Electronic Journal of Statistics
Vol. 6 (2012) 810–842
ISSN: 1935-7524
DOI: 10.1214/12-EJS695

Fixed and random effects selection in nonparametric additive mixed models

Randy C. S. Lai

Department of Statistics, University of California at Davis
4118 Mathematical Sciences Building, One Shields Avenue
Davis, CA 95616, USA
e-mail: rcslai@ucdavis.edu

Hsin-Cheng Huang*

Institute of Statistical Science
Academia Sinica, Taipei 115, Taiwan
e-mail: hchuang@stat.sinica.edu.tw

and

Thomas C. M. Lee†

Department of Statistics, University of California at Davis
4118 Mathematical Sciences Building, One Shields Avenue
Davis, CA 95616, USA
e-mail: tcmlee@ucdavis.edu

Abstract: This paper considers the problem of model selection in a nonparametric additive mixed modeling framework. The fixed effects are modeled nonparametrically using truncated series expansions with B-spline basis. Estimation and selection of such nonparametric fixed effects are simultaneously achieved by using the adaptive group lasso methodology, while the random effects are selected by a traditional backward selection mechanism. To facilitate the automatic selection of model dimension, computable expressions for the degrees of freedom for both the fixed and random effects components are derived, and the Bayesian Information criterion (BIC) is used to select the final model choice. Theoretically it is shown that this BIC model selection method is consistent, while computationally a practical algorithm is developed for solving the optimization problem involved. Simulation results show that the proposed methodology is often capable of selecting the correct significant fixed and random effects components, especially when the sample size and/or signal to noise ratio are not too small. The new method is also applied to two real data sets.

AMS 2000 subject classifications: Primary 62G08.
Keywords and phrases: Adaptive group lasso, additive mixed model, Bayesian information criterion, consistency.

Received March 2011.

*Supported in part by National Science Council Taiwan under grants 97-2118-M001-001-MY3 and 100-2628-M-001-004-MY3.
†Supported in part by NSF under Grant DMS 1007520.

1. Introduction

This paper considers the problem of joint selection of fixed and random effects in nonparametric additive mixed models. Suppose for the i-th subject we observe response Y_i and covariates $X_{i1}, \ldots, X_{iP}$ and $Z_{i1}, \ldots, Z_{iQ}$. It is assumed that these measurements are related by the following mixed model:

$$Y_i = \mu + \sum_{p=1}^{P} f_p(X_{ip}) + \sum_{q=1}^{Q} u_q Z_{iq} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where μ is the mean, f_p's are unknown smooth continuous functions that form the fixed component of the model, $\boldsymbol{u} = (u_1, \ldots, u_Q)'$ are random effects, and $\boldsymbol{\epsilon} = (\epsilon_i, \ldots, \epsilon_n)'$ are additive errors. For simplicity, write $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, and define $\boldsymbol{Z}$ as the matrix with Z_{iq} as its iq-th element, for $i = 1, \ldots, n$ and $q = 1, \ldots, Q$. It is further assumed that $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ follow the normal distribution and satisfy

$$\mathrm{E}\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \boldsymbol{0}, \quad \mathrm{Cov}\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \sigma^2 \begin{pmatrix} \boldsymbol{G}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}$$

and

$$\boldsymbol{V} = \mathrm{Cov}(\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}) = \sigma^2(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{I}),$$

where $\sigma^2 > 0$ and $\boldsymbol{G}(\boldsymbol{\theta})$ is a positive definite matrix with elements known upto the parameter vector $\boldsymbol{\theta}$; in the sequel we simply write $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\theta})$. For identifiability purposes, we assume $\mathrm{E}[f_p(X_{ip})] = 0$ for all i and p. We consider the case that the X_{ip}'s are random and sampled from some continuous density function.

The concept of fixed and random effects has been successfully applied to repeated measurement data to study the variations both within and between replicates of subjects; e.g., see Laird and Ware (1982); Zeger and Liang (1986); Liang and Zeger (1986); Diggle *et al.* (2002). These authors mainly consider the case when the fixed effects are linear, while our approach can be seen as an extension in which nonparametric fixed effects are allowed. The nonparametric additive mixed model for repeated measurements data can be written as:

$$Y_{ik} = \mu + \sum_{p=1}^{P} f_p(X_{ikp}) + \sum_{q=1}^{Q} u_{iq} Z_{ikq} + \epsilon_{ik}, \quad i = 1, \ldots, N, \quad k = 1, \ldots, n_i, \tag{2}$$

where Y_{ik} is the kth measurement of the ith subject, X_{ikp} and Z_{ikq} are covariates associated with Y_{ik}, N is the total number of subjects and n_i is the number of measurements taken from the ith subject. The random effects u_{iq}'s model the variation among different subjects while the fixed effects f_p's model the variation between the replicates. Equation (2) can be transformed to (1) by suppressing the replicate index k, and the total number of observations is $\sum_{i=1}^{N} n_i = n$. Similar formulations of modeling have been discussed for examples in Zhang

et al. (1998); Lin and Zhang (1999); Wand (2003); Fahrmeir and Lang (2001). The first three pieces of work consider L_2 minimization criteria while the last considers the problem from a Bayesian perspective.

The problem of fixed and random effects selection has been well studied for linear mixed models. Chen and Dunson (2003) develop a Bayesian approach for selecting random effects in linear mixed models, while Kinney and Dunson (2007) use Bayesian method for selecting both fixed and random effects in linear and logistic mixed models. Bondell *et al.* (2010) also study the problem of fixed and random effects selection in linear mixed models, but with a penalized likelihood approach. Lastly, Pu and Niu (2006) extended the generalized information criterion for choosing fixed and random effects in linear mixed models.

Given (1) (or (2)), the goal of this article is to identify which of the functions f_p's and which of the random effects u_q's are statistically significant. Our methodology begins with modeling the f_p's nonparametrically by B-spline basis expansions. As to be demonstrated in Section 2 below, the use of B-spline bases can transform (1) into a compact matrix representation. For parameter estimation, we develop an iterative algorithm that combines the adaptive group lasso and Newton-Raphson methods. This algorithm is presented in Section 3. Section 4 derives a Bayesian information criterion to perform simultaneous variable selection for both fixed and random effects. Theoretical properties of our methodology are studied in Section 5. We illustrate the practical performance of our approach via both simulation experiments and practical data analysis; see Sections 6 and 7. Technical details are deferred to the appendix.

2. B-spline modeling of nonparametric fixed effects

Loosely, a spline is a piecewise polynomial that is smoothly connected at its knots. Denote $\mathcal{S}(d_0, \boldsymbol{t})$ as the collection of all spline functions of order d_0 with knots $\boldsymbol{t} = (t_0, \ldots, t_{\tau+1})'$, where $t_0 \leq t_1 \leq \cdots \leq t_{\tau+1}$. There exists $m = d_0 + \tau$ B-spline basis functions $\{\phi_j(\cdot)\}_{j=1}^m$ for $\mathcal{S}(d_0, \boldsymbol{t})$. That is, all elements of $\mathcal{S}(d_0, \boldsymbol{t})$ can be expressed as linear combinations of $\{\phi_j(\cdot)\}_{j=1}^m$. These B-spline basis functions are normalized in the sense that $\sum_{j=1}^m \phi_j(x) = 1$ for all x. They play an important role in non-parametric additive modeling; e.g., see Stone (1985, 1986). For exact expressions of ϕ_j's, see de Boor (2001).

Denote $\boldsymbol{t}_p$ as the knot vector for f_p, and let $\{\phi_{pj}(\cdot)\}_{j=1}^{m_p}$ be the corresponding basis functions of $\mathcal{S}(d_0, \boldsymbol{t}_p)$. We model f_p by $\tilde{f}_p$:

$$\tilde{f}_p = \sum_{k=1}^{m_p} \beta_{pk} \phi_{pk}.$$

To make the model identifiable, we assume $\sum_{i=1}^n \tilde{f}_p(X_{ip}) = 0$, which is a sample analogue of $\mathrm{E}[f_p(X_{ip})] = 0$. This can be achieved by centering the basis functions, or equivalently, by assuming that

$$\sum_{k=1}^{m_p} \sum_{i=1}^{n} \phi_{pk}(X_{ip}) \beta_{pk} = 0. \tag{3}$$

Hence, the number of effective basis functions is m_p. For simplicity, we set $m_p = m$ for all p; the effect of this is minimal as long as m is large enough.

For $p = 1, \ldots, P$, let $\boldsymbol{\beta}_p = (\beta_{p1}, \ldots, \beta_{pm})'$ and $\boldsymbol{\beta} = (\mu, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_P')'$. Also, write $\boldsymbol{X}_p$ as the matrix with its ij-th element as $\phi_{pj}(X_{ip})$ for $j = 1, \ldots, m$ and $i = 1, \ldots, n$, and define $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_P]$. The matrix representation of our spline model for (1) is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} = \mu\mathbf{1} + \sum_{p=1}^{P} \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{4}$$

which is a common formulation of mixed models. However, in (4) those columns of $\boldsymbol{X}$ that correspond to the basis functions of the same f_p are naturally grouped together, in the sense that this whole group of basis functions should be kept or killed together during the fixed effects selection process. This concept of group variables has been considered in Lin and Zhang (2006); Yuan and Lin (2006); Wang *et al.* (2007b). We note that (4) can also be expressed as $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$, with the log-likelihood function given as

$$l(\boldsymbol{\beta}, \boldsymbol{V}) = -\frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

3. Parameter estimation

There are two types of parameters in (4): the fixed component parameters $\boldsymbol{\beta} = (\mu, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_P')$ and the random component parameters $\{\boldsymbol{G}, \sigma^2\}$. This section develops a method for estimating these parameters when the complexity of the model is pre-specified. Automatic choice of model complexity will be discussed in the next section. The main idea of our estimation method is to iterate the following two steps until convergence:

1. given a current set of estimates for $\{\boldsymbol{G}, \sigma^2\}$, estimate $\boldsymbol{\beta}$ via the adaptive group lasso methodology,
2. given a current estimate for $\boldsymbol{\beta}$, obtain estimates for $\{\boldsymbol{G}, \sigma^2\}$ using the Newton-Raphson method.

3.1. Fixed component estimation using adaptive group lasso

The lasso was introduced by Tibshirani (1996) as a methodology for simultaneous variable selection and shrinkage estimation. Later Zou (2006) developed a variant termed the adaptive lasso which was shown to possess superior theoretical properties. The use of lasso for grouped variables, the so-called group lasso, has been discussed for examples in Yuan and Lin (2006); Meier *et al.* (2008); Wei and Huang (2008); Ravikumar *et al.* (2009). In particular, Huang *et al.* (2010); Meier *et al.* (2009) consider the estimation of additive models using group lasso. Here we develop an adaptive group lasso procedure for estimating the fixed component parameters.

Suppose for now estimates for $\{\boldsymbol{G}, \sigma^2\}$ and hence $\boldsymbol{V}$ are available. Denote the estimate as $\hat{\boldsymbol{V}}$, and write the likelihood as $l(\boldsymbol{\beta}) = l(\boldsymbol{\beta}, \hat{\boldsymbol{V}})$. Firstly the estimate for μ is $\hat{\mu} = \bar{Y}$, the average of all Y_i's. Then our adaptive group lasso estimates for the remaining parameters in $\boldsymbol{\beta} = (\mu, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_P')$ are defined as the minimizer of

$$S(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{p=1}^{P} w_p \|\boldsymbol{\beta}_p\| \tag{5}$$

with the weights w_p given by

$$w_p = \begin{cases} (\|\tilde{\boldsymbol{\beta}}_p\|^{-1} - \nu)_+, & \text{if } \|\tilde{\boldsymbol{\beta}}_p\| > 0, \\ \infty, & \text{otherwise,} \end{cases} \qquad p = 1, \ldots, P, \tag{6}$$

where $\|\cdot\|$ is the usual Euclidean norm, $\lambda > 0$ and $\nu > 0$ are tuning parameters, and $\tilde{\boldsymbol{\beta}}_p$ is an initial estimate for $\boldsymbol{\beta}_p$. Both λ and ν are pre-specified before the minimization of (5) is carried out. When λ is zero, the corresponding estimates will be reduced to the generalized least square estimate. As λ becomes larger, more $\boldsymbol{\beta}_p$'s are shrunken to zero. The presence of ν is to ensure the consistency properties of our estimation method (see Section 5). Methods for choosing λ and ν will be provided in Section 4 below.

Due to the non-differentiable property of the penalty term, minimization of (5) is difficult. Different algorithms have been developed for minimizing such group lasso type likelihoods; e.g., Meier *et al.* (2008); Yuan and Lin (2006). For (5), as the B-spline basis functions are not orthogonal and the variance component $\boldsymbol{V}$ is not diagonal, the shooting algorithm of Yuan and Lin (2006) is not applicable, but the blockwise co-ordinate gradient descent (BCGD) algorithm of Meier *et al.* (2008) can be adopted. Below is a brief description of the BCGD algorithm customized to our settings.

Note that the log-likelihood $l(\boldsymbol{\beta})$ is not a quadratic function and the key idea of BCGD is to combine a quadratic approximation of $l(\boldsymbol{\beta})$ with a line search. First, for any $\boldsymbol{d} \in \mathbb{R}^{mP+1}$, define the following approximation $M(\boldsymbol{d})$ to $S(\boldsymbol{\beta}+\boldsymbol{d})$:

$$M(\boldsymbol{d}) = -\left\{ l(\boldsymbol{\beta}) + \boldsymbol{d}' \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \frac{1}{2} \boldsymbol{d}' \boldsymbol{H} \boldsymbol{d} \right\} + \lambda \sum_{p=1}^{P} w_p \|\boldsymbol{\beta}_p + \boldsymbol{d}_p\| \approx S(\boldsymbol{\beta} + \boldsymbol{d}),$$

where $\boldsymbol{H}$ is a diagonal matrix approximating the Hessian of $S(\boldsymbol{\beta} + \boldsymbol{d})$. Denote the diagonal elements of $\boldsymbol{H}$ as $(h_0, h_1 \boldsymbol{e}_m, h_2 \boldsymbol{e}_m, \ldots, h_p \boldsymbol{e}_m, \ldots, h_P \boldsymbol{e}_m)$, where $\boldsymbol{e}_m$ is a row vector of m ones. For $p > 0$, a possible choice of h_p is

$$h_p = \min \left[\operatorname{diag} \left\{ \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p \partial \boldsymbol{\beta}_p'} \right\} \right] = -\max[\operatorname{diag}(\boldsymbol{X}_p' \boldsymbol{V}^{-1} \boldsymbol{X}_p)].$$

For h_0, it vanishes after a differentiation operation and hence plays no role in the actual minimization algorithm. For simplicity, we set $h_0 = 0$.

