t_{ij}))^T \boldsymbol{\gamma}^*_k + \varepsilon_i(t_{ij}), \tag{2.4}$$

where we introduced the notation

$$\boldsymbol{\gamma}^*_k = \left(\gamma^*_{k,1}, \cdots, \gamma^*_{k,L_k}\right)^T \qquad \text{and} \qquad \boldsymbol{B}^{(k)}(t) = \left(B^{(k)}_1(t), \cdots, B^{(k)}_{L_k}(t)\right)^T$$

for these vectors of dimension $L_k \times 1$ and

$$\mathbf{Y} = (Y_1(t_{11}), \cdots, Y_1(t_{1N_1}), Y_2(t_{21}), \cdots, Y_n(t_{n1}), \cdots, Y_n(t_{nN_n}))^T$$
$$\boldsymbol{\varepsilon} = (\varepsilon_1(t_{11}), \cdots, \varepsilon_1(t_{1N_1}), \varepsilon_2(t_{21}), \cdots, \varepsilon_n(t_{n1}), \cdots, \varepsilon_n(t_{nN_n}))^T$$

for the latter vectors of dimension $\sum_{i=1}^{n} N_i \times 1 \equiv N \times 1$.

We further denote by $\mathbf{Z}_k$ the matrix of dimension $N \times L_k$ consisting of all elements

$$(\mathbf{Z}_k)_{ij,\ell} = X_i^{(k)}(t_{ij}) B_\ell^{(k)}(t_{ij}) \qquad i = 1, \cdots, n,\ j = 1, \cdots, N_i,\ \ell = 1, \cdots, L_k.$$

We have $(p+1)$ such matrices and stack these into one single big structure

$$\mathbf{Z} = [\mathbf{Z}_0 \mathbf{Z}_1 \mathbf{Z}_2 \cdots \mathbf{Z}_p],$$

of dimension $N \times \left(\sum_{k=0}^{p} L_k \right)$. Finally we denote

$$\boldsymbol{\gamma}^* = \left(\boldsymbol{\gamma}_0^*, \cdots, \boldsymbol{\gamma}_p^*\right)^T,$$

of dimension $(\sum_{k=0}^{p} L_k) \times 1$.

With all the notations introduced above we can write the (approximate) model of observations in (2.4) in matrix form as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}, \tag{2.5}$$

which is now a linear model in $\boldsymbol{\gamma}^*$ in which the variance-covariance matrix of the error term has the structure

$$\boldsymbol{\Sigma}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_N,$$

where $\boldsymbol{I}_N$ denotes the diagonal matrix of dimension $N \times N$ with ones on the diagonal.

In the sequel we work with the goodness-of-fit quantity

$$\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \left(Y_i(t_{ij}) - \sum_{k=0}^{p} \sum_{\ell=1}^{L_k} \gamma_{k,\ell} X_i^{(k)}(t_{ij}) B_\ell^{(k)}(t_{ij}) \right)^2 \equiv \|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2, \tag{2.6}$$

where we put $\widetilde{\mathbf{Y}} = \mathbf{W}^{1/2}\mathbf{Y}$ and $\widetilde{\mathbf{Z}} = \mathbf{W}^{1/2}\mathbf{Z}$ with $\mathbf{W}$ the matrix of dimension $N \times N$ consisting of all diagonal matrices $\mathbf{W}_i$ of dimension $N_i \times N_i$ containing N_i^{-1} on the diagonal elements. The weights in this weighted squared ℓ_2 goodness-of-fit measure allow us to treat each subject equally, while using $\widetilde{\mathbf{Y}} = \mathbf{W}^{1/2}\mathbf{Y}$ and $\widetilde{\mathbf{Z}} = \mathbf{W}^{1/2}\mathbf{Z}$ allows us to simplify the general presentation to the case where all weights are equal to 1.

A special situation occurs when, for all individuals, i, observations on the same time points $t_1, \cdots, t_{\widetilde{N}}$ are available, meaning that $N_i = \widetilde{N}$ for all $i = 1, \cdots, n$. In that case $N = n\widetilde{N}$.

Model (1.1) is also related to the model considered in [34] when studying smoothing ℓ_1-penalized estimators for high-dimensional time-course data. They consider linear models with slowly changing high-dimensional $p \times 1$ parameter vector $\boldsymbol{\beta}(t)$:

$$\mathbf{Y}(t_r) = \mathbf{X}(t_r)\boldsymbol{\beta}(t_r) + \boldsymbol{\varepsilon}(t_r), \quad r = 1, \ldots, \widetilde{N}, \tag{2.7}$$

where $\mathbf{X}(t)$ is an $n(t) \times p$ design matrix at time t, $\mathbf{Y}(t)$ is the $n(t)$ dimensional response vector at times t, that is for every time t one has data as in (2.7) with sample size $n(t)$, and finally the $\boldsymbol{\varepsilon}(t)$'s are independent with $\mathrm{E}(\boldsymbol{\varepsilon}(t)) = 0$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}(t)) = \sigma^2 I_{n(t)}$. Assuming that for all t they have the same number of observations $n(t) = \widetilde{N}$, they propose the smoothed LASSO for estimating sparsely $\boldsymbol{\beta}$.

The high-dimensional linear model (2.7) considered in [34] with $n(t) = \widetilde{N}$ is thus a special case of the varying coefficient model (2.2) studied by [18]. Indeed, this is easily seen by taking in (2.2), $N_i = \widetilde{N}$ for all $i = 1, \ldots, n$ (cross-sectional longitudinal data model) and assuming further that the error covariance structure is determined by $\mathrm{Cov}(\varepsilon(t), \varepsilon(s)) = \sigma^2 \delta_{st}$. This remark will be important when we are going to explore the various estimation procedures that have been designed in the literature to treat such models.

The interest is now to study the variable selection problem together with the estimation of the univariate functions $\beta_k(\cdot)$. Given our framework, this is equivalent to selecting and estimating some vector of coefficients $\boldsymbol{\gamma}$ in the linear model (2.5). In the next sections we discuss several grouped regularization methods for this task. For each of the methods we provide a brief discussion on their implementation and computational algorithms as well as on their possible limitations.

3 Grouped LASSO regularization

The first extension of the ideas of penalized regression to problems of grouped variables was proposed by [41] where rather than penalizing individual covariates they proposed penalizing norms of groups of coefficients and called their method the group LASSO. The grouped LASSO procedure in our context consists of minimizing the objective function

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \lambda \sum_{k=1}^{p} w_k \|\boldsymbol{\gamma}_k\|_2, \tag{3.1}$$

where $w_k = \sqrt{L_k}$, with respect to the vector of parameters $\boldsymbol{\gamma}$. Note that we are not penalizing the intercept function $\beta_0(\cdot)$ parameterized by the vector $\boldsymbol{\gamma}_0$, since this term is not to be selected. Denote by $\widehat{\boldsymbol{\gamma}}$ the solution of this optimization problem.

Minimization of (3.1) is equivalent to minimization of

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}}-\widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2+\sum_{k=1}^{p}\lambda_k w_k\|\boldsymbol{\gamma}_k\|_2^2+\nu\sum_{k=1}^{p}\frac{1}{\lambda_k}, \tag{3.2}$$

with the constraints that $\lambda_1,\cdots,\lambda_p>0$ and $\nu>0$.

