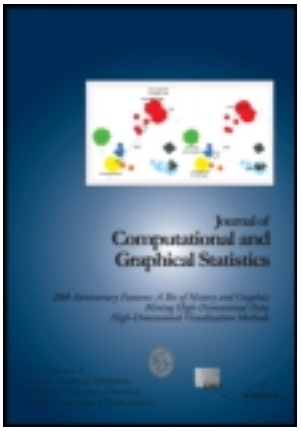


Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

# Variable Selection in Varying-Coefficient Models Using P-Splines

Anestis Antoniadis [a] , Irène Gijbels [b] & Anneleen Verhasselt [c]

[a] Laboratoire Jean Kuntzmann, Université Joseph Fourier, Grenoble, 38041, France

[b] Department of Mathematics and Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven (KU Leuven), 3001, Heverlee, Belgium

[c] Department of Mathematics and Computer Science, Universiteit Antwerpen, 2020, Antwerpen, Belgium

Accepted author version posted online: 24 May 2012.Version of record first published: 16 Aug 2012.

To cite this article: Anestis Antoniadis, Irène Gijbels & Anneleen Verhasselt (2012): Variable Selection in Varying-Coefficient Models Using P-Splines, Journal of Computational and Graphical Statistics, 21:3, 638-661

To link to this article: http://dx.doi.org/10.1080/10618600.2012.680826

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

# Variable Selection in Varying-Coefficient Models Using P-Splines

Anestis ANTONIADIS, Irène GIJBELS, and Anneleen VERHASSELT

In this article, we consider nonparametric smoothing and variable selection in varying-coefficient models. Varying-coefficient models are commonly used for analyzing the time-dependent effects of covariates on responses measured repeatedly (such as longitudinal data). We present the P-spline estimator in this context and show its estimation consistency for a diverging number of knots (or B-spline basis functions). The combination of P-splines with nonnegative garrote (which is a variable selection method) leads to good estimation and variable selection. Moreover, we consider APSO (additive P-spline selection operator), which combines a P-spline penalty with a regularization penalty, and show its estimation and variable selection consistency. The methods are illustrated with a simulation study and real-data examples. The proofs of the theoretical results as well as one of the real-data examples are provided in the online supplementary materials.

**Key Words:** Longitudinal data; Nonparametric smoothing; Penalized splines; Selecting variables; Varying regression coefficients.



## 1. INTRODUCTION

Varying-coefficient models (Hastie and Tibshirani 1993) are an extension of classical linear regression models in the sense that the regression coefficients vary in a smooth way with another variable (often time). We study varying-coefficient models of the following form:

$$Y(t) = \mathbf{X}(t)'\boldsymbol{\beta}(t) + \varepsilon(t) = \beta_0(t) + \sum_{p=1}^{d} X^{(p)}(t)\beta_j(t) + \varepsilon(t), \tag{1.1}$$

where $Y(t)$ is the response at time $t$ ($\in \mathcal{T} = [0, T]$); $\mathbf{X}(t) = (X^{(0)}(t), \ldots, X^{(d)}(t))'$ is the covariate vector at time $t$, with $X^{(0)}(t) \equiv 1$; $\boldsymbol{\beta}(t) = (\beta_0(t), \ldots, \beta_d(t))'$ is the vector of

Anestis Antoniadis is Professor, Laboratoire Jean Kuntzmann, Université Joseph Fourier, Grenoble 38041, France (E-mail: *anestis.antoniadis@imag.fr*). Irène Gijbels is Professor, Department of Mathematics and Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven (KU Leuven), 3001 Heverlee, Belgium (E-mail: *irene.gijbels@wis.kuleuven.be*). Anneleen Verhasselt is Assistant Professor, Department of Mathematics and Computer Science, Universiteit Antwerpen, 2020 Antwerpen, Belgium (E-mail: *Anneleen.Verhasselt.ua.ac.be*).


*Journal of Computational and Graphical Statistics, Volume 21, Number 3, Pages 638–661*
*DOI: 10.1080/10618600.2012.680826*

coefficients at time $t$; $\beta_0(t)$ is the baseline effect; and $\varepsilon(t)$ is a mean zero stochastic process at time $t$.

We consider longitudinal data, that is, samples with $n$ independent subjects or individuals, each measured repeatedly over a time period. The $j$th measurement for subject $i$ of $(t, Y(t), \mathbf{X}(t))$ is denoted by $(t_{ij}, Y_{ij}, \mathbf{X}_{ij})$, where $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant N_i$, $N_i$ is the number of repeated measurements of subject $i$, $t_{ij}$ is the measurement time, $Y_{ij}$ is the observed response at time $t_{ij}$, and $\mathbf{X}_{ij} = (X_{ij}^{(0)}, \ldots, X_{ij}^{(d)})'$. Let $N = \sum_{i=1}^{n} N_i$ be the total number of observations.