Next we consider minimizing $M(\boldsymbol{d})$ one "block at a time". More precisely, for each $p > 0$, we minimize $M(\boldsymbol{d})$ with respect to (w.r.t.)

$$\boldsymbol{d} = (0, 0\boldsymbol{e}_m, \ldots, 0\boldsymbol{e}_m, \boldsymbol{d}_p, 0\boldsymbol{e}_m, \ldots, 0\boldsymbol{e}_m),$$

where $\boldsymbol{d}_p \in \mathbb{R}^m$. Direct algebra shows that

$$\boldsymbol{d}_p = \begin{cases} -\boldsymbol{\beta}_p, & \text{if } \|\boldsymbol{S}_p\| < \lambda w_p \\ -\frac{1}{h_p}\{\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} - \lambda w_p \frac{\boldsymbol{S}_p}{\|\boldsymbol{S}_p\|}\}, & \text{otherwise} \end{cases}, \tag{7}$$

where $\boldsymbol{S}_p = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_p} - h_p \boldsymbol{\beta}_p$. Therefore, given $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{d}_p^{(t)}$ is the best direction of search w.r.t. the pth component for minimizing $S(\boldsymbol{\beta}^{(t)})$.

Lastly, we update the minimizer by $\boldsymbol{\beta}^{(t+1)} + a^{(t)}\boldsymbol{d}_p^{(t)}$, where $a^{(t)}$ is the step length of the search. This step length can be chosen by the Armijo rule: $a^{(t)}$ is chosen as the largest value that satisfies

$$S(\boldsymbol{\beta}^{(t+1)}) \leq S(\boldsymbol{\beta}^{(t)}) + c_1 a_p^{(t)} \Delta^{(t)}, \tag{8}$$

where $\Delta^{(t)} = -\boldsymbol{d}_p^{(t)\prime} \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \lambda\{\|\boldsymbol{\beta}_p^{(t)} + \boldsymbol{d}_p^{(t)}\| - \|\boldsymbol{\beta}^{(t)}\|\}$ and $c_1 > 0$. The value $\Delta^{(t)}$ is a linear approximation of the improvement in $S(\cdot)$ and we choose $c_1 = 0.1$ as suggested by Meier *et al.* (2008). Selecting the step length $a^{(t)}$ in this fashion ensures a sufficient decrease in the penalized likelihood.

In summary, given a current estimate $\hat{\boldsymbol{\beta}}^{(t)}$ of $\boldsymbol{\beta}$, the BCGD algorithm computes the next iterative estimate $\hat{\boldsymbol{\beta}}^{(t+1)}$ with the following steps: for $p = 1, \ldots, P$,

1. evaluates $\boldsymbol{d}_p^{(t)}$ with (7),
2. finds the largest $a^{(t)}$ that satisfies (8), and
3. set $\hat{\boldsymbol{\beta}}_p^{(t+1)} = \hat{\boldsymbol{\beta}}_p^{(t)} + a_p^{(t)}\boldsymbol{d}_p^{(t)}$.

This BCGD algorithm converged rapidly in all our numerical work.

3.2. Random component estimation using Newton Raphson

Now we describe the second iterative step of our estimation procedure: assume an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is available and estimate the variance components $\{\boldsymbol{G}, \sigma^2\}$. Setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the log-likelihood becomes

$$l(\hat{\boldsymbol{\beta}}, \boldsymbol{V}) = -\frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \boldsymbol{V}^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right). \tag{9}$$

The maximum likelihood estimates for $\{\boldsymbol{G}, \sigma^2\}$ can be obtained by maximizing (9). A major criticism of these maximum likelihood estimates is their bias caused by the loss in the degrees of freedom from the estimation of $\boldsymbol{\beta}$. To correct this, Harville (1974) introduced a restricted maximum likelihood approach in which a correction term $-\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|$ is added to (9). However, empirical results seem to suggest that there is no major difference between the two

likelihood approaches. In our work, we use maximum likelihood and the maximization of (9) is achieved via the Newton-Raphson (NR) algorithm (Lindstrom and Bates, 1988).

To further simplify our maximization problem, we replace σ^2 in (9) with its maximum likelihood estimate:

$$\hat{\sigma}^2 = \frac{1}{n}\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r} \quad \text{with} \quad \boldsymbol{V}^* = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{I} \quad \text{and} \quad \boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

Then our estimate for $\boldsymbol{G}$ is given as the minimizer of the following negative profile likelihood function

$$p(\boldsymbol{G}|\hat{\boldsymbol{\beta}}) = \log(|\boldsymbol{V}^*|) + n\log(\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r}). \tag{10}$$

Let $\boldsymbol{A} = \boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{r}(\boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{r})'$, $\boldsymbol{B} = \boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{Z}$. Then, as shown in Appendix A, the gradient and the Hessian matrix for the NR algorithm w.r.t. vec($\boldsymbol{G}$) are, respectively,

$$\frac{\partial p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G})} = \mathrm{vec}\boldsymbol{B} - \frac{1}{\hat{\sigma}^2}\mathrm{vec}\boldsymbol{A}$$

and

$$\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G})\partial \mathrm{vec}(\boldsymbol{G})'} = \frac{\partial \mathrm{vec}(\boldsymbol{G}')}{\partial \mathrm{vec}(\boldsymbol{G})}\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G}')\partial \mathrm{vec}(\boldsymbol{G})'},$$

where

$$\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G}')\partial \mathrm{vec}(\boldsymbol{G})'} = -\boldsymbol{B}\otimes\boldsymbol{B} - \frac{1}{n\hat{\sigma}^4}\mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{A})' + \frac{1}{\hat{\sigma}^2}\Big(\boldsymbol{A}\otimes\boldsymbol{B} + \boldsymbol{B}\otimes\boldsymbol{A}\Big).$$

In particular, denote the diagonal elements of $\boldsymbol{G}$ as $\boldsymbol{\theta}$, we have

$$\frac{\partial p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\theta}} = \frac{\partial \mathrm{vec}(\boldsymbol{G})}{\partial \boldsymbol{\theta}}\frac{\partial p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G})} = \mathrm{diag}(\boldsymbol{B}) - \frac{1}{\hat{\sigma}^2}\mathrm{diag}(\boldsymbol{A}) \tag{11}$$

and

$$\begin{aligned}\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\theta}'\partial \boldsymbol{\theta}} &= \frac{\partial \mathrm{vec}(\boldsymbol{G})}{\partial \boldsymbol{\theta}}\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G})\partial \mathrm{vec}(\boldsymbol{G})'}\Big(\frac{\partial \mathrm{vec}(\boldsymbol{G})}{\partial \boldsymbol{\theta}}\Big)' \\ &= -\boldsymbol{B}\odot\boldsymbol{B} - \frac{1}{n\hat{\sigma}^4}(\boldsymbol{A}\odot\boldsymbol{A}) + \frac{2}{\hat{\sigma}^2}(\boldsymbol{A}\odot\boldsymbol{B}),\end{aligned} \tag{12}$$

where $\odot$ represents the element-wise multiplication. With these expressions, the standard NR algorithm can be applied to maximize (10) and obtain estimates for $\boldsymbol{V}$ and $\boldsymbol{G}$. We note Woodbury's identity $\boldsymbol{V}^{*-1} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'$ can be applied to avoid heavy computations.

3.3. Combining the two algorithms

The above BCGD and NR algorithms can be applied iteratively to obtain estimates for both the fixed and random components parameters $\boldsymbol{\beta}$, $\boldsymbol{V}$ and $\boldsymbol{G}$. The combined algorithm begins with initial estimates $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{V}}^{(0)}$ for $\boldsymbol{\beta}$ and $\boldsymbol{V}$, respectively, and iterate, for $t = 0, 1, \ldots$, the following two steps until convergence:

1. *Fixed Effects Estimation*: Obtain the next iterative estimate $\hat{\boldsymbol{\beta}}^{(t+1)}$ with the BCGD Algorithm described in Section 3.1.
2. *Random Effects Estimation*: Obtain $\hat{\boldsymbol{V}}^{(t+1)}$ and $\hat{\boldsymbol{G}}^{(t+1)}$ by minimizing (10) with the NR algorithm described in Section 3.2.

Convergence of this combined algorithm can be determined by monitoring the successive changes of the iterative fitted value of $\boldsymbol{Y}$. In our implementation we set the initial estimates as $\hat{\boldsymbol{\beta}}^{(0)} = 0$ and $\hat{\boldsymbol{V}}^{(0)} = \boldsymbol{I}$.

Denote the final estimates as $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{G}}$. When the above iteration finishes, the estimated conditional expectation of $\boldsymbol{u}$ given $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ can be calculated as

$$\hat{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z} + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{13}$$

while the estimate for $\mathrm{E}[\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{u}]$ is given by $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}$. Also, $\hat{\boldsymbol{Y}}$ can be used as the prediction of $\boldsymbol{Y}$.

4. Selection of model complexity

This section develops a method for selecting the model complexity of (1). That is, to select the adaptive group lasso parameters (λ, ν) and to determine which random effects u_q's should enter the final model. Note that the selection of the fixed effects f_p's (which is equivalent to $\boldsymbol{\beta}_p$'s) is achieved by varying the values of (λ, ν).

4.1. Model selection criterion

We adopt the Bayesian information criterion (BIC) to select the model complexity. To be specific, we define the best fitting model as the one that minimizes the following objection function:

$$n\log(\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{u}}\|^2 + \mathrm{df}_{\hat{\boldsymbol{Y}}} \cdot \log(n), \tag{14}$$

where $\mathrm{df}_{\hat{\boldsymbol{Y}}}$ is the effective degrees of freedom of the fitted model. Computable expression for $\mathrm{df}_{\hat{\boldsymbol{Y}}}$ is given in the next subsection. Notice that the first two terms in (14) together form the conditional likelihood of $\boldsymbol{Y}$ given $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ (e.g., Ruppert *et al.*, 2003).

4.2. Calculating the degrees of freedom

This subsection derives a computable expression for $\mathrm{df}_{\hat{\boldsymbol{Y}}}$ so that (14) can be used in practice.

For simplicity, suppose for the moment that $\boldsymbol{V}$ is known or can be independently estimated. The degrees of freedom of any fitted model $\mathrm{df}_{\hat{\boldsymbol{Y}}}$ is calculated as the sum of the degrees of freedom contributed by the fixed and random effects parameters. If we denote the degrees of freedom of the fixed and random effects as, respectively, $\mathrm{df}_{\hat{\boldsymbol{\beta}}}$ and $\mathrm{df}_{\hat{\boldsymbol{u}}}$, we have $\mathrm{df}_{\hat{\boldsymbol{Y}}} = \mathrm{df}_{\hat{\boldsymbol{\beta}}} + \mathrm{df}_{\hat{\boldsymbol{u}}}$.

Following Shen and Ye (2002) and Ruppert *et al.* (2003), we define $\mathrm{df}_{\hat{\boldsymbol{\beta}}}$ and $\mathrm{df}_{\hat{\boldsymbol{u}}}$ as, respectively,

$$\mathrm{df}_{\hat{\boldsymbol{\beta}}} = \mathrm{E}\Big\{\mathrm{tr}\Big(\frac{\partial \boldsymbol{X}\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{Y}}\Big)\Big\} = 1 + \mathrm{E}\Big\{\mathrm{tr}\Big(\sum_{p=1}^{P} \frac{\partial \hat{\boldsymbol{\beta}}_p}{\partial \boldsymbol{Y}} \boldsymbol{X}_p'\Big)\Big\}, \tag{15}$$

and

$$\mathrm{df}_{\hat{\boldsymbol{u}}} = \mathrm{E}\Big\{\mathrm{tr}\Big(\frac{\partial \boldsymbol{Z}\hat{\boldsymbol{u}}}{\partial \boldsymbol{Y}}\Big)\Big\} = \mathrm{E}\Big[\mathrm{tr}\Big\{\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} - \sum_{p=1}^{P} \frac{\partial \hat{\boldsymbol{\beta}}_p}{\partial \boldsymbol{Y}} \boldsymbol{X}_p'\Big)\Big\}\Big], \tag{16}$$

where $\boldsymbol{J}$ is a matrix of ones. In these expressions the only unknown is $\frac{\partial \hat{\boldsymbol{\beta}}_p}{\partial \boldsymbol{Y}}$, and it can be calculated as follows.

By the KKT conditions of (5), we have, for $p = 1, \ldots, P$,

$$-\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \lambda w_p \frac{\hat{\boldsymbol{\beta}}_p}{\|\hat{\boldsymbol{\beta}}_p\|} = \boldsymbol{0} \quad \text{if} \quad \hat{\boldsymbol{\beta}}_p \neq \boldsymbol{0},$$

which gives

$$\frac{\partial \hat{\boldsymbol{\beta}}_p}{\partial \boldsymbol{Y}}\Big\{\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{X}_p + \frac{\lambda w_p}{\|\hat{\boldsymbol{\beta}}_p\|}\Big(\boldsymbol{I} - \frac{\hat{\boldsymbol{\beta}}_p\hat{\boldsymbol{\beta}}_p'}{\hat{\boldsymbol{\beta}}_p'\hat{\boldsymbol{\beta}}_p}\Big)\Big\} + \sum_{j \neq p} \frac{\partial \hat{\boldsymbol{\beta}}_j}{\partial \boldsymbol{Y}} \boldsymbol{X}_j'\boldsymbol{V}^{-1}\boldsymbol{X}_p = \Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\Big)\boldsymbol{V}^{-1}\boldsymbol{X}_p.$$

The above can be expressed in a more compact manner as follows:

$$\Big(\frac{\partial \hat{\boldsymbol{\beta}}_1}{\partial \boldsymbol{Y}}, \ldots, \frac{\partial \hat{\boldsymbol{\beta}}_P}{\partial \boldsymbol{Y}}\Big)\boldsymbol{C} = \Big(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_P\Big) = \boldsymbol{D},$$

where $\boldsymbol{C} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{E}$, $\boldsymbol{E} = \mathrm{diag}(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_p)$,

$$\boldsymbol{D}_p = \begin{cases} (\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{V}^{-1}\boldsymbol{X}_p, & \text{if } \hat{\boldsymbol{\beta}}_p \neq \boldsymbol{0} \\ \boldsymbol{0}, & \text{if } \hat{\boldsymbol{\beta}}_p = \boldsymbol{0} \end{cases}$$

and

$$\boldsymbol{E}_p = \begin{cases} \frac{\lambda w_p}{\|\hat{\boldsymbol{\beta}}_p\|}\left(\boldsymbol{I} - \frac{\hat{\boldsymbol{\beta}}_p\hat{\boldsymbol{\beta}}_p'}{\hat{\boldsymbol{\beta}}_p'\hat{\boldsymbol{\beta}}_p}\right), & \text{if } \hat{\boldsymbol{\beta}}_p \neq \boldsymbol{0} \\ 0, & \text{if } \hat{\boldsymbol{\beta}}_p = \boldsymbol{0} \end{cases}.$$

Therefore we have

$$\Big[\frac{\partial\hat{\boldsymbol{\beta}}_1}{\partial \boldsymbol{Y}},\ldots,\frac{\partial\hat{\boldsymbol{\beta}}_P}{\partial \boldsymbol{Y}}\Big] = \boldsymbol{D}\boldsymbol{C}^{-1}.$$

We note that $\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}$ is typically not of full rank, hence $\boldsymbol{C}$ may not be invertible in some cases (e.g., when $\lambda = 0$). However, $\boldsymbol{X}\boldsymbol{C}^{-1}\boldsymbol{D}'$ is invariant w.r.t. the choice of the pseudo inverse of $\boldsymbol{C}$. Hence, we can estimate (15) and (16) by

$$\hat{\mathrm{df}}_{\hat{\boldsymbol{\beta}}} = 1 + \mathrm{tr}\Big(\sum_{p=1}^{P}\frac{\partial\hat{\boldsymbol{\beta}}_p}{\partial \boldsymbol{Y}}\boldsymbol{X}_p'\Big), \tag{17}$$

and

$$\hat{\mathrm{df}}_{\hat{\boldsymbol{u}}} = \Big\{\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \hat{\boldsymbol{G}}^{-1})^{-1}\boldsymbol{Z}'\Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} - \sum_{p=1}^{P}\frac{\partial\hat{\boldsymbol{\beta}}_p}{\partial \boldsymbol{Y}}\boldsymbol{X}_p'\Big)\Big\}, \tag{18}$$

respectively. And the estimated degrees of freedom for the model equals $\hat{\mathrm{df}}_{\hat{\boldsymbol{Y}}} = \hat{\mathrm{df}}_{\hat{\boldsymbol{\beta}}} + \hat{\mathrm{df}}_{\hat{\boldsymbol{u}}}$. In practice we replace $\boldsymbol{V}$ by the $\hat{\boldsymbol{V}}$ obtained using the algorithm summarized in Section 3.3.