The above equivalent reformulation of the minimization problem given in (3.1) was first noticed and proved in the context of smoothing splines ANOVA (SS-ANOVA) models where the COmponent Selection Shrinkage Operator (COSSO) was introduced as a variable selection method in SS-ANOVA models (see [32]). The only difference with the COSSO smoothing spline approach is the penalty, which here is a weighted sum of the ℓ_2 norms of the vectors $\boldsymbol{\gamma}_k$ instead of a specific squared projection norm used in the COSSO method. The proposed new penalty penalizes the fitted model more straightforwardly through the norm of the vector of fitted coefficients of each group component. Such a penalty therefore encourages sparsity at the group level. The equivalence between (3.1) and (3.2) can be proved along the same lines as in Lemma 1 of [1] who studied additive models with P-splines. Note that the first penalty term in (3.2), namely $\sum_{k=1}^{p}\lambda_k w_k\|\boldsymbol{\gamma}_k\|_2^2$ involves $\|\boldsymbol{\gamma}_k\|_2^2$ instead of $\|\boldsymbol{\gamma}_k\|_2$ as in (3.1). In order to minimize (3.2) with respect to the vector $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_p)$ and the sequence of group coefficients $(\boldsymbol{\gamma}_k)_{k=0,\ldots,p}$ one iterates between minimizing (3.2) for fixed $\boldsymbol{\lambda}$ (ridge regression) and minimizing (3.2) for fixed $(\boldsymbol{\gamma}_k)_{k=0,\ldots,p}$ under the positivity constraint on the components of $\boldsymbol{\lambda}$ (nonnegative garrote). We will review in Section 6 how the above results may be exploited to provide effective algorithms for computing a minimizer of the original grouped LASSO minimization problem but also group bridge variable selection.

We would like to mention here a different and promising approach for solving the grouped LASSO regularization problem, that is based on an iterative projection method for structured sparsity regularization used in the machine learning community (see, e.g. [40]). The method in [40] is based on the spectral projected-gradient algorithm originally developed by [7]. Instead of the regularized version (3.1) of the grouped LASSO problem, they consider the following constrained version

$$\begin{aligned}&\min_{\boldsymbol{\gamma}}\frac{1}{2n}\|\widetilde{\mathbf{Y}}-\widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2\\&\sum_{k=1}^{p}\lambda_k w_k\|\boldsymbol{\gamma}_k\|_2\le\tau,\end{aligned} \tag{3.3}$$

where τ is a positive constrain parameter, that they iteratively solve using a spectral gradient projection. Basically, given a current iterate $\boldsymbol{\gamma}^{(j)}$ their iterated solution is defined as

$$\boldsymbol{\gamma}^{(j+1)}=\mathit{\Pi}\left(\boldsymbol{\gamma}^{(j)}+\alpha\widetilde{\mathbf{Z}}^t(\widetilde{\mathbf{Y}}-\widetilde{\mathbf{Z}}\boldsymbol{\gamma}^{(j)})\right), \tag{3.4}$$

where the step length $\alpha > 0$ is some step size to be optimized to ensure sufficient descent (with a backtracking line search, for example) and $\mathit{\Pi}(\cdot)$ is the projection operator defined as

$$\mathit{\Pi}(\mathbf{u}) = \left\{ \text{argmin}_{\mathbf{x}} \|\mathbf{u} - \mathbf{x}\|_2 \text{ subject to } \sum_{k=1}^{p} \lambda_k w_k \|\boldsymbol{\gamma}_k\|_2 \leq \tau \right\}.$$

This algorithm is simple to implement, has low memory requirements and seems to be competitive with the more elaborated Cosso-based algorithm that is usually used in the statistical literature (see [19]).

For the grouped LASSO problem in (3.1) we aim at applying the theoretical results established by [33]. We investigate the variable selection consistency as well as the estimation consistency, i.e. the asymptotic property that the method can correctly select important variables with probability approaching one and that the convergence rates for the nonzero coefficients are the same as the oracle estimator (the estimator when the important variables are known before carrying out statistical analysis). In order to apply the results in [33] we need to make sure that the columns of the matrix $\widetilde{\mathbf{Z}}$ are standardized. This is done easily by replacing $\widetilde{\mathbf{Z}}$ by $\widetilde{\mathbf{Z}}\boldsymbol{D}^{-1}_{\|\widetilde{\mathbf{Z}}\|_2}$ where $\boldsymbol{D}_{\|\widetilde{\mathbf{Z}}\|_2}$ is the diagonal matrix of dimension $(\sum_{k=0}^{p} L_k) \times (\sum_{k=0}^{p} L_k)$ consisting of all diagonal sub matrices of dimension $L_k \times L_k$ with $\|\widetilde{\mathbf{Z}}_k\|_2/\sqrt{n}$ on the diagonal where $\widetilde{\mathbf{Z}}_k = \mathbf{W}_k^{1/2}\mathbf{Z}_k$, for $k = 0, \cdots, p$. From now on we assume that the matrix $\widetilde{\mathbf{Z}}$ has been standardized from the start. Let us remark that the results in [33] are given for a loss function different from ours when N_j, $j = 1, \ldots, p$, are different. Nevertheless all their results can be obtained in our case.

We now explain with more details what is meant by variable selection consistency and estimation consistency in the presented framework. Denote by

$$S = \{k : \|\boldsymbol{\gamma}_k^*\|_\infty \neq 0, k = 1, \cdots, p\}, \tag{3.5}$$

the set of all varying coefficient variables that are non-null, where we used the standard notation $\|\boldsymbol{\gamma}_k\|_\infty = \max_{1 \leq \ell \leq L_k} |\gamma_{k,\ell}|$. Denote by $s_N = |S|$, the number of elements in S. Since $p = p_N$, s_N obviously depends on N. The sparsity assumption means that $s_N << p_N$. An estimator is said to be *variable selection consistent* if it can correctly recover the sparsity pattern with probability going to one, i.e.

$$P\left\{S\left(\widehat{\boldsymbol{\gamma}}\right) = S(\boldsymbol{\gamma}^*)\right\} \to 1, \quad \text{as } N \to \infty,$$

where $S(\boldsymbol{\gamma}^*) = S$ as defined in (3.5), and $S\left(\widehat{\boldsymbol{\gamma}}\right)$ is defined similarly using $\widehat{\boldsymbol{\gamma}}$.

An estimator is ℓ_2-estimation consistent if

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \xrightarrow{\text{P}} 0 \quad \text{as } N \to \infty.$$

Let us introduce $\rho_N^* = \min_{j \in S} \|\boldsymbol{\gamma}_j^*\|$. Further denote by $\widetilde{\mathbf{Z}}_S$ the large matrix formed by stacking the columns of $\widetilde{\mathbf{Z}}$ whose indexes belong to S. We need to introduce the following assumptions

(C1) $\Lambda_{\min}(\frac{1}{n}\widetilde{\mathbf{Z}}_S^T\widetilde{\mathbf{Z}}_S) \geq C > 0$, where $\Lambda_{\min}(\mathbf{A})$ denotes the minimum eigenvalue of the matrix $\mathbf{A}$.
(C2) $\exists\ 0 < \delta < 1$, $\max_{k\in S^C} \|(\widetilde{\mathbf{Z}}_k^T\widetilde{\mathbf{Z}}_S)(\widetilde{\mathbf{Z}}_S^T\widetilde{\mathbf{Z}}_S)^{-1}\|_{2,2} \leq 1-\delta$, where $\|\mathbf{A}\|_{a,b} = \sup_{\mathbf{x}} \|\mathbf{A}\mathbf{x}\|_{\ell_a}/\|\mathbf{x}\|_{\ell_b}$, $1 \leq a, b \leq \infty$.
(C3) $L_k \to +\infty$, $k = 0, \ldots, p_N$, and $\overline{L}_N = o(N)$, where $\overline{L}_N = \max_{k=0,\ldots,p_N} L_k$.
(C4)

$$\frac{\lambda_N^2 N}{\log((p_N - s_N)\overline{L}_N)} \to +\infty.$$

(C5)

$$\frac{1}{\rho_N^*}\left\{\sqrt{\frac{\log(s_N\overline{L}_N)}{N}} + \lambda_N\sqrt{\overline{L}_N}\left\|\left(\frac{1}{N}\widetilde{\mathbf{Z}}_S^T\widetilde{\mathbf{Z}}_S\right)^{-1}\right\|_{\infty,\infty}\right\} \to 0.$$

Note here that (C2) is a variant of the irrepresentable condition used in several papers involving the variable selection consistency. Applying Theorem 3.1 in [33], we obtain the following result.