4.3. Practical minimization of (14)

In practice a global minimization of the selection criterion (14) is difficult. Here we suggest using the following procedure to approximate the minimizer of (14). This procedure is relatively fast, and produces excellent numerical results in all our experimental work. The steps are:

1. *Initialization:* With the full model (i.e., with all fixed and random effects), apply the ordinary group lasso to obtain initial parameter estimates. That is, the initial estimates are chosen as the minimizers of (5) with $w_p = 1$ for all p. The tuning parameter of this ordinary group lasso can be chosen for example by BIC.
2. *Fixed Effects Selection:* Using the estimates obtained in Step 1 as initial estimates and weights (6), apply the proposed parameter estimation method summarized in Section 3.3 to the full model. The adaptive group lasso parameters λ and ν are chosen by the BIC criterion (14) through a two-dimensional grid search.
3. *Random Effects Selection:* Perform the following backward selection strategy to the fitted model obtained in Step 2. First from this model remove the random effect that has the smallest coefficient magnitude and re-calculate its BIC (14) value, denoted as BIC_1. If BIC_1 is larger than the BIC value of the model from Step 2, then the model from Step 2 is taken as the final model. Otherwise, further remove the random effect that has the second smallest magnitude and re-calculate its corresponding BIC value; denoted it as BIC_2. If $\mathrm{BIC}_2 > \mathrm{BIC}_1$, then the earlier model with one random effect removed is taken as the final model. Otherwise, repeat

this process by removing more random effects until there is no decrease in the BIC values.

One could iterate Steps 2 and 3 above, but our experience suggests that the improvement, if any, is minimal.

5. Theoretical results

This section establishes the consistency properties of the proposed adaptive group lasso estimator, as well as the BIC criterion (14).

5.1. Consistency property of adaptive group lasso

Denote $A_0 \equiv \{p : \|f_p\| = 0, p = 1, \ldots, P\}$ and $A_T \equiv \{p : \|f_p\| > 0, p = 1, \ldots, P\}$. That is, A_0 is the set of all non-significant f_p's, while A_T is the set of all significant f_p's. Let $\boldsymbol{M}$ be any matrix in $\mathbb{R}^{m\times n}$. Define the matrix norm of $\boldsymbol{M}$ as $\|\boldsymbol{M}\| = (\max_{\boldsymbol{x}\in\mathbb{R}^n, \|\boldsymbol{x}\|=1} \boldsymbol{x}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{x})^{1/2}$. An important property of this matrix norm is that, for any $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{M}\boldsymbol{x}\| \le \|\boldsymbol{M}\|\|\boldsymbol{x}\|$. We consider the following regularity conditions (A.1-7):

(A.1) The number of elements in A_T, denoted as $|A_T|$, does not depend on n and P.

(A.2) For the knots $(t_i, i = 0, \ldots, \tau + 1)$, let $h_i = t_{i+1} - t_i$ and $h = \max_i h_i$. We assume that $h/\min_i h_i = O(1)$ and $\max_i |h_{i+1} - h_i| = o(m^{-1})$.

(A.3) Without loss of generality, we assume that the design points $X_{ip} \in [0, 1]$. Moreover, we assume that X_{ip} has a nonzero continuous density function.

(A.4) Let $\mathcal{F}$ be the collection of functions satisfying Lipschitz condition:

$$|f^{(k)}(s) - f^{(k)}(t)| \le C|s - t|^{\alpha}, \quad s, t \in [0, 1], \text{ for some } C > 0$$

where k is non-negative integer, $0 < \alpha \le 1$ and $0.5 < k + \alpha = d$ is called the order of smoothness. For $p = 1, \ldots, P$, $f_p \in \mathcal{F}$ and $\mathrm{E}f_p(X_{ip}) = 0$.

(A.5) The initial estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ satisfies the following conditions as $n \to \infty$:

(a) $r \max_{p\in A_0} \|\tilde{\boldsymbol{\beta}}_p\| = O_p(1)$ with $r \to \infty$.

(b) $P(\min_{p\in A_T} \|\tilde{\boldsymbol{\beta}}_p\| > \nu^{-1}) \to 1$ for any $\nu > 0$.

(A.6) As $n \to \infty$:

(a) $m \to \infty$, $\dfrac{m}{n^{1/2}} \to 0$, $\dfrac{m \log(2mP)}{n} \to 0$.

(b) $\dfrac{n}{\lambda r m^{d+1/2}} \to 0$, $\dfrac{\sqrt{n \log(2mP)}}{\lambda r} \to 0$.

(A.7) The norm of the covariance matrix $\|\boldsymbol{V}\|$ is finite.

We need more notation to proceed. For any $A \subset \{1, \ldots, P\}$, denote $\boldsymbol{X}_A$ and $\boldsymbol{\beta}_A$, respectively, as the design matrix and the subset of $\boldsymbol{\beta}$ corresponding to the

model with fixed effects f_p's indexed by A. Then the generalized least squares estimate of $\boldsymbol{\beta}_A$ based on $\boldsymbol{X}_A$ is $\hat{\boldsymbol{b}}_A = (\hat{\boldsymbol{b}}'_{A,p}, p = 1, \ldots, P)'$, where $(\hat{\boldsymbol{b}}'_{A,p}, p \in A)' = (\boldsymbol{X}'_A \boldsymbol{V}^{-1} \boldsymbol{X}_A)^{-1} \boldsymbol{X}'_A \boldsymbol{V}^{-1} \boldsymbol{Y}$ and $(\hat{\boldsymbol{b}}'_{A,p}, p \notin A)' = \boldsymbol{0}$. Without loss of generality, we assume $\mu = 0$.

The following theorem shows that our adaptive group lasso estimate of $\boldsymbol{\beta}$ is not only capable of identifying the correct variables, but is also capable of recovering the predictive performance of the generalized least squares estimate based on the correct variables. The proof can be found in Appendix C.

Theorem 1 ($\boldsymbol{G}$ known)**.** *Under (A.1)-(A.7), as* $n \to \infty$,

1. $P\big(sign(\|\hat{\boldsymbol{\beta}}_p\|) = sign(\|\boldsymbol{\beta}_p\|), p = 1, \ldots, P\big) \to 1$.
2. $P(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{b}}_{A_T}) \to 1$.
3. $\displaystyle\sum_{p=1}^{P} \|\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p\|^2 = O_p\Big(\frac{m^2}{n}\Big) + O_p\Big(\frac{1}{m^{2d-1}}\Big) + O_p\Big(\frac{m^2\lambda^2}{n^2}\Big) o_p(1)$.
4. $\displaystyle\sum_{p=1}^{P} \|\hat{f}_p - f_p\|^2 = O_p\Big(\frac{m}{n}\Big) + O_p\Big(\frac{1}{m^{2d}}\Big) + O_p\Big(\frac{m\lambda^2}{n^2}\Big) o_p(1)$.

Corollary 1 ($\boldsymbol{G}$ unknown)**.** *Under the conditions of Theorem 1, except that* $\boldsymbol{G}$ *is replaced by a consistent estimate* $\hat{\boldsymbol{G}}$*, Theorem 1 continue to hold.*

Proof. The proof is essentially the same as that for Theorem 1 after applying Slutsky's theorem: $h(\hat{\boldsymbol{G}}) \to h(\boldsymbol{G})$ in probability for any real-value continuous function h. □

5.2. Consistency property of Bayesian information criterion

First assume $\boldsymbol{G}$ is known. Then the maximum likelihood estimate of σ^2 based on model A is $\hat{\sigma}^2_A \equiv \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}}_A)' \boldsymbol{V}^{*-1} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}}_A)$. To stress the dependence of the parameter estimates on (λ, ν), denote $\hat{\boldsymbol{\beta}}_{\lambda,\nu}$ and $\hat{\boldsymbol{u}}_{\lambda,\nu}$ as the adaptive grouped lasso estimates with weights (6) and tuning parameters (λ, ν). Also let $\hat{\sigma}^2_{\lambda,\nu} \equiv \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})' \boldsymbol{V}^{*-1} (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})$ be the corresponding estimate of σ^2. With this, the Bayesian information criterion (14) can be re-written as:

$$\mathrm{BIC}(\lambda, \nu) = \log(\hat{\sigma}^2_{\lambda,\nu}) + \frac{1}{n\hat{\sigma}^2_{\lambda,\nu}} \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu} - \boldsymbol{Z}\hat{\boldsymbol{u}}_{\lambda,\nu}\|^2 + \hat{\mathrm{df}}_{\hat{\boldsymbol{Y}}}(\lambda, \nu) \frac{\log(n)}{n}, \quad (19)$$

where $\hat{A}_{\lambda,\nu} \equiv \{p : \|\hat{\boldsymbol{\beta}}_{\lambda,\nu,p}\| > 0\}$. We consider the following technical conditions:

(B.1) For any $A \in \mathcal{A}$, there exists $\sigma^2_A > 0$ such that $\hat{\sigma}^2_A \to \sigma^2_A$ in probability.
(B.2) For any A with $A \cap A_T \neq A_T$, $\sigma^2_A > \sigma^2_{A_T}$.
(B.3) As $n \to \infty$, $\lambda/n \to 0$.

Appendix C establishes the following theorem.

Theorem 2 (***G*** known)**.** *Under (A.1)-(A.7) and (B.1)-(B.3). Assume that (A.5) and (A.6) hold for some sequences* $\{\lambda_n^*\}$ *and* $\{\nu_n^*\}$*. Let*

$$(\hat{\lambda}, \hat{\nu}) = \arg\min_{(\lambda,\nu)} BIC(\lambda, \nu).$$

Then

$$P(\hat{\boldsymbol{\beta}}_{\hat{\lambda},\hat{\nu}} = \hat{\boldsymbol{b}}_{A_T}) \to 1 \quad \text{as } n \to \infty.$$

As similar to Corollary 1, applying Slutsky's theorem leads to the following Corollary.

Corollary 2 (***G*** unknown)**.** *Under the conditions of Theorem 2, except that* $\boldsymbol{G}$ *is replaced by a consistent estimate* $\hat{\boldsymbol{G}}$*, Theorem 2 continues to hold.*

6. Simulations

In this section, we investigate the empirical performance of our proposed methodology via numerical experiments. The covariates X_{ip}'s were generated from the uniform$(0, 1)$ distribution. The test functions (i.e., fixed effects) used in the experiments are listed in Table 1; these functions have been used by Meier *et al.* (2009). Note that the functions listed in Table 1 are not centered, while in all our numerical work, they were, however, all centered.

We tested our method with three different types of models:

Model 1 (categorical variables as mixed-effects):

$$\begin{aligned} Y_i =& f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + \sum_{p=5}^{20} f_p(X_{ip}) \\ &+ \sum_{k=1}^{5} u_{1k} I(Z_{i1} = k) + \sum_{k=1}^{5} u_{2k} I(Z_{i2} = k) \\ &+ \sum_{k=1}^{3} u_{3k} I(Z_{i3} = k) + \sum_{k=1}^{3} u_{4k} I(Z_{i4} = k) + \epsilon_i, \end{aligned}$$

where u_{qk}'s are i.i.d. $N(0, \theta_q^2 \sigma^2)$ with $\theta_1 = 3$, $\theta_2 = 4$, and $\theta_3 = \theta_4 = 0$, and ϵ_i's are i.i.d. $N(0, \sigma^2)$. The Z_{iq}'s are generated from the discrete uniform[1,...,5] distribution for $q = 1, 2$ and discrete uniform[1,2,3] for $q = 3, 4$.

TABLE 1
Test functions used in the simulation, all were centered in the simulation

test function	
f_1	$6x$
f_2	$5(2x-1)^2$
f_3	$4\sin(2\pi x)/\{2 - \sin(2\pi x)\}$
f_4	$3\{0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^2(2\pi x) + 0.5\sin^3(2\pi x)\}$
f_p, $p \geq 5$	0

Model 2 (random intercept):

$$Y_{ik} = u_i + f_1(X_{ik1}) + f_2(X_{ik2}) + f_3(X_{ik3}) + f_4(X_{ik4}) + \sum_{p=5}^{20} f_p(X_{ikp}) + \epsilon_{ik},$$

where u_i's are i.i.d. $N(0, \theta^2\sigma^2)$ with $\theta = 3$ and ϵ_{ik}'s are i.i.d. $N(0, \sigma^2)$. The response Y_{ik} represents the kth measurement of the ith subject, where $1 \le i \le N$, $1 \le k \le n_i$ and $\sum_{i=1}^{N} n_i = n$. Here N is the total number of subjects and n_i is the number of measurements taken from the ith subject.