Theorem 1 *Under conditions* **(C1-C5)**, *the grouped* LASSO *estimator is variable selection consistent.*

For the estimation consistency we need to add and replace condition (C4) by the following assumption:

(C6) $$\kappa = \min_{S_0\subseteq\{1,\ldots,p\}:|S_0|\leq s_N} \min_{\{\boldsymbol{\gamma}:\sum_{j\in S_0^c}\sqrt{L_j}\|\boldsymbol{\gamma}_j\|_2\leq 3\sum_{j\in S_0}\sqrt{L_j}\|\boldsymbol{\gamma}_j\|_2\}} \frac{\left\|\widetilde{\mathbf{Z}}\boldsymbol{\gamma}\right\|_2}{\sqrt{n}\sqrt{\sum_{j\in S_0} L_j\|\boldsymbol{\gamma}_j\|_2^2}} > 0.$$

This assumption on the Gram matrix $\widetilde{\mathbf{Z}}$ is very similar to the restricted eigenvalue assumption as in [6] that is needed to guarantee nice statistical properties of the Lasso selector under a sparsity scenario. One can find in the paper cited above some simple sufficient conditions for such an assumption to hold. Note however, that such conditions are generally computationally intractable.

We now may apply Theorem 4.3 together with Remark 4.4 in [33] to obtain the following result.

Theorem 2 *Under condition* **(C6)**, *let* $\varepsilon_i(t_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, *be independent identically normal distributed of mean 0 and variance* σ^2. *If*

$$\lambda_N = A\sigma\sqrt{\frac{\log\sum_{k=1}^{p_N} L_k}{N}},$$

for some $A > 2\sqrt{2}$, *then with probability at least* $1-(\sum_{k=1}^{p_N} L_k)^{1-A^2/8}$, *we have*

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 \leq \frac{144A^4\sigma^4 s_N^2\overline{L}_N^2}{\kappa^4}\frac{\log\sum_{k=1}^{p_N} L_k}{N}.$$

For example, assuming that $N_i = \tilde{N}$, for $i = 1, \ldots, n$, if one takes $p_N = O(n^\beta)$ and $L_k = O(\tilde{N}^\alpha)$ for all k with $0 < \alpha, \beta < 1/2$, in Theorem 2, one guarantees the asymptotic consistency of $\hat{\boldsymbol{\gamma}}$. As noted by a referee, while according to the above theorems the grouped LASSO enjoys nice properties in terms of estimation consistency and variable selection, it can not be both at the same time, i.e. it does not possess the oracle property. Indeed, looking carefully at condition (C4) and the rate required for λ_N in Theorem 2, it is clear that in order that the two conditions hold simultaneously one must require that $(p_N - s_N) \to 0$ as $N \to \infty$, meaning that the sparsity assumption is not true. The grouped LASSO therefore behaves similarly to the standard LASSO. One should however note that Huang and Zhang [27] showed recently that the grouped LASSO can be better than the standard LASSO under an assumption of strong group sparsity together with a group sparse eigenvalue condition on the design matrix.

Since we are estimating some functional coefficient β_k, it is natural to study the rate of convergence of the grouped LASSO estimator. When using our B-splines setup, the following holds ([12]):

$$\begin{aligned}\left\|\hat{\beta}_k - \beta_k\right\|_{L_2}^2 &\leq \left\|\sum_{\ell=1}^{L_k}(\hat{\gamma}_{k\ell} - \gamma^*_{k\ell})B_\ell^{(k)}\right\|_{L_2}^2 + \left\|\beta_k - \sum_{\ell=1}^{L_k}\gamma^*_{k\ell}B_\ell^{(k)}\right\|_{L_2}^2 \\ &= \frac{1}{L_k}\left\|\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}^*_k\right\|_2^2 + O(L_k^{-4}),\end{aligned}$$

where $\|g\|_{L2}$ denotes $\{\int g^2(x)dx\}^{1/2}$ the L_2-norm of a function g.

Using Theorem 3.2 of [33], we obtain the following theorem:

Theorem 3 *Under conditions of Theorem 3.2 in [33], we have*

$$\left\|\hat{\beta}_k - \beta_k\right\|_{L_2}^2 = O_P\left(\frac{s_N^2\overline{L}_N^2\log\sum_{k=1}^{p_N}L_k}{NL_k} + L_k^{-4}\right).$$

Note that, assuming again that $N_i = \tilde{N} = \exp(n^{1-\eta})$, for $i = 1, \ldots, n$, if one takes $p_N = O(n^\beta)$ and $L_k = O(\tilde{N}^{1/5})$ for all k with $0 < \eta < \beta < 1/2$, in Theorem 3, one guarantees the optimal nonparametric asymptotic rate for the functional coefficients.

Let us remark that [46] propose a grouped LASSO type of variable selection method in the context of varying coefficient models. They consider however a penalty term that is equal to $\sqrt{\sum_{k=1}^{p}\boldsymbol{\gamma}_k^T\mathbf{R}_k\boldsymbol{\gamma}_k}$ with $\mathbf{R}_k$ being an $L_k \times L_k$ symmetric positive definite matrix. In the context of B-splines approximations and for components β_k that are twice continuously differentiable the evaluation of these quadratic forms R_k involves the inner products of the B-spline basis functions.

Similar quadratic forms, but involving up to the second order derivatives of the B-spline basis functions, have been used in the work by [3] . The numerical evaluation of these quadratic forms is not as easy task as it appears. Indeed,

when the knots that are used for the B-splines approximation are equi-spaced then one may use a recursive difference equation defining the elements of the B-spline basis and then an appropriate algorithm described by [5] for the numerical evaluation of the $\mathbf{R}_k$'s. In [46] it is shown that, under appropriate conditions, this grouped LASSO procedure, with penalty term $\sqrt{\sum_{k=1}^{p} \boldsymbol{\gamma}_k^T \mathbf{R}_k \boldsymbol{\gamma}_k}$, selects a model of the right order of dimensionality and is estimation consistent. However, this procedure is (under their assumptions) in general not selection consistent. In order to improve the selection results, they propose to apply an adaptive grouped LASSO penalty, based on a given initial estimator. But they need to add some critical condition on this initial estimator used in the weights of the adaptive grouped LASSO penalty to have the oracle selection property (see condition (C5) in [46]). Indeed this condition is very difficult to establish. For these two reasons, we decided not to compare our approach with the method of [46].

4 Grouped SCAD regularization

We have seen previously that the group LASSO asymptotically suffers from the same drawbacks as the LASSO, due to the lack of the oracle property. This is not true for the group SCAD or group BRIDGE to be discussed hereafter, which share the oracle property even when the dimension of the predictive variables is large.

A SCAD procedure for variable selection in nonparametric varying coefficient models has been discussed in [45]. An application to microarray gene expression data can be found in [44].

The approach taken in both papers is as follows. Denote by $p_\lambda(v)$ the function providing the SCAD penalty. For $v \geq 0$, the penalty is defined as

$$p_\lambda(v) = \begin{cases} \lambda v & \text{if } 0 \leq v \leq \lambda, \\ -\dfrac{v^2 - 2a\lambda v + \lambda^2}{2(a-1)} & \text{if } \lambda < v < a\lambda, \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } v \geq a\lambda. \end{cases} \tag{4.1}$$

A common choice for a is 3.7. A Taylor expansion of $p_\lambda(v)$ for v around v_0 leads to

$$p_\lambda(v) \approx p_\lambda(v_0) + \frac{1}{2}\frac{p'_\lambda(v_0)}{v_0}(v^2 - v_0^2), \tag{4.2}$$

as explained in [15].

A grouped SCAD procedure under the model (2.5) is defined by minimizing

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{k=1}^{p} p_\lambda(\omega_k \|\boldsymbol{\gamma}_k\|_2), \tag{4.3}$$

with $p_\lambda(\cdot)$ the SCAD penalty function in (4.4), and $\omega_k = \sqrt{L_k}$ as before.

It is worth mentioning here that the grouped MCP (Minimax Concave Penalization) method introduced in [48]) is similar to grouped SCAD; this method has the same form as (4.3), only with the SCAD penalty replaced with the MC penalty defined by

$$p_\lambda(v) = \begin{cases} \lambda v - \frac{v^2}{2a} & \text{if } 0 \le v \le a\lambda, \\ \dfrac{a\lambda^2}{2} & \text{if } v > a\lambda. \end{cases} \tag{4.4}$$

Substitution of (4.2) into (4.3) then leads to

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{k=1}^{p}\left\{ p_\lambda(\omega_k\|\boldsymbol{\gamma}_k^{(0)}\|_2) + \frac{1}{2}\frac{p'_\lambda(\omega_k\|\boldsymbol{\gamma}_k^{(0)}\|_2)}{\|\boldsymbol{\gamma}_k^{(0)}\|_2}\omega_k\left(\|\boldsymbol{\gamma}_k\|_2^2 - \|\boldsymbol{\gamma}_k^{(0)}\|_2^2\right)\right\},$$

which as minimization problem is equivalent to the minimization problem

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \frac{1}{2}\sum_{k=1}^{p}\frac{p'_\lambda(\omega_k\|\boldsymbol{\gamma}_k^{(0)}\|_2)}{\|\boldsymbol{\gamma}_k^{(0)}\|_2}\omega_k\|\boldsymbol{\gamma}_k\|_2^2, \tag{4.5}$$

with $\boldsymbol{\gamma}_k^{(0)}$ starting vectors.

Defining a diagonal matrix

$$\boldsymbol{V}_\lambda\left(\boldsymbol{\gamma}^{(0)}\right) = \begin{pmatrix} \omega_1\dfrac{p'_\lambda(\omega_1\|\boldsymbol{\gamma}_1^{(0)}\|_2)}{\|\boldsymbol{\gamma}_1^{(0)}\|_2}\boldsymbol{I}_{L_0} & & 0 \\ & \ddots & \\ 0 & & \omega_p\dfrac{p'_\lambda(\omega_p\|\boldsymbol{\gamma}_p^{(0)}\|_2)}{\|\boldsymbol{\gamma}_p^{(0)}\|_2}\boldsymbol{I}_{L_p} \end{pmatrix}$$

of dimension $\left(\sum_{k=1}^{p} L_k\right)\times\left(\sum_{k=1}^{p} L_k\right)$, and letting $\boldsymbol{D}_\lambda\left(\boldsymbol{\gamma}^{(0)}\right) = \mathrm{diag}(0, \boldsymbol{V}_\lambda\left(\boldsymbol{\gamma}^{(0)}\right))$, the $(p+1)\times(p+1)$ diagonal matrix, we rewrite (4.5) as a Ridge-regression problem

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \frac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{D}_\lambda(\boldsymbol{\gamma}^{(0)})\boldsymbol{\gamma}. \tag{4.6}$$

Minimization of (4.6) with respect to $\boldsymbol{\gamma}$ yields

$$\widehat{\boldsymbol{\gamma}} = \frac{1}{2n}\left(\frac{1}{2n}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} + \frac{1}{2}\boldsymbol{D}_\lambda(\boldsymbol{\gamma}^{(0)})\right)^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Y}},$$

and the fitted values

$$\widehat{\mathbf{Y}} = \widetilde{\mathbf{Z}}\widehat{\boldsymbol{\gamma}} = \frac{1}{2n}\widetilde{\mathbf{Z}}\left(\frac{1}{2n}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} + \frac{1}{2}\boldsymbol{D}_\lambda(\boldsymbol{\gamma}^{(0)})\right)^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Y}},$$

which leads to the (approximate) Hat matrix

$$\mathbf{H}(\lambda) = \frac{1}{2n}\widetilde{\mathbf{Z}}\left(\frac{1}{2n}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} + \frac{1}{2}\boldsymbol{D}_\lambda(\widehat{\boldsymbol{\gamma}})\right)^{-1}\widetilde{\mathbf{Z}}^T. \tag{4.7}$$

Relying on this expression for the Hat matrix one can then use Generalized Cross Validation (GCV) techniques for selecting the smoothing parameter λ, as explained in [45].

The above treatment of the grouped SCAD regularization problem hence results into an approximate Ridge regression problem. Of course, for high-dimensional cases where p_N can be larger than N, this ridge alike procedure produces reasonable local minimizers (reasonable in the sense that the resulting estimators while biased have small variance).

Regarding the asymptotic properties of the above group SCAD method, one can show that under similar conditions as (C1), (C2) and (C3) of the previous Section, but assuming further that p and s are fixed and that the p covariate processes as well as the variance of the noise process, are uniformly bounded, then with a choice of $\lambda_N \to 0$ and $N/n^{2/5}\lambda_N \to \infty$ the group SCAD is shown in [45] to be both variable selection and estimation consistent with oracular least squares asymptotic rates. A completely different approach for tackling the grouped SCAD optimization problem in (4.3) is inspired by the work of [29] that is using a so-called ConCave Convex Procedure (CCCP) type of algorithm. However, there has been no investigation of asymptotic properties of estimators derived by this method in the context of high-dimensional models.

5 Grouped Bridge regularization

In the group LASSO the estimates are obtained by applying an ℓ_1 penalty to the ℓ_2 norms of the groups, while for the group SCAD a SCAD penalty is applied to the ℓ_2 norms of the groups. This fundamentally differs from the group BRIDGE where the penalty is applied to the ℓ_1 norms of the groups.

More precisely, in Bridge regression the penalty function equals, for $v > 0$,

$$p_\lambda(v) = \lambda|v|^q \qquad \text{with } 0 < q < 1. \tag{5.1}$$

The grouped Bridge approach then consists of minimizing the objective function

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{k=1}^{p} p_\lambda(\omega_k\|\boldsymbol{\gamma}_k\|_1), \tag{5.2}$$

and an algorithm for solving this optimization problem is obtained from [9].