Model 3 (random intercepts and trends):

$$Y_{ik} = a_i + b_{i1}X_{ik1} + f_3(X_{ik1}) + b_{i2}X_{ik2} + f_4(X_{ik2}) + \sum_{p=3}^{6}\{b_{ip}X_{ikp} + f_{p+2}(X_{ikp})\} + \epsilon_{ik},$$

where $a_i \sim N(0, \theta_1^2\sigma^2)$, $b_{i1} \sim N(0, \theta_2^2\sigma^2)$, $b_{i2} \sim N(0, \theta_3^2\sigma^2)$ with $\theta_i = i + 1$ and $b_{ip} = 0$ for $p \ge 3$, and ϵ_{ik}'s are i.i.d. $N(0, \sigma^2)$. As above, Y_{ik} represents the kth measurement of the ith subject, where $1 \le i \le N$, $1 \le k \le n_i$ and $\sum_{i=1}^{N} n_i = n$.

We tested the proposed method with different combinations of n, N and signal-to-noise ratio (SNR, defined as SNR $= \mathrm{Var}\{\mathrm{E}(Y|\boldsymbol{X})\}^{\frac{1}{2}}/\sigma$). The number of Monte-Carlo runs tested for each experimental configuration was 200. We use the following mean squared error (MSE) to measure the estimation accuracy for the additive components: MSE $= \sum_{p=1}^{P} \|f_p - \hat{f}_p\|^2$.

For each simulated data set, various estimates were obtained by three different estimation methods:

- Method I: applying the proposed method to the full data set.
- Method II: applying the ordinary group lasso method to the full data set (i.e., with both fixed and random effects covariates). No random effects selection is performed as all the random effects are always included in the final model. The tuning parameter λ for the group lasso selection is chosen by BIC. This method is implemented in a similar fashion as the approach described in Section 3.3, except no random effects selection is done.
- Method III: this method is used here as an "oracle method" for benchmark comparison, as only the significant variables are considered (i.e., assuming the true model is known). It iterates between the following two steps: use generalized least squares to estimate the fixed effects parameters and use the NR algorithm to estimate the random effects parameters.

For each estimated model, we counted the number of non-zero $\hat{f}_p$, and the number of non-zero random effects, intercepts and/or trends. We also calculated the above-defined MSE, and recorded the parameter estimates for the random components. Lastly, we noted if the estimated model is exactly the same as the true model, or if it is a superset of the true model. These results are summarized in Tables 2 to 7.

TABLE 2
Simulation results for Model 1. The true numbers of non-zero f_p's and random effects are 4 and 2 respectively

SNR = 3		number of non-zero $\hat{f}_p$				MSE		number of random effects			number of times true model is	
configuration	method	4	5	6	7+	mean	S.E.	2	3	4	included	selected
$n = 128$ $m = 7$	I	193	5	2	0	0.3748	0.0067	161	36	3	200	155
	II	39	48	40	73	0.6884	0.0136	—	—	—	—	—
	III	200	0	0	0	0.3753	0.0069	—	—	—	—	—
$n = 256$ $m = 8$	I	200	0	0	0	0.2467	0.0037	167	30	3	200	167
	II	118	52	18	12	0.4284	0.0071	—	—	—	—	—
	III	200	0	0	0	0.2082	0.0033	—	—	—	—	—
$n = 512$ $m = 9$	I	200	0	0	0	0.1351	0.0019	186	12	2	200	186
	II	156	41	3	0	0.2152	0.0035	—	—	—	—	—
	III	200	0	0	0	0.1105	0.0014	—	—	—	—	—

SNR = 4		number of non-zero $\hat{f}_p$				MSE		number of random effects			number of times true model is	
configuration	method	4	5	6	7+	mean	S.E.	2	3	4	included	selected
$n = 128$ $m = 7$	I	200	0	0	0	0.2605	0.0038	161	34	5	200	161
	II	58	49	36	57	0.5021	0.009	—	—	—	—	—
	III	200	0	0	0	0.2461	0.004	—	—	—	—	—
$n = 256$ $m = 8$	I	200	0	0	0	0.1835	0.0025	170	29	1	200	170
	II	123	53	20	4	0.3073	0.0045	—	—	—	—	—
	III	200	0	0	0	0.1501	0.0019	—	—	—	—	—
$n = 512$ $m = 9$	I	200	0	0	0	0.0926	0.0011	182	16	2	200	182
	II	154	40	4	2	0.1434	0.0018	—	—	—	—	—
	III	200	0	0	0	0.0747	9e-04	—	—	—	—	—

TABLE 3
Estimates of random components for Model 1

SNR = 3		$\hat{\theta}_1$ ($\theta_1 = 3$)			$\hat{\theta}_2$ ($\theta_2 = 4$)			$\hat{\sigma}$ ($\sigma \approx 1$)		
configuration	method	median	mean	s.e.	median	mean	s.e.	median	mean	s.e.
$n = 128$ $m = 7$	I	2.874	2.888	0.071	3.621	3.767	0.092	1.029	1.028	0.006
	II	2.579	2.709	0.089	3.271	3.556	0.121	1.131	1.132	0.009
	III	2.936	2.983	0.075	3.75	3.89	0.095	0.994	0.996	0.006
$n = 256$ $m = 8$	I	2.596	2.623	0.063	3.429	3.523	0.087	1.093	1.093	0.005
	II	2.393	2.45	0.06	3.193	3.299	0.084	1.176	1.173	0.005
	III	2.684	2.723	0.066	3.558	3.655	0.09	1.053	1.055	0.005
$n = 512$ $m = 9$	I	2.597	2.645	0.069	3.361	3.414	0.09	1.101	1.1	0.003
	II	2.499	2.55	0.066	3.252	3.295	0.088	1.139	1.138	0.003
	III	2.656	2.707	0.07	3.446	3.493	0.092	1.075	1.075	0.003

SNR = 4		$\hat{\theta}_1$ ($\theta_1 = 3$)			$\hat{\theta}_2$ ($\theta_2 = 4$)			$\hat{\sigma}$ ($\sigma \approx 0.8$)		
configuration	method	median	mean	s.e.	median	mean	s.e.	median	mean	s.e.
$n = 128$ $m = 7$	I	2.717	2.735	0.07	3.477	3.727	0.122	0.811	0.818	0.005
	II	2.423	2.505	0.08	3.101	3.551	0.19	0.911	0.917	0.007
	III	2.82	2.835	0.072	3.606	3.764	0.093	0.778	0.789	0.005
$n = 256$ $m = 8$	I	2.601	2.631	0.071	3.481	3.61	0.082	0.839	0.84	0.004
	II	2.432	2.437	0.066	3.233	3.343	0.075	0.91	0.907	0.004
	III	2.689	2.735	0.074	3.627	3.753	0.085	0.808	0.809	0.003
$n = 512$ $m = 9$	I	2.537	2.614	0.064	3.257	3.335	0.078	0.837	0.835	0.003
	II	2.477	2.53	0.062	3.153	3.225	0.075	0.868	0.863	0.003
	III	2.606	2.677	0.066	3.319	3.415	0.08	0.817	0.816	0.002

TABLE 4
Simulation results for Model 2. The true number of f_p's is 4, and N_I denotes "number of times the random intercept is selected"

SNR = 3		number of non-zero $\hat{f}_p$				MSE			number of times true model is	
configuration	method	4	5	6	7+	mean	S.E.	N_I	included	selected
$n = 128$	I	195	5	0	0	0.3944	0.0068	200	200	195
$m = 7$	II	34	65	46	55	0.7134	0.0138	—	—	—
$N = 16$	III	200	0	0	0	0.4202	0.0077	—	—	—
$n = 256$	I	200	0	0	0	0.2519	0.0041	200	200	200
$m = 8$	II	109	66	18	7	0.4376	0.0074	—	—	—
$N = 16$	III	200	0	0	0	0.2166	0.0033	—	—	—
$n = 512$	I	200	0	0	0	0.1344	0.0018	200	200	200
$m = 9$	II	168	32	0	0	0.2203	0.0031	—	—	—
$N = 16$	III	200	0	0	0	0.112	0.0016	—	—	—
$n = 256$	I	200	0	0	0	0.2659	0.0036	200	200	200
$m = 8$	II	103	65	26	6	0.4744	0.0074	—	—	—
$N = 32$	III	200	0	0	0	0.2341	0.0035	—	—	—
$n = 512$	I	200	0	0	0	0.14	0.002	200	200	200
$m = 9$	II	157	37	5	1	0.226	0.0036	—	—	—
$N = 32$	III	200	0	0	0	0.1148	0.0015	—	—	—

SNR = 4		number of non-zero $\hat{f}_p$				MSE			number of times true model is	
configuration	method	4	5	6	7+	mean	S.E.	N_I	included	selected
$n = 128$	I	197	3	0	0	0.2763	0.0041	200	200	197
$m = 7$	II	40	51	42	67	0.5337	0.009	—	—	—
$N = 16$	III	200	0	0	0	0.265	0.0046	—	—	—
$n = 256$	I	200	0	0	0	0.1783	0.0023	200	200	200
$m = 8$	II	131	50	17	2	0.3039	0.0046	—	—	—
$N = 16$	III	200	0	0	0	0.1503	0.0018	—	—	—
$n = 512$	I	200	0	0	0	0.0959	0.0012	200	200	200
$m = 9$	II	160	36	3	1	0.1484	0.002	—	—	—
$N = 16$	III	200	0	0	0	0.0765	0.001	—	—	—
$n = 256$	I	200	0	0	0	0.1859	0.0026	200	200	200
$m = 8$	II	115	52	27	6	0.3191	0.0049	—	—	—
$N = 32$	III	200	0	0	0	0.1578	0.0021	—	—	—
$n = 512$	I	200	0	0	0	0.0945	0.0012	200	200	200
$m = 9$	II	171	27	2	0	0.1505	0.0021	—	—	—
$N = 32$	III	200	0	0	0	0.0764	0.001	—	—	—

The simulation results seem to suggest that our proposed method can often correctly select the correct significant fixed effects components f_p's. It is also capable of selecting the significant random effects. The empirical performances of the proposed method improves as the sample size and/or SNR increase. The simulation results also show that our method outperforms the ordinary group lasso in terms of fixed effects selection.

When estimating the variance components (e.g., θ_1 and θ_2 in Model 1), a bias pattern seems to exist: the bias increases as n increases. This is not too surprising, as it is known that the maximum likelihood estimates of the variance components are biased, as they do not account for the loss in degrees of freedom

TABLE 5
Estimates of random components for Model 2

SNR = 3		$\hat{\theta}$ ($\theta = 3$)			$\hat{\sigma}$ ($\sigma \approx 1$)		
configuration	method	median	mean	s.e.	median	mean	s.e.
$n = 128$	I	3.085	3.171	0.046	1.044	1.036	0.007
$m = 7$	II	2.763	2.84	0.044	1.163	1.16	0.008
$N = 16$	III	3.169	3.267	0.048	1.007	1.004	0.006
$n = 256$	I	2.843	2.828	0.042	1.093	1.099	0.005
$m = 8$	II	2.662	2.631	0.039	1.174	1.182	0.005
$N = 16$	III	2.943	2.934	0.044	1.054	1.059	0.005
$n = 512$	I	2.884	2.891	0.037	1.098	1.095	0.003
$m = 9$	II	2.786	2.782	0.036	1.14	1.138	0.003
$N = 16$	III	2.955	2.957	0.038	1.073	1.071	0.003
$n = 256$	I	2.948	2.98	0.031	1.102	1.098	0.005
$m = 8$	II	2.719	2.748	0.028	1.197	1.19	0.005
$N = 32$	III	3.076	3.091	0.032	1.063	1.059	0.005
$n = 512$	I	2.954	2.961	0.026	1.097	1.097	0.003
$m = 9$	II	2.846	2.855	0.026	1.138	1.139	0.004
$N = 32$	III	3.026	3.028	0.027	1.071	1.073	0.003

SNR = 4		$\hat{\theta}$ ($\theta = 3$)			$\hat{\sigma}$ ($\sigma \approx 0.8$)		
configuration	method	median	mean	s.e.	median	mean	s.e.
$n = 128$	I	2.948	3.478	0.179	0.805	0.802	0.005
$m = 7$	II	2.603	4.357	0.369	0.918	0.912	0.007
$N = 16$	III	3.035	3.074	0.044	0.778	0.777	0.005
$n = 256$	I	2.819	2.811	0.039	0.844	0.836	0.004
$m = 8$	II	2.576	2.585	0.036	0.913	0.91	0.004
$N = 16$	III	2.929	2.926	0.04	0.81	0.804	0.004
$n = 512$	I	2.797	2.797	0.039	0.833	0.831	0.003
$m = 9$	II	2.666	2.693	0.038	0.864	0.863	0.003
$N = 16$	III	2.861	2.864	0.04	0.813	0.811	0.003
$n = 256$	I	2.917	2.912	0.029	0.834	0.835	0.004
$m = 8$	II	2.701	2.675	0.027	0.909	0.91	0.004
$N = 32$	III	3.037	3.025	0.03	0.805	0.803	0.004
$n = 512$	I	2.86	2.911	0.027	0.828	0.83	0.003
$m = 9$	II	2.726	2.794	0.026	0.864	0.865	0.003
$N = 32$	III	2.926	2.98	0.027	0.81	0.811	0.003

when estimating the fixed effects (Harville, 1974). However, for Model 1, we believe that it is more important to look at the estimates $\hat{\theta}_1\hat{\sigma}$ and $\hat{\theta}_2\hat{\sigma}$ rather than $\hat{\theta}_1$ and $\hat{\theta}_2$. If we look at the estimates $\hat{\theta}_1\hat{\sigma}$ and $\hat{\theta}_2\hat{\sigma}$, this bias pattern disappears.

By comparing the MSEs obtained from the proposed method (I) and the oracle method (III), one can see that, in terms of prediction power, the proposed method is not too far behind from the oracle method. As expected, as sample size increases, the MSE values of the proposed method decrease, which supports the claim that the spline approximation for the fixed effects components f_p's is consistent. We also note that the MSE values decrease as SNR increases, which indicates that the noise level affects the convergence rate of the spline approximation. In general, the random effects selection based on BIC improves

TABLE 6
Simulation results for Model 3. The true numbers of non-zero f_p and random trends are both 2. The notation N_I denotes "number of times the random intercept is selected"

SNR = 3		number of non-zero $\hat{f}_p$			MSE			number of random trends			number of times true model is	
configuration	method	2	3	4+	mean	S.E.	N_I	≤ 1	2	≥ 3	included	selected
$n = 128$	I	200	0	0	0.2676	0.0079	147	66	125	9	128	100
$m = 7$	II	106	54	40	0.4079	0.0104	—	—	—	—	—	—
$N = 16$	III	200	0	0	0.2593	0.0086	—	—	—	—	—	—
$n = 256$	I	200	0	0	0.1825	0.0055	190	15	176	9	185	161
$m = 8$	II	150	47	3	0.3093	0.0086	—	—	—	—	—	—
$N = 16$	III	200	0	0	0.1773	0.0056	—	—	—	—	—	—
$n = 512$	I	200	0	0	0.2092	0.0068	199	1	191	8	199	179
$m = 9$	II	173	24	3	0.2989	0.0076	—	—	—	—	—	—
$N = 16$	III	200	0	0	0.1345	0.0061	—	—	—	—	—	—
$n = 256$	I	200	0	0	0.1732	0.0043	134	74	117	9	126	108
$m = 8$	II	155	38	7	0.251	0.0054	—	—	—	—	—	—
$N = 32$	III	200	0	0	0.153	0.0039	—	—	—	—	—	—
$n = 512$	I	200	0	0	0.1254	0.0042	196	4	186	10	196	180
$m = 9$	II	169	29	2	0.1865	0.0054	—	—	—	—	—	—
$N = 32$	III	200	0	0	0.0905	0.0032	—	—	—	—	—	—

SNR = 4		number of non-zero $\hat{f}_p$			MSE			number of random trends			number of times true model is	
configuration	method	2	3	4+	mean	S.E.	N_I	≤ 1	2	≥ 3	included	selected
$n = 128$	I	200	0	0	0.2049	0.0051	153	77	110	13	120	90
$m = 7$	II	120	51	29	0.3013	0.0067	—	—	—	—	—	—
$N = 16$	III	200	0	0	0.1871	0.0047	—	—	—	—	—	—
$n = 256$	I	200	0	0	0.144	0.0037	188	19	171	10	181	154
$m = 8$	II	151	44	5	0.2227	0.0054	—	—	—	—	—	—
$N = 16$	III	200	0	0	0.1274	0.0033	—	—	—	—	—	—
$n = 512$	I	200	0	0	0.1369	0.0045	200	0	194	6	200	180
$m = 9$	II	149	46	5	0.1917	0.0054	—	—	—	—	—	—
$N = 16$	III	200	0	0	0.088	0.0038	—	—	—	—	—	—
$n = 256$	I	200	0	0	0.1338	0.0026	135	77	116	7	123	108
$m = 8$	II	151	40	9	0.1838	0.0038	—	—	—	—	—	—
$N = 32$	III	200	0	0	0.117	0.0022	—	—	—	—	—	—
$n = 512$	I	200	0	0	0.0827	0.0021	198	4	183	13	196	175
$m = 9$	II	157	39	4	0.1183	0.003	—	—	—	—	—	—
$N = 32$	III	200	0	0	0.0646	0.0016	—	—	—	—	—	—

when sample size and/or SNR increase. Lastly, for the repeated measurement model, when the number of replicates per subject increases, our method results in a more accurate random effects selection. However, the MSE values for the fixed effects components do not decrease when the number of replicates per subject increases, unless the total sample size increases.

In conclusion, our proposed method using BIC model selection is very capable of recovering the true model, especially when the sample size and/or SNR are not too small.

TABLE 7
Estimates of random components for Model 3

SNR = 3		$\hat{\theta}_1$ ($\theta_1 = 2$)			$\hat{\theta}_2$ ($\theta_2 = 3$)			$\hat{\theta}_3$ ($\theta_3 = 4$)			$\hat{\sigma}$ ($\sigma \approx 1$)		
configu-ration	method	median	mean	s.e.	median	mean	s.e.	median	mean	s.e.	median	mean	s.e.
$n = 128$	I	1.979	2.032	0.028	2.857	2.903	0.046	3.412	3.541	0.054	0.849	0.851	0.007
$m = 7$	II	1.83	1.86	0.047	2.899	2.947	0.065	3.568	3.673	0.063	0.821	0.811	0.008
$N = 16$	III	1.832	1.851	0.037	2.837	2.85	0.053	3.561	3.665	0.057	0.798	0.799	0.006
$n = 256$	I	1.891	1.927	0.027	2.806	2.878	0.045	3.666	3.658	0.058	0.829	0.829	0.004
$m = 8$	II	1.969	2.004	0.032	3.026	3.055	0.049	3.612	3.646	0.06	0.839	0.834	0.004
$N = 16$	III	1.86	1.866	0.029	2.769	2.818	0.046	3.699	3.705	0.059	0.81	0.809	0.003
$n = 512$	I	2.026	2.062	0.03	3.074	3.077	0.05	3.779	3.817	0.052	0.806	0.806	0.002
$m = 9$	II	2.136	2.167	0.032	3.234	3.248	0.05	3.819	3.827	0.052	0.807	0.809	0.003
$N = 16$	III	1.866	1.869	0.026	2.82	2.789	0.043	3.78	3.803	0.052	0.797	0.797	0.002
$n = 256$	I	2.041	2.058	0.019	2.767	2.831	0.032	3.728	3.694	0.045	0.854	0.879	0.008
$m = 8$	II	1.851	1.85	0.028	2.792	2.825	0.039	3.778	3.74	0.048	0.829	0.835	0.005
$N = 32$	III	1.908	1.912	0.027	2.794	2.837	0.036	3.881	3.88	0.047	0.801	0.806	0.004
$n = 512$	I	2.007	2.041	0.021	2.922	2.949	0.03	3.787	3.837	0.037	0.808	0.812	0.003
$m = 9$	II	2.055	2.077	0.022	3.017	3.048	0.031	3.786	3.829	0.038	0.806	0.809	0.003
$N = 32$	III	1.974	1.99	0.021	2.861	2.89	0.029	3.832	3.875	0.038	0.797	0.8	0.003

SNR = 4		$\hat{\theta}_1$ ($\theta_1 = 2$)			$\hat{\theta}_2$ ($\theta_2 = 3$)			$\hat{\theta}_3$ ($\theta_3 = 4$)			$\hat{\sigma}$ ($\sigma \approx 0.8$)		
configu-ration	method	median	mean	s.e.	median	mean	s.e.	median	mean	s.e.	median	mean	s.e.
$n = 128$	I	1.839	1.938	0.031	2.689	2.745	0.043	3.363	3.459	0.058	0.661	0.669	0.006
$m = 7$	II	1.709	1.706	0.05	2.564	2.595	0.06	3.367	3.441	0.073	0.659	0.655	0.007
$N = 16$	III	1.773	1.795	0.038	2.616	2.606	0.052	3.532	3.601	0.062	0.621	0.622	0.004
$n = 256$	I	1.815	1.83	0.026	2.712	2.766	0.044	3.545	3.545	0.049	0.638	0.644	0.003
$m = 8$	II	1.839	1.865	0.031	2.88	2.895	0.048	3.506	3.534	0.05	0.648	0.649	0.003
$N = 16$	III	1.784	1.781	0.028	2.69	2.71	0.045	3.618	3.63	0.049	0.623	0.625	0.002
$n = 512$	I	2.039	2.064	0.029	3.09	3.086	0.045	3.754	3.731	0.049	0.615	0.613	0.002
$m = 9$	II	2.166	2.195	0.03	3.303	3.274	0.046	3.796	3.749	0.049	0.615	0.615	0.002
$N = 16$	III	1.871	1.874	0.027	2.796	2.784	0.04	3.75	3.721	0.05	0.609	0.607	0.002
$n = 256$	I	1.969	1.991	0.018	2.746	2.768	0.031	3.446	3.503	0.039	0.664	0.682	0.006
$m = 8$	II	1.825	1.831	0.026	2.748	2.76	0.038	3.471	3.57	0.041	0.642	0.645	0.004
$N = 32$	III	1.901	1.888	0.024	2.78	2.783	0.036	3.647	3.712	0.039	0.62	0.622	0.003
$n = 512$	I	1.926	1.946	0.019	2.888	2.917	0.03	3.653	3.754	0.037	0.616	0.617	0.002
$m = 9$	II	1.977	1.992	0.02	2.995	2.999	0.032	3.665	3.745	0.037	0.616	0.616	0.002
$N = 32$	III	1.9	1.902	0.018	2.806	2.848	0.03	3.721	3.796	0.037	0.609	0.609	0.002

7. Real data sets

7.1. *Prostate cancer data*

This data set has been studied by Tibshirani (1996). It examines the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. Table 8 lists the variables in this data set. The full model is

$$\begin{aligned}\texttt{lpsa} =& \mu + f_1(\texttt{lcavol}) + f_2(\texttt{lweight}) + f_3(\texttt{age}) + f_4(\texttt{lbph}) + f_5(\texttt{gleason}) \\ &+ f_6(\texttt{pgg45}) + I(\texttt{svi} = 0)u_0 + I(\texttt{svi} = 1)u_1 + \epsilon,\end{aligned}$$

where μ is the overall mean, f_p are the additive components, $u_0, u_1 \sim N(0, \theta^2\sigma^2)$ are the random effects corresponding to variable `svi`, and ϵ is $N(0, \sigma^2)$ random error. In this model, we mapped the domains of all continuous predictors into $[0, 1]$.

TABLE 8
Variables of the Prostate Cancer Data Set

Variable Name	Description	Type
`lpsa`	log of prostate specific antigen	Continuous
`lcavol`	log cancer volume	Continuous
`lweight`	log prostate weight	Continuous
`age`	age	Continuous
`lbph`	log of benign prostatic hyperplasia amount	Continuous
`lcp`	log of capsular penetration	Continuous
`gleason`	gleason score	Continuous
`pgg45`	Gleason score 4 or 5	Continuous
`svi`	seminal vesicle invasion	Binary

The analysis by Tibshirani (1996) concludes that variables `lcavol`, `lweight` and `svi` are important in predicting the level of prostate specific antigen. Our proposed method gives similar results: we confirm that `lcavol` and `lweight` are related to `lpsa`, the binary predictor `svi` is significant, and all the remaining variables are insignificant. Plots of the estimated fixed effects f_p's and fitted values can be found in Figure 1.

7.2. Depression data

It is of interest to examine the degree to which the effects of time-varying drug plasma levels on the change in depression levels over time. In this study each patient was treated for 4 weeks with imipramine. Blood samples were drawn twice per week, 15 hours after the last drug intake, and imipramine and desipramine concentrations were measured. Using this data set, Reisby *et al.* (1977) examine the relationship amongst imipramine (`imi`), desipramine (`dmi`) and clinical response in 66 depressed inpatients (37 endogenous and 29 non-endogenous). The variables of this data set is summarized in Table 9.

The full model is

$$\begin{aligned}\texttt{hamd}_{ik} =& \mu + u_i + f_1(\texttt{week}_{ik}) + f_2(\texttt{week}_{ik} \times \texttt{endog}_i) + f_3(\texttt{imi}_{ik}) \\ &+ f_4(\texttt{imi}_{ik} \times \texttt{endog}_i) + f_5(\texttt{dmi}_{ik}) + f_6(\texttt{dmi}_{ik} \times \texttt{endog}_i) + \epsilon_{ik},\end{aligned}$$

where $i = 1, \ldots, 66$, $k = 1, \ldots, 4$, μ is the overall mean, f_p's are the additive components, $u_i \sim N(0, \theta^2\sigma^2)$'s are the random effects corresponding to different inpatients and ϵ_{ik}'s are i.i.d. $N(0, \sigma^2)$ errors. Note that we introduced a random intercept u_i for each subject, and interaction terms of `endog` with other continuous predictors. We have also mapped the domains of all the continuous predictors into $[0, 1]$.

With this data set, we demonstrate that our proposed method can handle repeated measurement data. The variation among different patients can be explained by a random intercept term. Variable `id` represents an important characteristics of repeated measurement data and is kept in our model. And the group lasso procedure can help to choose the additive components.

Hedeker and Gibbons (2006) have also investigated this data set using linear mixed model and we obtained similar conclusions: larger `dmi` values lead to

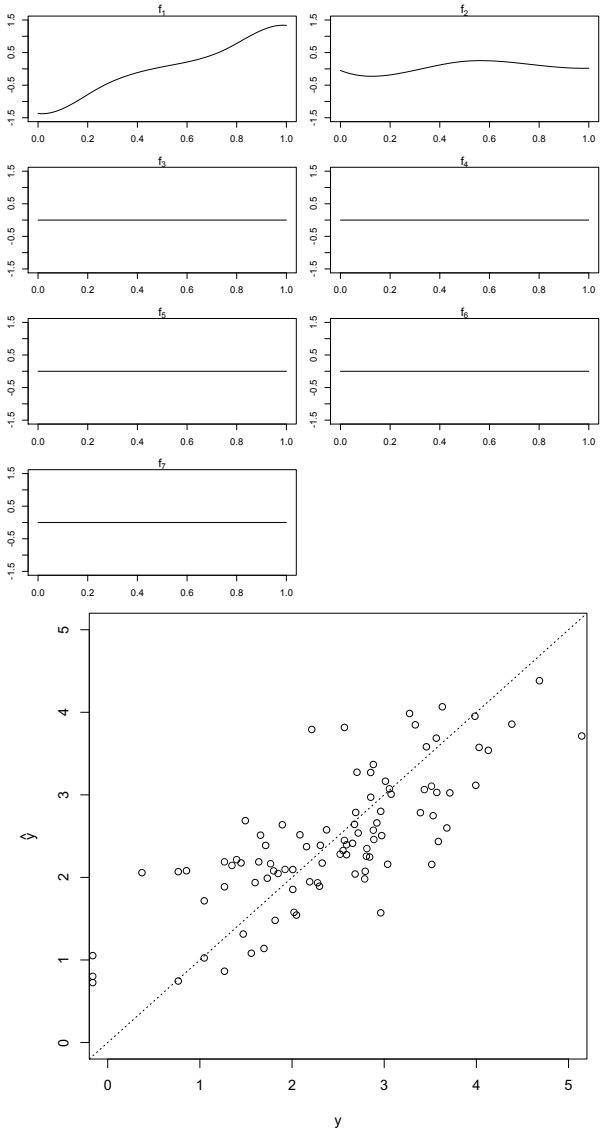

FIG 1. *Top: estimated fixed effects f_p's of the Prostate Cancer Data Set. Bottom: fitted values verses response.*

TABLE 9
Variables of the Depression Data Set

Variable Name	Description	Type
`id`	subject number	Discrete (1 to 66)
`hamd`	Hamilton Depression Scores	Continuous
`endog`	endogenous (=1) or non-endogenous (=0)	Binary
`week`	week	Continuous
`imi`	imipramine drug-plasma levels ($\mu g/l$)	Continuous
`dmi`	desipramine drug-plasma levels ($\mu g/l$)	Continuous

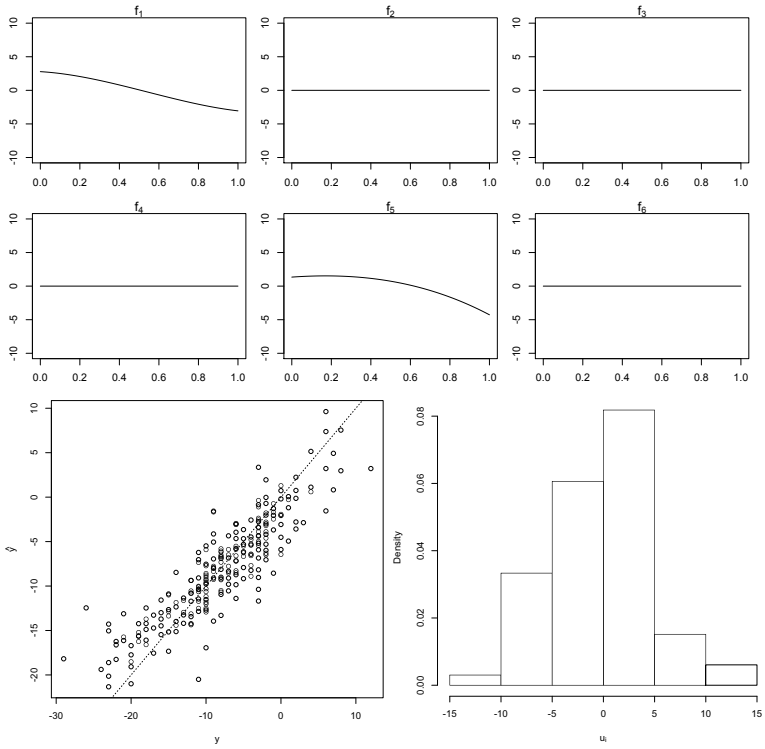

FIG 2. *Top two rows: estimated functions of the depression data set. Bottom left: fitted values verses response. Bottom right: histogram for the predicted* u_i.

greater improvement in depression (i.e., more negative `hamd` scores), while the drug `imi` is not significantly related to the `hamd` scores. Plots of estimated fixed effects f_p's, histograms of the estimated u_i's, together with fitted values can be found in Figure 2.