Of course, one could instead minimize the objective function

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \sum_{k=1}^{p} p_\lambda(\omega_k\|\boldsymbol{\gamma}_k\|_2), \tag{5.3}$$

and produce meaningful estimators but this problem is not addressed in the present paper. For a brief discussion on such concave 2-norm group selection methods the reader may refer to a recent review article [26] by Huang *et al.*

Asymptotic properties of Bridge estimators with $0 < q < 1$ when the number of covariates p_N may increase to infinity with N have been studied by [23] extending the results of [30] to infinite-dimensional parameter settings. They show that, for $0 < q < 1$, the Bridge estimators can correctly select covariates with nonzero coefficients and that, under appropriate conditions on the growth rates of p_N and λ_N, the estimators of nonzero coefficients have the same asymptotic distribution as they would have if the zero coefficients were known in advance. Therefore, Bridge estimators have the oracle property of [15] and [16]. The permitted rate of growth of p_N depends on the penalty function form specified by q. The above authors require that $p_N < N$ that is, the number of covariates must be smaller than the sample size, which is needed for identification and consistent estimation of the regression parameters. However, if there is a special suitable structure in the covariate matrix (the partial orthogonality condition), they show that it is possible to achieve consistent variable selection and estimation, even in the case $p_N > N$. The estimation is performed in two steps: first, they use a marginal bridge estimator to select the covariates with nonzero coefficients; and then they estimate the regression model with these selected covariates. The interested reader is referred to their paper for further details.

6 Grouped Cosso regularization

The grouped Cosso regularization procedure consists of minimizing the objective function

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \lambda \sum_{k=1}^{p} w_k \|\boldsymbol{\gamma}_k\|_2, \tag{6.1}$$

where w_k are positive fixed weights. Note that considering weights in the block penalty norm is important in practice as those have an influence regarding the consistency of the estimator (see [4]). Note also that with probability tending to one, if for example $\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}$ is invertible, there is a unique minimum. Efficient exact algorithms exist for the regular lasso, i.e., for the case where all group dimensions, and therefore the weights w_k, are equal to one. They are based on the piecewise linearity of the set of solutions as a function of the regularization parameter λ (see [14]). For the grouped lasso, however, the path is only piecewise differentiable, and following such a path is not as efficient as for the lasso. Other algorithms have been designed to solve problem (6.1) for a single value of λ. The grouped cosso like algorithm relies upon the equivalent cosso formulation of (6.1),

$$\frac{1}{2n}\|\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}\|_2^2 + \mu \sum_{k=1}^{p} \lambda_k w_k \|\boldsymbol{\gamma}_k\|_2^2 + \nu \sum_{k=1}^{p} \frac{1}{\lambda_k}, \tag{6.2}$$

and the algorithm that one may use can be summarized as follows:

1. Fix $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p) = (1, \ldots, 1)$, find the best μ to minimize GCV in the corresponding ridge like criterion, say μ_0 and let $\boldsymbol{\gamma}_{\mu_0}$ the corresponding coefficients.
2. For fixed $\boldsymbol{\gamma}$, solve for $\boldsymbol{\lambda}$ via quadratic programming.
3. For fixed $\boldsymbol{\lambda}$, solve for $\boldsymbol{\gamma}$ using the normal equation or the Gaussian profile likelihood if $\boldsymbol{\lambda}$ contains zero entries.
4. Iterate between steps 2 and 3 until convergence. The final solution corresponds to ν that gives the minimum GCV score.

Such grouped COSSO algorithm is also discussed in details in the recent paper [25] devoted to group BRIDGE variable selection. We do not consider further the grouped COSSO regularization method and this algorithm, since the results obtained with such an approach are very unstable (see for example [1]).

7 Numerical study

In this section, we first carry out a simulation study to compare the performances of the grouped Bridge, the grouped SCAD, the grouped MCP and three implementations of the grouped LASSO methods.

The first grouped LASSO method, denoted by `grlasso 1`, was implemented by [8] and the second grouped LASSO method, called `grlasso 2`, was implemented by [35]. The third consits of post-model selection which apply ordinary least squares to the model selected by first-step `grlasso 1` penalized estimators. We call this method `grlasso-ols`. In [9], the authors describe a general penalization approach based on a local coordinate descent algorithm. This is further developed in [26] where the authors propose a procedure that also includes the methods `gbridge` (with $q = 1/2$), `grscad` and `grMCP`.

In a second section, we illustrate the use of the grouped LASSO method `grlasso-ols`, on two real data examples: a data set concerning the study of AIDS and the Boston Housing data set.

To chose the tuning parameter λ, we use a BIC-type criterion with effective number of model parameters estimated as in [8]. More precisely we use the following criterion

$$\log\left(\frac{\mathrm{RSS}_\lambda}{\Sigma_{i=1}^n N_i}\right) + \frac{\log(\Sigma_{i=1}^n N_i)}{\Sigma_{i=1}^n N_i}\mathrm{df}_\lambda, \tag{7.1}$$

where the residual sum of squares (RSS_λ) is the sum of squares of residuals associated with the estimate $\widehat{\boldsymbol{\gamma}}$ and df_λ is the number of nonzero coefficients of $\widehat{\boldsymbol{\gamma}}$. It is worth noting that in [8], one uses a criterion without the logarithm applied to the normalized RSS_λ. In our simulation study we have observed however that such a criterion (without the log function) leads in some situations to very bad results in comparison with the criterion involving the logarithm. Moreover the criterion (7.1) leads to results similar to those obtained when applying the LSA (Least squares approximation) with a BIC type penalty as proposed by [42].

7.1 Simulation study and comparison

Table 1 *Selection model ability. First column (λ): mean value of λ. Second one (S): mean of number of variables selected. Third one (FP): mean of number of false positives (truly zero variables that were selected). Fourth one (FN): mean of number of false negatives (truly nonzero variables that were not selected). Fifth one (CF): percentage of the experiments with model perfectly identified. Sixth one (ME): mean of the model error and in brackets, its standard deviation.*

	λ	S	FP	FN	CF	ME
$\sigma = 1$						
`gbridge`	0.0180	4.0420	0.0420	0.0000	0.9580	0.0122 (0.0033)
`grscad`	0.0280	4.0000	0.0000	0.0000	1.0000	0.0106 (0.0027)
`grMCP`	0.0280	4.0000	0.0000	0.0000	1.0000	0.0106 (0.0027)
`grlasso 1`	0.0250	5.5440	1.5440	0.0000	0.2420	0.2410 (0.0351)
`grlasso 2`	0.0590	4.1340	0.1340	0.0000	0.8700	3.1292 (0.0430)
`grlasso-ols`	0.7670	4.0000	0.0000	0.0000	1.0000	0.0106 (0.0027)
$\sigma = 2$						
`gbridge`	0.0400	4.0580	0.0580	0.0000	0.9480	0.0482 (0.0138)
`grscad`	0.0560	4.0000	0.0000	0.0000	1.0000	0.0423 (0.0106)
`grMCP`	0.0560	4.0000	0.0000	0.0000	1.0000	0.0423 (0.0106)
`grlasso 1`	0.0550	4.3040	0.3040	0.0000	0.7220	0.5754 (0.0624)
`grlasso 2`	0.1100	4.3900	0.3900	0.0000	0.6760	3.5590 (0.1258)
`grlasso-ols`	0.7620	4.0000	0.0000	0.0000	1.0000	0.0424 (0.0106)
$\sigma = 6$						
`gbridge`	0.2090	4.0900	0.0900	0.0000	0.9180	0.4434 (0.1298)
`grscad`	0.1590	4.6280	0.6280	0.0000	0.6800	1.4763 (1.0481)
`grMCP`	0.1650	4.1940	0.1940	0.0000	0.8840	0.6769 (0.7342)
`grlasso 1`	0.1700	4.0220	0.0220	0.0000	0.9780	1.8663 (0.2772)
`grlasso 2`	0.3350	4.1200	0.1200	0.0000	0.8820	6.2692 (0.4700)
`grlasso-ols`	0.7640	4.0000	0.0000	0.0000	1.0000	0.3809 (0.0953)

We consider a model similar to the one used by [24] and [45], given by

$$Y_i(t_{ij}) = \beta_0(t_{ij}) + \sum_{k=1}^{23} \beta_k(t_{ij}) X_i^{(k)}(t_{ij}) + \varepsilon_i(t_{ij}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, \widetilde{N}.$$

The coefficients $\beta_k(t)$, $k = 0, \ldots, 3$, correspond to the intercept term and the three true relevant variables and are given by

$$\beta_0(t) = 15 + 20 \sin\left(\tfrac{\pi t}{60}\right), \qquad \beta_1(t) = 2 - 3 \cos\left(\tfrac{\pi(t-25)}{15}\right),$$
$$\beta_2(t) = 6 - 0.2t, \qquad \beta_3(t) = -4 + \tfrac{(20-t)^3}{2000}, \quad t \in [1, 30].$$

The remaining coefficients are given by $\beta_k(t) = 0$, $k = 4, \ldots, 23$. The time points t_{ij} are $1, 2, \ldots, 30$ ($\widetilde{N} = 30$) and $n = 100$. The three relevant variables $X_i^{(k)}(t)$, $k = 1, \ldots, 3$, are simulated in the following way. At any point t, the variable $X_i^{(1)}(t)$ is sampled uniformly from $[t/10, 2 + t/10]$. Conditioning on $X_i^{(1)}(t)$, the variable $X_i^{(2)}(t)$ is a centered Gaussian random variable with variance given by $(1 + X_i^{(1)}(t))/(2 + X_i^{(1)}(t))$. The variable $X_i^{(3)}(t)$ is independent

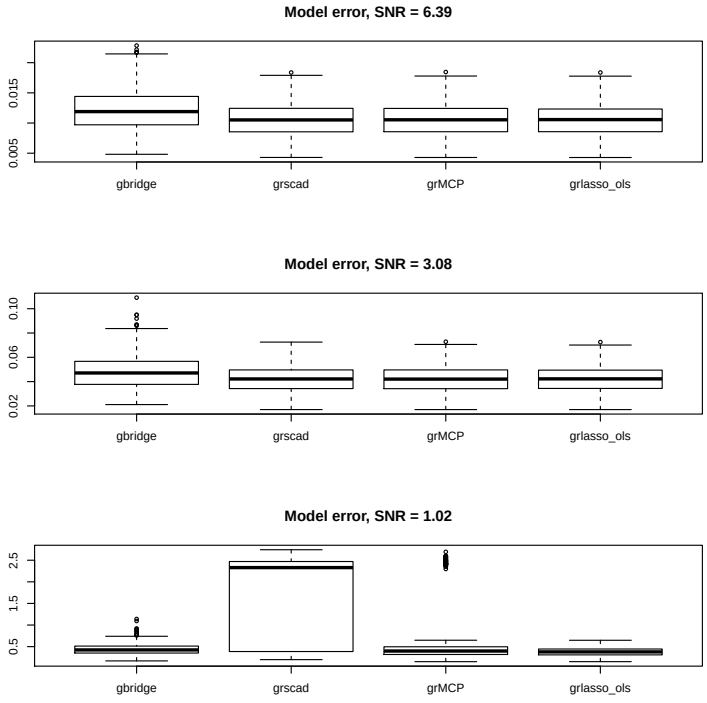

Fig. 7.1 *Boxplot of the model errors for four methods.*

of $X_i^{(1)}$ and $X_i^{(2)}$ and is a Bernoulli random variable with success rate equal to 0.6. The irrelevant variables $X_i^{(k)}$, $k = 4, \ldots, 23$ are paths of centered Gaussian process with covariance function $\mathrm{Cov}(X_i^{(k)}(t), X_i^{(k)}(s)) = 4\exp(-|t-s|)$; they are independent between them as well as independent of the other first three variables. We chose several levels of noise, $\sigma = 1$, 2 and 6, for the random error. These noise levels correspond to signal-to-noise ratios (SNR) given respectively by 6.39, 3.08 and 1.02. The SNR is defined by $\boldsymbol{\gamma}^{*T}\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma}^*/N$ (see [13]).

For each simulated data set, we use cubic splines with five equidistant internal knots. We repeat the simulations 500 times. The simulation results are summarized in Table 1. We present the mean value of the tuning parameter λ, the average number of variables selected, the average number of truly zero variables that were selected (false positives), the average number of truly nonzero variables that were not selected and the mean and standard deviation of the model error. The model error is defined similarly as in [8] and is given by $(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^T\mathbf{Z}^T\mathbf{Z}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)/N$. Figure 7.1 depicts the boxplots of the model errors for all the methods except for `grlasso 1` and `grlasso 2`, since the errors for these methods are very (too) high compared to these for the other four methods. See also Table 1. In Figure 7.2 we present typical curve

(median-performing curve over all the simulations) estimates of the first four coefficients for a signal-to-noise ratio given by 1.25.

Looking at Table 1 we can see that for all the signal-to-noise ratios, the method `grlasso-ols` gives the best results in terms of selection ability and model error compared to the other methods. The `grlasso-ols` is the only method that each time has selected the exact model whatever the level of the noise. The procedures `gbridge`, `grscad` and `grMCP` leads to similar results, except for `grscad` whose results seem to deteriorate when the noise level increases. Finally the `grlasso 1` and `grlasso 2` procedures give relatively correct result in selection model but leads to very bad results in terms of the model error criterion (especially for `grlasso 2`). At first sight, one would believe that the `grlasso-ols` should have exactly the same selection properties as `grlasso 1` but this is not the case. The optimal value of the regularisation parameter is searched over a fine grid of λ values. For each value of the penalty parameter λ on the given λ-grid one applies a model selection using `grlasso 1` and then re-estimates by ordinary least squares the coefficients already selected by `grlasso 1` in the first step. Now the above estimator is reinjected into the BIC-type criterion defined in eq.(7.1), leading to a BIC value that differs from the one obtained in the first step. The optimal λ and the corresponding model selected is the one minimising this BIC criterion over the chosen grid. Therefore the two methods are different in terms of model selection.

7.2 Data analysis

In this subsection we demonstrate the effectiveness of the `grlasso-ols` method in selecting the variables and in estimating the varying coefficients, by considering results from the analysis of two real data sets: the AIDS data set (see [28]) and the Boston Housing data set (see [20]). Also, as suggested by a referee, we have computed pointwise variability bands for the estimated varying coefficients by bootstrap resampling from the original data in a way similar to the one used in [24]. We have treated the selected varying coefficients as if they were known in advance (because of the Oracle property) in order to estimate the bias. However a general and concrete constructive procedure to achieve exact uniform confidence bands for the estimated varying coefficients, with the requested coverage probabilility, is beyond the scope of the present paper. For each data set, together with the estimated varying coefficients we display, for each t, a 95% confidence interval based on a normal approximation with a sample standard error of the estimated varying coefficient computed from 100 bootstrap samples.