8. Summary

In this paper we studied the problem of variable selection in the context of nonparametric additive mixed modeling. The mixed modeling framework provides a method to jointly handle additive fixed effects as well as random effects. The additive fixed effects component is approximated by truncated series expansions with B-spline bases, with consistency properties of the spline approximation established. We have considered fitting the nonparametric fixed components with the adaptive group lasso methodology. This adaptive group lasso procedure yields sparse estimates for the coefficients corresponding to the B-spline bases. An model selection criterion derived from BIC, which is proven to be consistent, is presented for selecting the tuning parameters (λ and ν) of the adaptive group lasso procedure. Empirical results show that the new methodology is very

efficient in fitting high dimensional data with sparse solution, especially for the case of repeated measurement model with a number of continuous covariates.

Acknowledgment

The authors are grateful to the reviewer and the editor for many constructive comments, which led to a much improved version of the paper.

Appendix

Appendix A: Derivation of (11) and (12)

This appendix derives (11) and (12). In below $\otimes$ denotes the Kronecker product operator and vec$\boldsymbol{X}$ denotes the vector obtained by stacking the columns of the matrix $\boldsymbol{X}$. We begin by calculating

$$\begin{aligned}\frac{\partial \boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r}}{\partial \mathrm{vec}\boldsymbol{G}} &= \frac{\partial \mathrm{vec}\boldsymbol{V}^*}{\partial \mathrm{vec}\boldsymbol{G}}\frac{\partial \boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r}}{\partial \mathrm{vec}\boldsymbol{V}^*}\\ &= -\boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{r}\otimes \boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{r}\\ &= -\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1'}\boldsymbol{Z}),\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r}}{\partial \mathrm{vec}(\boldsymbol{G}')\partial \mathrm{vec}(\boldsymbol{G})'} &= (\boldsymbol{Z}'\otimes\boldsymbol{Z}')\Big\{(\boldsymbol{V}^{*-1'}\boldsymbol{r}\otimes\boldsymbol{V}^{*-1}\boldsymbol{Z})(\boldsymbol{r}'\boldsymbol{V}^{*-1'}\boldsymbol{Z}\otimes\boldsymbol{I})\\ &\quad + (\boldsymbol{V}^{*-1'}\boldsymbol{Z}\otimes\boldsymbol{V}^{*-1}\boldsymbol{r})(\boldsymbol{I}\otimes\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{Z})\Big\}\\ &= \boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1'}\boldsymbol{Z}\otimes\boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{Z}\\ &\quad + \boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{Z}\otimes\boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{Z},\end{aligned}$$

$$\frac{\partial \log(|\boldsymbol{V}^*|)}{\partial \mathrm{vec}\boldsymbol{G}} = (\boldsymbol{Z}'\otimes\boldsymbol{Z}')\mathrm{vec}(\boldsymbol{V}^{*-1'}) = \mathrm{vec}(\boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{Z})$$

and

$$\begin{aligned}\frac{\partial^2 \log(|\boldsymbol{V}^*|)}{\partial \mathrm{vec}(\boldsymbol{G}')\partial \mathrm{vec}(\boldsymbol{G})'} &= -(\boldsymbol{Z}'\otimes\boldsymbol{Z}')(\boldsymbol{V}^{*-1'}\boldsymbol{Z}\otimes\boldsymbol{V}^{*-1}\boldsymbol{Z})\\ &= -\boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{Z}\otimes\boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{Z}.\end{aligned}$$

Let $\boldsymbol{A} = \boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{r}(\boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{r})'$ and $\boldsymbol{B} = \boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{Z}$. Since $\boldsymbol{G}$ is symmetric, the gradient and Hessian matrix w.r.t. vec$\boldsymbol{G}$ are

$$\begin{aligned}\frac{\partial p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G})} &= \mathrm{vec}(\boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{Z}) - \frac{n}{\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r}}\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{V}^{*-1'}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1'}\boldsymbol{Z})\\ &= \mathrm{vec}\boldsymbol{B} - \frac{1}{\hat{\sigma}^2}\mathrm{vec}\boldsymbol{A}\end{aligned}$$

and

$$\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G})\partial \mathrm{vec}(\boldsymbol{G})'} = \frac{\partial \mathrm{vec}(\boldsymbol{G}')}{\partial \mathrm{vec}(\boldsymbol{G})}\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G}')\partial \mathrm{vec}(\boldsymbol{G})'},$$

where

$$\begin{aligned}
&\frac{\partial^2 p(\boldsymbol{G}|\hat{\boldsymbol{\beta}})}{\partial \mathrm{vec}(\boldsymbol{G}')\partial \mathrm{vec}(\boldsymbol{G})'} \\
&= -\boldsymbol{Z}'\boldsymbol{V}^{*-1\prime}\boldsymbol{Z} \otimes \boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{Z} \\
&\quad - \frac{n}{(\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r})^2}\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{V}^{*-1\prime}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1\prime}\boldsymbol{Z})\mathrm{vec}(\boldsymbol{Z}'\boldsymbol{V}^{*-1\prime}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1\prime}\boldsymbol{Z})' \\
&\quad + \frac{n}{\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{r}}\Big(\boldsymbol{Z}'\boldsymbol{V}^{*-1\prime}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1\prime}\boldsymbol{Z} \otimes \boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{Z} \\
&\quad + \boldsymbol{Z}'\boldsymbol{V}^{*-1\prime}\boldsymbol{Z} \otimes \boldsymbol{Z}'\boldsymbol{V}^{*-1}\boldsymbol{r}\boldsymbol{r}'\boldsymbol{V}^{*-1}\boldsymbol{Z}\Big) \\
&= -\boldsymbol{B} \otimes \boldsymbol{B} - \frac{1}{n\hat{\sigma}^4}\mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{A})' + \frac{1}{\hat{\sigma}^2}\Big(\boldsymbol{A} \otimes \boldsymbol{B} + \boldsymbol{B} \otimes \boldsymbol{A}\Big).
\end{aligned}$$

With the above (11) and (12) can be obtained.

Appendix B: Lemmas

This appendix presents a few lemmas which will be required for proofing our major theoretical results.

Lemma 1. *Under (A.3) and (A.4), for every $f_p \in \mathcal{F}$, there exists $\tilde{f}_p \in \mathcal{S}(d, \boldsymbol{t}_p)$ s.t.*

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\tilde{f}_p(X_{ip}) - f_p(X_{ip})\Big\}^2 = O_p(m^{-d}).$$

Proof. See lemma 1 of Huang *et al.* (2010). □

Lemma 2. *Let $\boldsymbol{w} = (w_p\frac{\hat{\boldsymbol{\beta}}_p'}{\|\hat{\boldsymbol{\beta}}_p\|}, p \in A_T)'$. Under (A.1) and (A.5)(b) we have $\|\boldsymbol{w}\| = o_p(1)$.*

Proof.

$$\|\boldsymbol{w}\|^2 = \sum_{p\in A_T} w_p^2 = \sum_{p\in A_T}(\|\tilde{\boldsymbol{\beta}}_p\|^{-1} - \nu)_+^2 \le |A_T| \max_{p\in A_T}(\|\tilde{\boldsymbol{\beta}}_p\|^{-1} - \nu)_+^2 = o_p(1).$$

□

Lemma 3. *For $p = 1, \ldots, P, j = 1, \ldots, m$, let $\boldsymbol{a}_{pj} = [\phi_{pj}(X_{ip})]_{i=1}^n$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V})$, then*

$$\|\boldsymbol{a}_{pj}\| = O_p\Big(\sqrt{\frac{n}{m}}\Big), \quad \boldsymbol{a}_{pj}'\boldsymbol{V}^{-1}\boldsymbol{e} = O_p\Big(\sqrt{\frac{n}{m}}\Big) \quad \text{and} \quad \|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{e}\| = O_p(\sqrt{n}).$$

Proof. By properties of B-splines, there exist c_1 and c_2 s.t.

$$\|\boldsymbol{a}_{pj}\|_\infty \le \|[\phi_{pj}(X_{ip})]_{i=1}^n\|_\infty \le c_1 \quad \text{and} \quad \mathrm{E}[\phi_{pj}^2(X_{ip})] \le \frac{c_2}{m}.$$

Then,

$$\phi_{pj}^2(X_{ip}) - \mathrm{E}\phi_{pj}^2(X_{ip}) \le 2c_1^2 \quad \text{and} \quad \mathrm{Var}\Big[\phi_{pj}^2(X_{ip})\Big] \le 4c_1^4\frac{c_2}{m}. \tag{20}$$

By law of large number,

$$\frac{1}{n}\|\boldsymbol{a}_{pj}\|^2 = \frac{1}{n}\sum_{i=1}^n \phi_{pj}^2(X_{ip}) = O_p\Big(\frac{1}{m}\Big).$$

Therefore,

$$\boldsymbol{a}_{pj}'\boldsymbol{V}^{-1}\boldsymbol{e} = O_p\left(\sqrt{\frac{n}{m}}\right) \quad \text{and} \quad \|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{e}\| \le \sqrt{m}\max_{1\le j\le m}\boldsymbol{a}_{pj}'\boldsymbol{V}^{-1}\boldsymbol{e} = O_p(\sqrt{n}).$$

The proof completes. □

Lemma 4. *Let $\boldsymbol{X}_A$ be the design matrix corresponding to A. If $|A|$ is bounded, then*

$$\|\boldsymbol{X}_A\| = O_p\left(\sqrt{\frac{n}{m}}\right), \quad \textit{and} \quad \|(\boldsymbol{X}_A'\boldsymbol{V}^{-1}\boldsymbol{X}_A)^{-1}\| = O_p\Big(\frac{m}{n}\Big).$$

Proof. The result holds under (A.2) and (A.3). Please referring lemma 3 of Huang *et al.* (2010) and lemma 6.2 of Zhou *et al.* (1998) for details. □

Lemma 5. *For $p = 1, \ldots, P, j = 1, \ldots, m$, let $\boldsymbol{a}_{pj} = [\phi_{pj}(X_{ip})]_{i=1}^n$. Then*

$$E\Big[\max_{pj}\|\boldsymbol{a}_{pj}\|\Big] = O_p\left(\sqrt{\frac{n}{m}}\right).$$

Proof. By equation (20) and lemma A.1 of Van De Geer (2008),

$$\mathrm{E}\Big[\max_{pj}\Big|\frac{1}{n}\sum_{i=1}^n\{\phi_{pj}^2(X_{ip}) - \mathrm{E}\phi_{pj}^2(X_{ip})\}\Big|\Big] \le \sqrt{\frac{8c_1^4c_2\log(2mP)}{mn}} + 2c_1^2\frac{\log(2mP)}{n}.$$

Also,

$$\begin{aligned}\mathrm{E}\Big[\frac{1}{n}\max_{pj}\|\boldsymbol{a}_{pj}\|^2\Big] &= \mathrm{E}\Big[\max_{pj}\Big|\frac{1}{n}\sum_{i=1}^n\{\phi_{pj}^2(X_{ip})\}\Big|\Big]\\ \le &\mathrm{E}\Big[\max_{pj}\Big|\frac{1}{n}\sum_{i=1}^n\{\phi_{pj}^2(X_{ip}) - \mathrm{E}\phi_{pj}^2(X_{ip})\}\Big|\Big] + \max_{pj}\mathrm{E}\phi_{pj}^2(X_{ip})\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\Big[\max_{pj}\|\boldsymbol{a}_{pj}\|\Big] &\le \sqrt{\mathrm{E}\Big[\max_{pj}\|\boldsymbol{a}_{pj}\|^2\Big]}\\ &\le \Big\{\sqrt{\frac{8c_1^4c_2n\log(2mP)}{m}} + 2c_1^2\log(2mP) + \frac{c_2n}{m}\Big\}^{1/2}.\end{aligned}$$

Particularly, under (A.6)(a), we have $\frac{m\log(2mP)}{n} \to 0$, so $\mathrm{E}[\max_{pj}\|\boldsymbol{a}_{pj}\|] = O_p(\sqrt{\frac{n}{m}})$. Therefore, the result follows. □

Lemma 6. *Let* $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V})$, *for any* $A \subset \{1, \ldots, P\}$

$$E(\max_{p\in A}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{e}\|) = O_p(\sqrt{n\log(2mP)})$$