7.2.1 AIDS data

In [28] the authors reported on a Multicenter AIDS Cohort Study conducted, in which one obtain repeated measurements of physical examinations, labo-

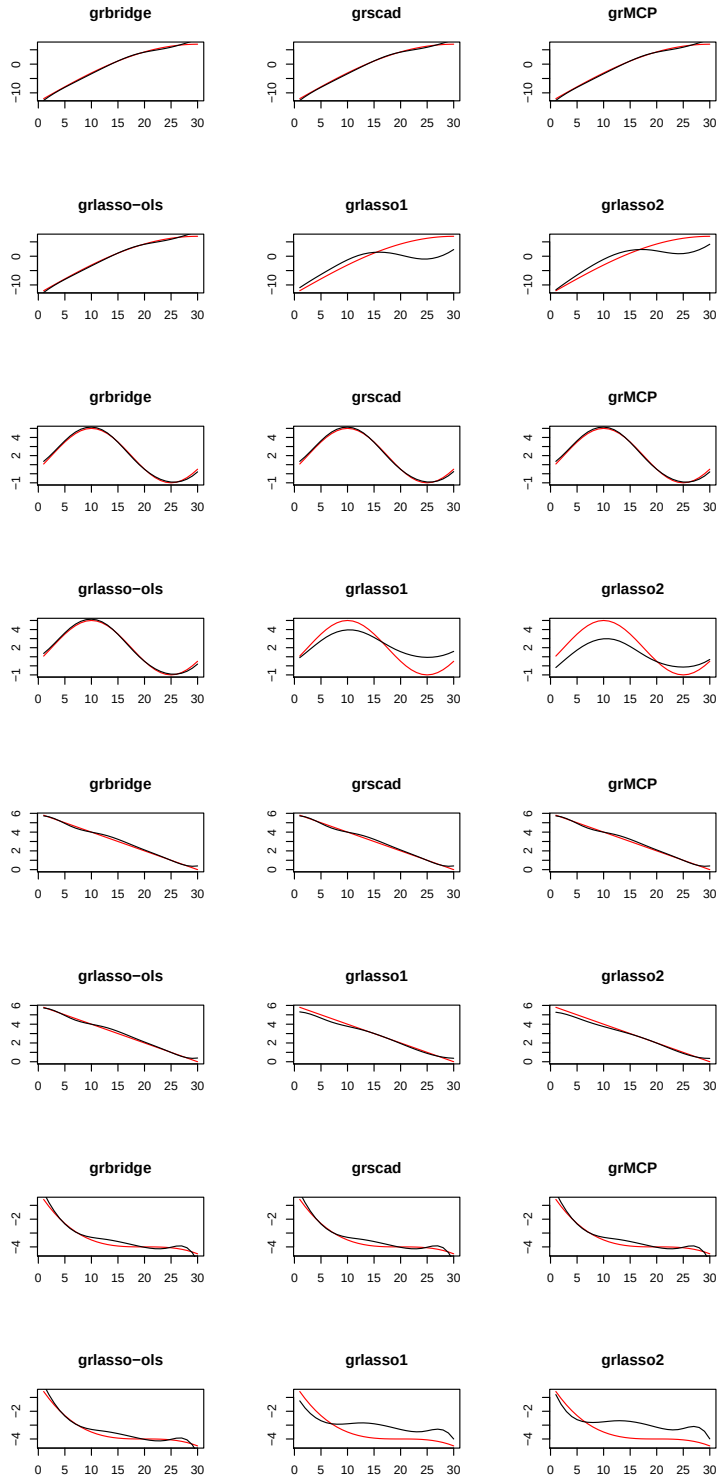

Fig. 7.2 *Typical curve (median-performing curve over all the simulations) estimates of the intercept and the three coefficients corresponding to relevant variables for the five methods ($\sigma = 2$). Red curves correspond to the true coefficient functions and black ones to the estimated coefficient functions.*

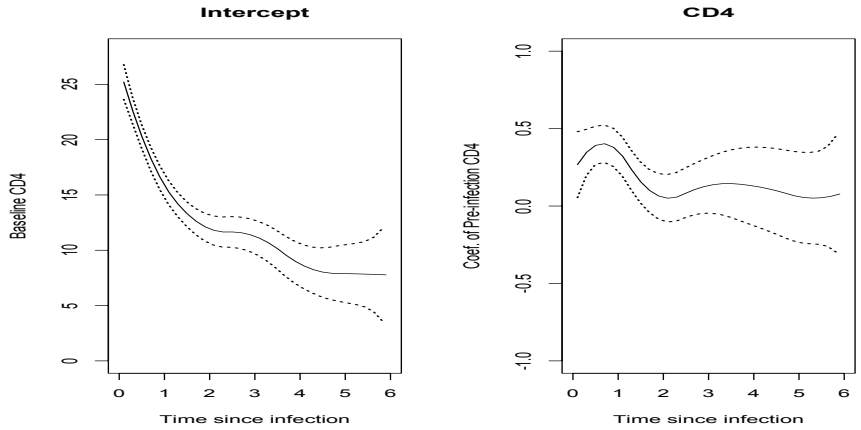

Fig. 7.3 *Application of the* `grlasso-ols` *method to the AIDS data: estimates of the relevant coefficient functions for the intercept and pre-infection CD4 percentages together with their pointwise 95% variability bands.*

ratory results, and CD4 cell counts and percentages of homosexual men who became human immunodeficiency virus (HIV)-positive during 1984 and 1991. This data set is also analyzed by [45] using a varying coefficient model. They use the `gscad` method to select and estimate the varying coefficients. All individuals were scheduled to undergo measurements at semi-annual visits, but because many individuals missed some of their scheduled visits and the HIV infections occured randomly during the study, there were unequal numbers of repeated measurements and different measurement times for each individual. Their analysis focused on the 283 homosexual men who become HIV-positive and aimed to evaluate the effects of cigarette smoking, pre-HIV infection CD4 cell percentage, and age at HIV infection on the mean CD4 percentage after infection. This data set is available in the R package timereg (data(cd4)). As in [45], the `gLASSO 1` method identified two nonzero coefficients (the intercept and the pre-infection CD4 percentages). That indicates that cigarette smoking and age at HIV infection have no effect on the post-infection CD4 percentage. Figure 7.3 shows the two fitted relevant varying coefficients.

7.2.2 Boston data

The Boston Housing data set was analyzed by [20], with the aim to find out whether 'clean air' had an influence on house prices. This data set is available in the R package mlbench (data(BostonHousing)) with 14 variables and 506 cases. As in [43], we consider the median value of owner occupied homes (MEDV) as the response of interest and the proportion of population that has a lower status (LSTAT) as the index variable. We consider the following predictors as: per capita crime rate by town (CRIM), nitric oxides concentration (parts per 10 million, NOX), average number of rooms per dwelling