Proof. For $p = 1, \ldots, P, j = 1, \ldots, m$, let $\boldsymbol{a}_{pj} = [\phi_{pj}(X_{ip})]_{i=1}^n$. Condition on $\boldsymbol{a}_{pj}$, by lemma 2.2.1 and 2.2.2 of Van der Vaart and Wellner (1996),

$$\mathrm{E}(\max_{1\le p\le P, 1\le j\le m}\boldsymbol{a}_{pj}'\boldsymbol{V}^{-1}\boldsymbol{e}|\boldsymbol{a}_{pj}) \le C_1\sqrt{\log(2mP)}\cdot\max_{pj}\|\boldsymbol{a}_{pj}\boldsymbol{V}^{-1/2}\|,$$

for some $C_1 > 0$. Therefore, for some $C_2 > 0$,

$$\begin{aligned}\mathrm{E}(\max_{1\le p\le P, 1\le j\le m}\boldsymbol{a}_{pj}'\boldsymbol{V}^{-1}\boldsymbol{e}) &\le C_1\sqrt{\log(2mP)}\cdot\mathrm{E}\Big[\max_{pj}\|\boldsymbol{a}_{pj}\|\Big]\cdot\|\boldsymbol{V}^{-1/2}\| \\ &\le C_2\sqrt{\frac{n\log(2mP)}{m}}.\end{aligned}$$

Moreover,

$$\mathrm{E}(\max_{p\in A}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{e}\|) \le \sqrt{m}\cdot\mathrm{E}(\max_{1\le p\le P, 1\le j\le m}\boldsymbol{a}_{pj}'\boldsymbol{V}^{-1}\boldsymbol{e}) \le C_2\sqrt{n\log(2mP)}.$$

It completes the proof. □

Appendix C: Proofs of theorems

Proof of theorem 1. The proof follows that main arguments of Theorem 3 of Huang *et al.* (2010), but with some modified technical details. Recall the necessary and sufficient conditions for $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \ldots, \hat{\boldsymbol{\beta}}_P')'$ to be the solution of (5) are

$$\begin{cases}-\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \lambda w_p\frac{\hat{\boldsymbol{\beta}}_p}{\|\hat{\boldsymbol{\beta}}_p\|} = \boldsymbol{0}, & \text{if } \hat{\boldsymbol{\beta}}_p \ne \boldsymbol{0}; \\ \|\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\| \le \lambda w_p, & \text{if } \hat{\boldsymbol{\beta}}_p = \boldsymbol{0}.\end{cases} \tag{21}$$

Let $\boldsymbol{w} = (w_p\frac{\hat{\boldsymbol{\beta}}_p'}{\|\hat{\boldsymbol{\beta}}_p\|}, p \in A_T)'$. Define

$$\hat{\boldsymbol{\gamma}}_{A_T} = (\hat{\boldsymbol{\gamma}}_{A_T,p}', p \in A_T)' = (\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\boldsymbol{X}_{A_T})^{-1}(\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\boldsymbol{Y} - \lambda\boldsymbol{w}).$$

If

$$\begin{cases}\|\hat{\gamma}_{A_T,p}\| > 0, & \forall p \in A_T; \\ \|\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_{A_T}\hat{\boldsymbol{\gamma}}_{A_T})\| \le \lambda w_p, & \forall p \in A_0,\end{cases}$$

then $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\gamma}}_{A_T}', \boldsymbol{0}')'$ satisfies (21), $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}_{A_T}\hat{\boldsymbol{\gamma}}_{A_T}$ and $\mathrm{sign}(\|\hat{\boldsymbol{\beta}}_p\|) = \mathrm{sign}(\|\boldsymbol{\beta}_p\|)$ for $p = 1, \ldots, P$. Therefore, $\mathrm{sign}(\|\hat{\boldsymbol{\beta}}_p\|) = \mathrm{sign}(\|\boldsymbol{\beta}_p\|)$ for $p = 1, \ldots, P$ if

$$\begin{cases}\|\boldsymbol{\beta}_p\| - \|\hat{\boldsymbol{\gamma}}_{A_T,p}\| < \|\boldsymbol{\beta}_p\|, & \forall p \in A_T; \\ \|\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_{A_T}\hat{\boldsymbol{\gamma}}_{A_T})\| \le \lambda w_p, & \forall p \in A_0.\end{cases}$$

Now

$$\begin{aligned} &P(\text{sign}(\|\hat{\boldsymbol{\beta}}_p\|) \neq \text{sign}(\|\boldsymbol{\beta}_p\|), \exists p) \\ \leq &P(\|\hat{\boldsymbol{\gamma}}_{A_T,p} - \boldsymbol{\beta}_p\| \geq \|\boldsymbol{\beta}_p\|, \exists p \in A_T) \\ &+ P(\|\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_{A_T}\hat{\boldsymbol{\gamma}}_{A_T})\| > \lambda w_p, \exists p \in A_0). \end{aligned} \tag{22}$$

Next we show that the right hand side of (22) tends to zero, by using claim 1 and claim 2 below.

Claim 1:

$$P(\|\hat{\boldsymbol{\gamma}}_{A_T,p} - \boldsymbol{\beta}_p\| \geq \|\boldsymbol{\beta}_p\|, \exists p \in A_T) \to 0. \tag{23}$$

Let $\boldsymbol{T}_p = [\boldsymbol{T}_{p1}, \boldsymbol{T}_{p2}, \ldots, \boldsymbol{T}_{p|A_T|}]$, where $\boldsymbol{T}_{pj} = \begin{cases} \boldsymbol{I}, p = j; \\ \boldsymbol{0}, p \neq j. \end{cases}$

And let $\boldsymbol{C}_A = \boldsymbol{X}_A'\boldsymbol{V}^{-1}\boldsymbol{X}_A$, $\boldsymbol{e} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$, $\boldsymbol{\delta} = [\delta_i]_{i=1}^n$, where $\delta_i = \sum_{p=1}^P f(X_{ip}) - \boldsymbol{X}_{(i)}\boldsymbol{\beta}$ and $\boldsymbol{X}_{(i)}$ is the i^{th} row of $\boldsymbol{X}$. We have

$$\hat{\boldsymbol{\gamma}}_{A_T,p} - \boldsymbol{\beta}_p = \boldsymbol{T}_p\boldsymbol{C}_{A_T}^{-1}\Big\{\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}(\boldsymbol{e} + \boldsymbol{\delta}) - \lambda\boldsymbol{w}\Big\}.$$

By the triangle inequality,

$$\|\hat{\boldsymbol{\gamma}}_{A_T,p} - \boldsymbol{\beta}_p\| \leq \|\boldsymbol{T}_p\boldsymbol{C}_{A_T}^{-1}\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\boldsymbol{e}\| + \|\boldsymbol{T}_p\boldsymbol{C}_{A_T}^{-1}\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\boldsymbol{\delta}\| + \lambda\|\boldsymbol{T}_p\boldsymbol{C}_{A_T}^{-1}\boldsymbol{w}\|.$$

By lemma 4, $\|\boldsymbol{C}_{A_T}^{-1}\| = O_p(m/n)$ and $\|\boldsymbol{X}_{A_T}\| = O_p(\sqrt{n/m})$. In addition, by lemma 3, the first term

$$\|\boldsymbol{T}_p\boldsymbol{C}_{A_T}^{-1}\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\boldsymbol{e}\| \leq \|\boldsymbol{C}_{A_T}^{-1}\| \cdot \|\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\boldsymbol{e}\| = O_p(m/\sqrt{n}) = o_p(1).$$

By lemma 1, the second term

$$\|\boldsymbol{T}_p\boldsymbol{C}_{A_T}^{-1}\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\boldsymbol{\delta}\| \leq \|\boldsymbol{C}_{A_T}^{-1}\| \cdot \|\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}\| \cdot \|\boldsymbol{\delta}\| = O_p(m^{-d+1/2}) = o_p(1).$$

By lemma 2, the third term

$$\lambda\|\boldsymbol{T}_p\boldsymbol{C}_{A_T}^{-1}\boldsymbol{w}\| \leq \lambda\|\boldsymbol{C}_{A_T}^{-1}\| \cdot \|\boldsymbol{w}\| = O_p(\lambda m/n)o_p(1) = o_p(1).$$

With probability tending to one, $\|\hat{\boldsymbol{\gamma}}_{A_T,p} - \boldsymbol{\beta}_p\| = o_p(1)$. Thus, claim 1 is proved.

Claim 2:

$$P(\|\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_{A_T}\hat{\boldsymbol{\gamma}}_{A_T})\| > \lambda w_p, \exists p \in A_0) \to 0. \tag{24}$$

Let $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}_{A_T}\boldsymbol{C}_{A_T}^{-1}\boldsymbol{X}_{A_T}'\boldsymbol{V}^{-1}$. We have $\|\boldsymbol{H}\| = O_p(1)$ and

$$\frac{1}{\lambda w_p}\boldsymbol{X}_p'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_{A_T}\hat{\boldsymbol{\gamma}}_{A_T}) = \frac{1}{\lambda w_p}\boldsymbol{X}_p'\boldsymbol{V}^{-1}(\boldsymbol{H}\boldsymbol{e} + \boldsymbol{H}\boldsymbol{\delta} + \lambda\boldsymbol{X}_{A_T}\boldsymbol{C}_{A_T}^{-1}\boldsymbol{w}).$$

The second term

$$\frac{1}{\lambda w_p}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{H}\boldsymbol{\delta}\| \leq \frac{1}{\lambda w_p}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\| \cdot \|\boldsymbol{H}\| \cdot \|\boldsymbol{\delta}\| = O_p\Big(\frac{n}{\lambda r m^{d+1/2}}\Big) = o_p(1).$$

The third term

$$\frac{1}{w_p}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{X}_{A_T}\boldsymbol{C}_{A_T}^{-1}\boldsymbol{w}\| \le \frac{1}{w_p}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{X}_{A_T}\boldsymbol{C}_{A_T}^{-1}\|\cdot\|\boldsymbol{w}\| = O_p(r^{-1})o_p(1) = o_p(1).$$

By lemma 6,

$$\begin{aligned}
&P(\|\boldsymbol{X}_p{}'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}_{A_T}\hat{\boldsymbol{\gamma}}_{A_T})\| > \lambda w_p, \exists p \in A_0)\\
\to &P(\frac{1}{\lambda w_p}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{H}\boldsymbol{e}\| > 1, \exists p \in A_0)\\
= &P(\max_{p\in A_0}\frac{1}{\lambda w_p}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{H}\boldsymbol{e}\| > 1)\\
\le &\mathrm{E}\Big[\max_{p\in A_0}\frac{1}{\lambda w_p}\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}\boldsymbol{H}\boldsymbol{e}\|\Big] = O_p\Big(\frac{\sqrt{n\log(2mP)}}{\lambda r}\Big) = o_p(1).
\end{aligned}$$

Thus, claim 2 is proved. This together with claim 1 proves that

$$P\left(\mathrm{sign}(\|\hat{\boldsymbol{\beta}}_p\|) = \mathrm{sign}(\|\boldsymbol{\beta}_p\|), p = 1,\ldots,P\right) \to 1.$$

□

Proof of part 2 of theorem 1. Define the event

$$F \equiv \bigcap_{p\in A_0}\{\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{b}}_{A_T})\| \le \lambda w_p\} \cap \Big\{\min_{p\in A_T}\|\tilde{\boldsymbol{\beta}}_p\| \ge \nu^{-1}\Big\}.$$

It is straightforward to check that $\hat{\boldsymbol{b}}_{A_T}$ satisfies (21) on F. Therefore,

$$\begin{aligned}
P\left(\hat{\boldsymbol{\beta}} \ne \hat{\boldsymbol{b}}_{A_T}\right) = &P(\|\boldsymbol{X}_p'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{b}}_{A_T})\|\\
&> \lambda w_p, \exists p \in A_0) + P\Big(\min_{p\in A_T}\|\tilde{\boldsymbol{\beta}}_p\| < \nu^{-1}\Big). \qquad (25)
\end{aligned}$$

By a similar argument as claim 2, put $\boldsymbol{w} = \boldsymbol{0}$, we can show that the first term goes to zero as $n \to \infty$. This together with (A.5b) implies that the righthand side of (25) goes to zero as $n \to \infty$. This completes the proof of part 2. □

Proof of parts 3 and 4 of theorem 1. Since $\hat{\boldsymbol{\beta}}$ is the minimizer of the quantity $\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{p=1}^{P} w_p\|\boldsymbol{\beta}_p\|$, we have

$$\frac{1}{2}\left\|\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\right\|^2 + \lambda\sum_{p=1}^{P} w_p\|\hat{\boldsymbol{\beta}}_p\| \le \frac{1}{2}\left\|\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right\|^2 + \lambda\sum_{p=1}^{P} w_p\|\boldsymbol{\beta}_p\|.$$

Therefore,

$$\frac{1}{2}\left\|\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\right\|^2 - \frac{1}{2}\left\|\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right\|^2 \le \lambda\sum_{p=1}^{P} w_p(\|\boldsymbol{\beta}_p\| - \|\hat{\boldsymbol{\beta}}_p\|).$$

Since $\|\boldsymbol{\beta}_p\| = 0$ for $p \in A_0$, the left hand side

$$\begin{aligned}\lambda \sum_{p=1}^{P} w_p(\|\boldsymbol{\beta}_p\| - \|\hat{\boldsymbol{\beta}}_p\|) \leq & \lambda \sum_{p \in A_T} w_p(\|\boldsymbol{\beta}_p\| - \|\hat{\boldsymbol{\beta}}_p\|) \\ \leq & \lambda \Big(\sum_{p \in A_T} w_p^2 \Big)^{1/2} \Big(\sum_{p \in A_T} \|\boldsymbol{\beta}_p - \hat{\boldsymbol{\beta}}_p\|^2 \Big)^{1/2} \\ \leq & \lambda \Big(\sum_{p \in A_T} w_p^2 \Big)^{1/2} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|.\end{aligned}$$

Recall that $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{e} + \boldsymbol{\delta}$. On the other side, let $\boldsymbol{\eta} = \boldsymbol{V}^{-1/2}(\boldsymbol{e} + \boldsymbol{\delta})$, $\boldsymbol{\xi} = \boldsymbol{V}^{-1/2}\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$,

$$\boldsymbol{V}^{-1/2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{\eta} - \boldsymbol{\xi}$$

and

$$\|\boldsymbol{V}^{-1/2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|^2 = \|\boldsymbol{\eta}\|^2 - 2\boldsymbol{\eta}'\boldsymbol{\xi} + \|\boldsymbol{\xi}\|^2.$$

Therefore,

$$\|\boldsymbol{\xi}\|^2 \leq 2\lambda \Big(\sum_{p \in A_T} w_p^2 \Big)^{1/2} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| + 2\boldsymbol{\eta}'\boldsymbol{\xi}. \tag{26}$$