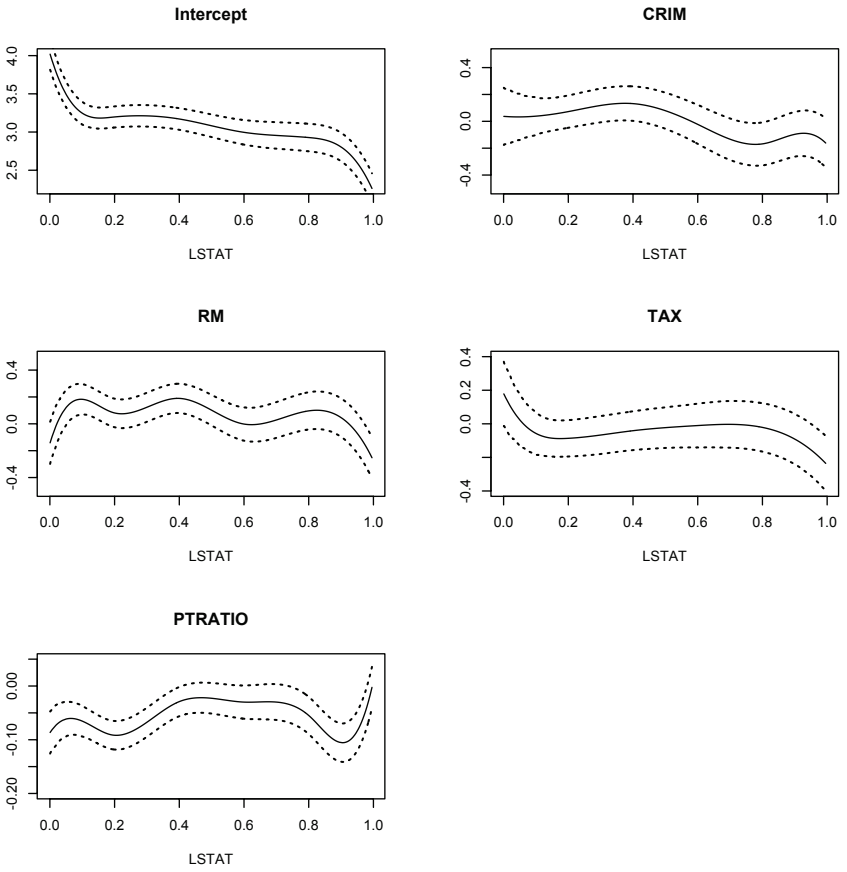

Fig. 7.4 *Application of the* `glrasso-ols` *method to the Boston data: estimates of relevant coefficient functions for the intercept, the crime rate ratio, the average number of rooms, the full-value property-tax rate and the pupil-teacher ratio together with their pointwise 95% variability bands.*

(RM), proportion of owner-occupied units built prior to 1940 (AGE), full-value property-tax rate per 10,000 (TAX) and pupil-teacher ratio by town (PTRATIO). Moreover as in [43] before applying our procedure, both the response and the predictors (except the intercept) are transformed so that their marginal distribution is approximately centered and reduced to an approximate normal distribution. We use Box-Cox transformations. The index variable LSTAT is transformed so that its marginal distribution is approximately uniform on $[0, 1]$.

The authors [43] propose to combine local polynomial smoothing and grouped LASSO to select and estimate the varying coefficients. We obtain results similar to theirs except that we select one supplementary variable (TAX).

Figure 7.4 shows the five fitted relevant varying coefficients together with their 95% pointwise confidence bands.

Acknowledgements The authors thank the editor and two reviewers for their detailed reading of the manuscript and their valuable comments and suggestions that led to a considerable improvement of the paper. Support from the IAP research network nr. P6/03 and P7/06 of the Federal Science Policy, Belgium, is acknowledged. The second author also gratefully acknowledges financial support by the projects GOA/07/04 and GOA/12/014 of the Research Fund KULeuven and the FWO-project G.0328.08N of the Flemish Science Foundation.

References

1. A. Antoniadis, I. Gijbels, and A. Verhasselt. Variable selection in additive models using P-splines. *Technometrics*, Volume **54**, Issue 4, 425–438, 2012.
2. A. Antoniadis, I. Gijbels, and A. Verhasselt. Variable selection in varying coefficient models using P-splines. *Journal of Computational and Graphical Statistics*, 21, issue 3, 638–661, 2012.
3. M. Avalos, Y. Grandvalet, and C. Ambroise. Regularization methods for additive models. *Lecture Notes in Computer Science, Advances in Intelligent Data Analysis V*, 2810:509–520, 2003.
4. F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.
5. M. Bhatti and P. Bracken. The calculation of integrals involving b-splines by means of recursion relations. *Applied Mathematics and Computation*, 172:91–100, 2006.
6. P. J. Bickel, Y. Ritov and A. Tsybakov. A. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, **37**, 4: 1705–1732, 2009.
7. E. G. Birgin, J. Martinez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10:1196–1211, 2000.
8. P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2:369–380, 2009.
9. P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5: 32–253, 2011.
10. B. Brumback and J. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, 93:961–994, 1998.
11. R. Chen and R. S. Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88:298–308, 1993.
12. C. de Boor. *A pratical guide to splines*. Springer, New York, 1978.
13. D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.
14. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–489, 2004.
15. J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.
16. J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annal Statistic*, 32, 2004.
17. J. Fan, C. Zhang, and J. Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics*, 29:153–193, 2001.
18. J. Fan and J.-T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B*, 62:303–322, 2000.
19. M. A. T. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal on Selected Topics in Signal Processing*, 1:586–597, 2007.

20. D. Harrison and D. Rubinfeld. Hedonic prices and the demand for clean air. *J. Environ. Economics & Management*, 5:81–102, 1978.
21. T. J. Hastie and R. J. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55:757–796, 1993.
22. D. Hoover, J. Rice, C. Wu, and L.-P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85:809–822, 1998.
23. J. Huang, J. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Stat.*, 36:587–613, 2007.
24. J. Z. Huang, C. O. Wu, and L. Zhou. Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika*, 89:111–128, 2002.
25. J. Huang, S. Ma, H. Xie, and C.-H. Zhang. A group bridge approach for variable selection. *Biometrika*, **96**, 2, 339–355, 2009.
26. J. Huang, P. Breheny and S. Ma. A Selective Review of Group Selection in High Dimensional Models. *Statistical Science*, **27**, 4, 481–499, 2012.
27. J. Huang, and T. Zhang. The benefit of group sparsity *Ann. Stat.* , **38**, 1978–2004, 2010.
28. R. A. Kaslow, D. G. Ostrow, R. Detels, J. P. Phair, B. F. Polk, and C. R. Rinaldo. The multicenter aids cohort study: Rationale, organization and selected characteristics of the participants. *American Journal of Epi- demiology*, 126:310–318, 1987.
29. Y. Kim, H. Choi, and H. Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103:1665–1673, 2008.
30. K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28:1356–1378, 2000.
31. R. Li and H. Liang. Variable selection in semiparametric regression modeling. *The Annals of Statistics*, 36:261–286, 2008.
32. B. Lin and H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 32:2272–2297, 2006.
33. H. Liu and J. Zhang. On the ℓ_1–ℓ_q regularized regression. *Technical Report*, 2008.
34. L. Meier and P. Bühlman. Smoothing ℓ_1-penalized estimators for high-dimensional time-course data. *Electronic Journal of Statistics*, 1:597–615, 2007.
35. L. Meier, S. van de Geer, and P. Bühlman. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.
36. G. Nürnberger. *Approximation by Spline Functions*. Springer-Verlag, New York, 1989.
37. T. Qingguo and C. Longsheng. Componentwise B-spline estimation for varying coefficient models with longitudinal data. *Statistical Papers*, **53**, 3, 629–652, 2012.
38. J. Ramsay and B. Silverman. *The Analysis of Functional Data*. Springer-Verlag, Berlin, 1997.
39. J. Rice. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14:631–647, 2004.
40. E. van den Berg, M. Schmidt, M. Friedlander, and K. Murphy. Group sparsity via linear-time projection. University of British Columbia, Department of Computer Science, 2008.
41. M. Yuan, and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2006.
42. H. Wang and C. Leng. Unified lasso estimation with least squares approximation. *Journal of American Statistical Association*, 102:1039–1048, 2007.
43. H. Wang and Y. Xia. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104:747–757, 2009.
44. L. Wang, G. Chen, and H. Li. Group scad regression analysis for microarray time course gene expression. *Bioinformatics*, 23:1486–1494, 2007.
45. L. Wang, H. Li, and J. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103:1556–1569, 2008.
46. X. Wei, J. Huang, and H. Li. Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21:1515–1540, 2011.
47. C. Wu, K. Yu, and C. Chiang. A two-step smoothing method for varying coefficient models with repeated measurements. *Annals of the Institute of Statistical Mathematics*, 52:519–543, 2000.
48. C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist*, 38:894–942, 2010.