Let $A_1 = \{p : \|\boldsymbol{\beta}_p\| > 0 \text{ or } \|\hat{\boldsymbol{\beta}}_p\| > 0\}$, $\boldsymbol{P} = \boldsymbol{V}^{-1/2}\boldsymbol{X}_{A_1}\boldsymbol{C}_{A_1}^{-1}\boldsymbol{X}'_{A_1}\boldsymbol{V}^{-1/2}$, $\boldsymbol{P}\boldsymbol{\xi} = \boldsymbol{\xi}$. By Cauchy-Schwartz inequality and completing square,

$$2|\boldsymbol{\eta}'\boldsymbol{\xi}| = 2|\boldsymbol{\eta}'\boldsymbol{P}\boldsymbol{\xi}| \leq 2\|\boldsymbol{P}\boldsymbol{\eta}\| \cdot \|\boldsymbol{\xi}\| \leq 2\|\boldsymbol{P}\boldsymbol{\eta}\|^2 + \frac{1}{2}\|\boldsymbol{\xi}\|^2.$$

Then,

$$\|\boldsymbol{\xi}\|^2 \leq 4\lambda \Big(\sum_{p \in A_T} w_p^2 \Big)^{1/2} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| + 4\|\boldsymbol{P}\boldsymbol{\eta}\|^2.$$

There exists $C > 0$, s.t. $\|\boldsymbol{\xi}\|^2 \geq C\frac{n}{m}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$. By Completing square,

$$\begin{aligned}\frac{nC}{m}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq & \frac{4m}{nC}\lambda^2 \Big(\sum_{p \in A_T} w_p^2 \Big) + \frac{nC}{2m}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 + 4\|\boldsymbol{P}\boldsymbol{\eta}\|^2 \\ \leq & \frac{8m}{nC}\lambda^2 \Big(\sum_{p \in A_T} w_p^2 \Big) + 8\|\boldsymbol{P}\boldsymbol{\eta}\|^2 \\ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq & \frac{8m^2}{n^2C^2}\lambda^2 \Big(\sum_{p \in A_T} w_p^2 \Big) + \frac{8m}{nC}\|\boldsymbol{P}\boldsymbol{\eta}\|^2.\end{aligned}$$

By lemma 2, $\sum_{p \in A_T} w_p^2 = o_p(1)$. Since $\sqrt{|A_1|} = O_p(1)$, by similar arguments in claim 1,

$$\|\boldsymbol{P}\boldsymbol{\eta}\|^2 \leq \|\boldsymbol{P}\boldsymbol{V}^{-1/2}\boldsymbol{e}\|^2 + \|\boldsymbol{P}\boldsymbol{V}^{-1/2}\boldsymbol{\delta}\|^2 = O_p(m) + O_p(m^{-2d}).$$

So,

$$\sum_{p=1}^{P}\|\hat{\boldsymbol{\beta}}_p-\boldsymbol{\beta}_p\|^2=\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2=O_p\Big(\frac{m^2}{n}\Big)+O_p\Big(\frac{1}{m^{2d-1}}\Big)+O_p\Big(\frac{m^2\lambda^2}{n^2}\Big)o_p(1).$$

By spline properties, there exist constants c_1 and $c_2>0$,

$$c_1m^{-1}\|\hat{\boldsymbol{\beta}}_p-\boldsymbol{\beta}_p\|^2\le\|\hat{f}_p-f\|^2\le c_2m^{-1}\|\hat{\boldsymbol{\beta}}_p-\boldsymbol{\beta}_p\|^2.$$

Therefore,

$$\sum_{p=1}^{P}\|\hat{f}_p-f_p\|^2=O_p\Big(\frac{m}{n}\Big)+O_p\Big(\frac{1}{m^{2d}}\Big)+O_p\Big(\frac{m^2\lambda^2}{n^2}\Big)o_p(1).$$

□

Proof of theorem 2. We adopt the idea of Wang *et al.* (2007a). Let $\Omega_-\equiv\{(\lambda,\nu):\hat{A}_{\lambda,\nu}\cap A_T\neq A_T\}$ and $\Omega_+\equiv\{(\lambda,\nu):\hat{A}_{\lambda,\nu}\supsetneq A_T\}$. It is enough to show that as $n\to\infty$, the following hold:

$$P\Big(\mathrm{BIC}(\lambda_n^*,\nu_n^*)=\log(\hat{\sigma}^2_{A_T})+\frac{1}{n\hat{\sigma}^2_{A_T}}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{b}}_{A_T}-\boldsymbol{Z}\hat{\boldsymbol{u}}_{A_T}\|^2+\hat{\mathrm{df}}_{A_T}\frac{\log(n)}{n}\Big)\to1,\tag{27}$$

$$P\Big(\inf_{(\lambda,\nu)\in\Omega_-}\mathrm{BIC}(\lambda,\nu)>\mathrm{BIC}(\lambda_n^*,\nu_n^*)\Big)\to1,\tag{28}$$

$$P\Big(\inf_{(\lambda,\nu)\in\Omega_+}\mathrm{BIC}(\lambda,\nu)>\mathrm{BIC}(\lambda_n^*,\nu_n^*)\Big)\to1,\tag{29}$$

where $\hat{\mathrm{df}}_{A_T}$ is the degrees of freedom of $\hat{\boldsymbol{Y}}_{A_T}=\boldsymbol{X}\hat{\boldsymbol{b}}_{A_T}+\boldsymbol{Z}\hat{\boldsymbol{u}}_{A_T}$.

For (27), it follows from theorem 1 that $P\big(\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*}=\hat{b}_{A_T}\big)\to1$, as $n\to\infty$. By (27), (B.1) and (B.2) with $(\lambda,\nu)\in\Omega_-$, we have as $n\to\infty$,

$$\begin{aligned}
&\mathrm{BIC}(\lambda_n^*,\nu_n^*)\\
&=\log(\hat{\sigma}^2_{\hat{A}_{\lambda^*,\nu^*}})+\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda^*,\nu^*})'\boldsymbol{V}^{*-2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*})}{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*})'\boldsymbol{V}^{*-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*})}+o_p(1)\\
&\ge\log(\hat{\sigma}^2_{\hat{A}_{\lambda^*,\nu^*}})+\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*})'\boldsymbol{V}^{*-1/2}(\boldsymbol{I}-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}')\boldsymbol{V}^{*-1/2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*})}{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*})'\boldsymbol{V}^{*-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_n^*,\nu_n^*})}\\
&\quad+o_p(1)\\
&\to\log(\sigma^2_{A_T})+1+o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{BIC}(\lambda,\nu)\\
&=\log(\hat\sigma^2_{\lambda,\nu})+\frac{1}{n\hat\sigma^2_{\lambda,\nu}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu}-\boldsymbol{Z}\hat{\boldsymbol{u}}_{\lambda,\nu})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu}-\boldsymbol{Z}\hat{\boldsymbol{u}}_{\lambda,\nu})+o_p(1)\\
&\geq\log(\hat\sigma^2_{\hat A_{\lambda,\nu}})+\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})'\boldsymbol{V}^{*-2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})}{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})'\boldsymbol{V}^{*-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})}+o_p(1)\\
&\geq\log(\hat\sigma^2_{\hat A_{\lambda,\nu}})+\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})'\boldsymbol{V}^{*-1/2}(\boldsymbol{I}-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}')\boldsymbol{V}^{*-1/2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})}{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})'\boldsymbol{V}^{*-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu})}\\
&\quad+o_p(1)\\
&\to\min_{A:A\cap A_T\neq A_T}\log(\sigma^2_A)+1>\log(\sigma^2_{A_T})+1,
\end{aligned}$$

which imply (28). It remains to show (29). Under (B.3),

$$\Big\{\hat{\mathrm{df}}_{\hat{\boldsymbol{Y}}}(\lambda,\nu)-\hat{\mathrm{df}}_{\hat{\boldsymbol{Y}}}(\lambda^*_n,\nu^*_n)\Big\}-\Big\{\hat{\mathrm{df}}_{\hat A_{\lambda,\nu}}-\hat{\mathrm{df}}_{A_T}\Big\}\to 0\text{ in probability,}$$

Therefore, as $n\to\infty$,

$$\begin{aligned}
&n\Big(\mathrm{BIC}(\lambda,\nu)-\mathrm{BIC}(\lambda^*_n,\nu^*_n)\Big)\\
=&n\log\Big(\frac{\hat\sigma^2_{\lambda,\nu}}{\hat\sigma^2_{\lambda^*_n,\nu^*_n}}\Big)+\frac{\hat\sigma^2_{A_T}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu}-\boldsymbol{Z}\hat{\boldsymbol{u}}_{\lambda,\nu}\|^2}{\hat\sigma^2_{\lambda,\nu}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{b}}_{A_T}-\boldsymbol{Z}\hat{\boldsymbol{u}}_{A_T}\|^2}\\
&+\Big\{\hat{\mathrm{df}}_{\hat{\boldsymbol{Y}}}(\boldsymbol{G},\lambda,\nu)-\hat{\mathrm{df}}_{\hat{\boldsymbol{Y}}}(\boldsymbol{G},\lambda^*_n,\nu^*_n)\Big\}\log(n)\\
\to&\hat\sigma^{-2}_{A_T}n(\hat\sigma^2_{\lambda,\nu}-\hat\sigma^2_{A_T})+o_p(1)+1+\Big\{\hat{\mathrm{df}}_{\hat A_{\lambda,\nu}}-\hat{\mathrm{df}}_{A_T}\Big\}\log(n).
\end{aligned}$$

It follows that

$$\begin{aligned}
\inf_{(\lambda,\nu)\in\Omega_+}n\Big(\mathrm{BIC}(\lambda,\nu)-\mathrm{BIC}(\lambda^*_n,\nu^*_n)\Big)\geq&\hat\sigma^{-2}_{A_T}\min_{A:A\supsetneq A_T}n(\hat\sigma^2_A-\hat\sigma^2_{A_T})+1\\
&+(\hat{\mathrm{df}}_{\hat A_{\lambda,\nu}}-\hat{\mathrm{df}}_{A_T})\log(n)+o_p(1).\quad(30)
\end{aligned}$$

Note that $n(\hat\sigma^2_{A_T}-\hat\sigma^2_A)=O_p(1)$ follows non-central chi-square distribution. By (17) and (18),

$$\hat{\mathrm{df}}_A=\mathrm{tr}\Big[\{\boldsymbol{I}-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\}\boldsymbol{V}^{-1}\boldsymbol{H}_A+\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\Big]$$

and

$$\hat{\mathrm{df}}_{\hat A_{\lambda,\nu}}-\hat{\mathrm{df}}_{A_T}=\mathrm{tr}\Big[\{\boldsymbol{I}-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\}\boldsymbol{V}^{-1}\{\boldsymbol{H}_{\hat A_{\lambda,\nu}}-\boldsymbol{H}_{A_T}\}\Big],$$

where $\boldsymbol{H}_A=\boldsymbol{X}_A(\boldsymbol{X}'_A\boldsymbol{V}^{-1}\boldsymbol{X}_A)^{-1}\boldsymbol{X}'_A$.

It can be shown that $\boldsymbol{I}-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'$ and $\boldsymbol{H}_{\hat A_{\lambda,\nu}}-\boldsymbol{H}_{A_T}$ are strictly positive definite, therefore $\hat{\mathrm{df}}_{\hat A_{\lambda,\nu}}-\hat{\mathrm{df}}_{A_T}>0$ for $(\lambda,\nu)\in\Omega_+$. The righthand side of (30) diverges to $+\infty$ as $n\to\infty$, and hence (29) is satisfied. This completes the proof. □

References

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010) Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, **66**, 1069–1077. MR2758494

de Boor, C. (2001) *A practical guide to splines*. New York: Springer Verlag. MR1900298

Chen, Z. and Dunson, D. B. (2003) Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769. MR2025100

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002) *Analysis of longitudinal data*. USA: Oxford University Press. MR2049007

Fahrmeir, L. and Lang, S. (2001) Bayesian inference for generalized additive mixed models based on markov random field priors. *Applied Statistics*, 201–220. MR1833273

Harville, D. A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385. MR0368279

Hedeker, D. and Gibbons, R. (2006) *Longitudinal data analysis*. New York: Wiley. MR2284230

Huang, J., Horowitz, J. and Wei, F. (2010) Variable selection in nonparametric additive models. *The Annals of Statistics*, **38**, 2282–2313. MR2676890

Kinney, S. K. and Dunson, D. B. (2007) Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**, 690–698. MR2395705

Laird, N. and Ware, J. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13. MR0836430

Lin, X. and Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B*, **61**, 381–400. MR1680318

Lin, Y. and Zhang, H. (2006) Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, **34**, 2272. MR2291500

Lindstrom, M. and Bates, D. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 1014–1022. MR0997577

Meier, L., van de Geer, S. and Buhlmann, P. (2008) The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, **70**, 53. MR2412631

Meier, L., Van de Geer, S., Buhlmann, P. and Zurich, E. (2009) High-dimensional additive modeling. *The Annals of Statistics*, **37**, 3779–3821. MR2572443

Pu, W. and Niu, X.-F. (2006) Selecting mixed-effects models based on a generalized information criterion. *JMA*, **97**, 733–758. MR2236499

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse additive models. *Journal of the Royal Statistical Society Series B*, **71**, 1009–1030. MR2750255

Reisby, N., Gram, L., Bech, P., Nagy, A., Petersen, G., Ortmann, J., Ibsen, I., Dencker, S., Jacobsen, O., Krautwald, O. *et al.* (1977) Imipramine: clinical effects and pharmacokinetic variability. *Psychopharmacology*, **54**, 263–272.

Ruppert, D., Wand, M. and Carroll, R. (2003) *Semiparametric regression*. New York: Cambridge University Press. MR1998720

Shen, X. and Ye, J. (2002) Adaptive model selection. *Journal of the American Statistical Association*, **97**, 210–221. MR1947281

Stone, C. (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705. MR0790566

Stone, C. (1986) The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 590–606. MR0840516

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267–288. MR1379242

Van der Vaart, A. and Wellner, J. (1996) *Weak convergence and empirical processes: with applications to statistics*. New York: Springer Verlag. MR1385671

Van De Geer, S. (2008) High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, **36**, 614. MR2396809

Wand, M. (2003) Smoothing and mixed models. *Computational Statistics*, **18**, 223–250.

Wang, H., Li, R. and Tsai, C. (2007a) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568. MR2410008

Wang, L., Chen, G. and Li, H. (2007b) Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486.

Wei, F. and Huang, J. (2008) Consistent group selection in high-dimensional linear regression. *Tech. rep.*, Department of Statistics and Actuarial Science, University of Iowa.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68**, 49–67. MR2212574

Zeger, S. and Liang, K. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998) Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, **93**. MR1631369

Zhou, S., Shen, X. and Wolfe, D. (1998) Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26**, 1760–1782. MR1673277

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429. MR2279469