Several nonparametric smoothing techniques, such as local polynomials and smoothing splines (e.g., see Hoover et al. 1998) have been proposed for estimating the coefficient curves $\boldsymbol{\beta}(t)$ for longitudinal data. Lu, Zhang, and Zhu (2008) considered a penalized spline estimator in varying-coefficient models with time-independent covariates $\mathbf{X}$. In this special case, they showed the consistency and asymptotic normality of the estimates for a fixed number of knots. Recently, Wang, Li, and Huang (2008) used a basis function expansion and the Scad penalty (Fan and Li 2001) to estimate $\boldsymbol{\beta}(t)$ and select the relevant covariates $X^{(p)}(t)$. We use P-splines for estimating the coefficients, combined with the nonnegative garrote (Breiman 1995) to select the relevant variables and an extension of APSO (additive P-spline selection operator) of Antoniadis, Gijbels, and Verhasselt (forthcoming) to varying-coefficient models. For both methods, we show the estimation and variable selection consistency for the number of knots (thus also the number of B-splines in the basis expansion) tending to infinity with the total number of observations. The estimation consistency of P-splines is based on the estimation consistency of regular regression with B-splines as proved in Huang, Wu, and Zhou (2004). The variable selection consistency of the nonnegative garrote with P-splines in this context is, on the other hand, based on a consistency result for the nonnegative garrote of Yuan (2007). The idea behind the nonnegative garrote is very simple. It starts with a good initial estimator $\widehat{\boldsymbol{\beta}}^{\text{init}}(\cdot)$ for $\boldsymbol{\beta}(\cdot)$—we will use P-splines—for the regression coefficients and then shrinks some of the regression coefficients. As in Antoniadis, Gijbels, and Verhasselt (forthcoming), the APSO can also in this context be written as a grouped Lasso problem (Yuan and Lin 2006) and the consistency, therefore, follows from the consistency of the grouped Lasso (Bach 2008).

The article is organized as follows. In Section 2, we introduce P-splines in the varying-coefficient model context and show their consistency and asymptotic normality. The nonnegative garrote and its variable selection consistency are discussed in Section 3. In Section 4, the APSO is considered. We compare the methods with existing methods in the literature in Section 5 on simulated and real data examples (one of which is in the supplementary materials). Finally, in Section 6, some conclusions and a discussion are presented. The proofs of the theoretical results are deferred to the Appendix in the online supplementary materials.

## 2. P-SPLINE ESTIMATOR

### 2.1 UNIVARIATE P-SPLINE ESTIMATOR

P-splines were first introduced by Eilers and Marx (1996) in the univariate nonparametric smoothing context:

$$Y_i = f(X_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the $\varepsilon_i$'s are iid zero mean random variables with finite variance $\sigma^2$. P-splines are an extension of regression splines with a penalty on the coefficients of adjacent B-splines. Suppose that we have data $(X_i, Y_i)$, for $i = 1, \ldots, n$, with $X_i \in [a, b]$. To estimate $f(\cdot)$, we use a regression spline model $f(x) = \sum_{j=1}^{m} B_j(x; q)\alpha_j$, where $\{B_j(\cdot; q) : j = 1, \ldots, K + q = m\}$ is the $q$th-degree B-spline basis (using normalized B-splines such that $\sum_j B_j(x; q) = 1$) with $K + 1$ equidistant knot points $\xi_0 = a, \xi_1 = a + \frac{b-a}{K}, \ldots, \xi_K = b$ in $[a, b]$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ is the unknown column vector of regression coefficients. The penalized least-square estimator $\hat{\boldsymbol{\alpha}} = (\alpha_1, \ldots, \alpha_m)'$ is the minimizer of

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \Big( Y_i - \sum_{j=1}^{m} B_j(x_i; q)\alpha_j \Big)^2 + \lambda \sum_{j=k+1}^{m} (\Delta^k \alpha_j)^2, \tag{2.1}$$

where $\lambda > 0$ is the smoothing parameter and $\Delta$ is the differencing operator, that is, $\Delta^k \alpha_j = \sum_{t=0}^{k} (-1)^t \binom{k}{t} \alpha_{j-t}$, with $k \in \mathbb{N}$. In particular, for $k = 1$ and $k = 2$, this is $\Delta^1 \alpha_j = \alpha_j - \alpha_{j-1}$ and $\Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$, respectively. Rewriting expression (2.1) in matrix notation, we obtain

$$S(\boldsymbol{\alpha}) = (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha})'(\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}' \mathbf{D}_k' \mathbf{D}_k \boldsymbol{\alpha}, \tag{2.2}$$

where the elements $B_{ij}$ of $\mathbf{B}$ ($\in \mathbb{R}^{n \times m}$) are $B_j(x_i; q)$, $\mathbf{D}_k$ ($\in \mathbb{R}^{(m-k) \times m}$) is the matrix representation of the $k$th-order differencing operator $\Delta^k$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

### 2.2 P-Splines in Varying-Coefficient Models

Now, we discuss the extension of P-splines to varying-coefficient models (see Lu, Zhang, and Zhu 2008): suppose that for each $p = 0, \ldots, d$, $\beta_p(t)$ can be approximated by a B-spline basis expansion, that is, $\beta_p(t) = \sum_{l=1}^{m_p} B_{pl}(t; q_p)\alpha_{pl}$, where $\{B_{pl}(\cdot; q_p) : l = 1, \ldots, K_p + q_p = m_p\}$ is the $q_p$th-degree B-spline basis with $K_p + 1$ equidistant knots $\xi_{p0}, \ldots, \xi_{pK_p}$ for the $p$th component [which is a basis of the space $\mathbb{G}_p$ of spline functions on $\mathcal{T}$, with fixed degree $q_p$ and knot sequence $\Xi_p = (\xi_{p0}, \ldots, \xi_{pK_p})$]. In our consistency results, the number of knots $K_p + 1$ (and thus $m_p$) will grow with $n$, in contrast to the results of Lu, Zhang, and Zhu (2008), which are based on a fixed number of knots. Let

$$m_{\max} = \max_{0 \leqslant p \leqslant d} m_p,$$

the maximal size of the B-spline basis of the various components.

We then obtain the P-spline estimates of the regression coefficients $\alpha_{pl}$ by minimizing $S(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \ldots, \boldsymbol{\alpha}_d')' \in \mathbb{R}^{\dim \times 1}$, where $\boldsymbol{\alpha}_p = (\alpha_{p1}, \ldots, \alpha_{pm_p})'$ and $\dim = \sum_p m_p$:

$$\begin{aligned} S(\boldsymbol{\alpha}) &= \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \Big( Y_{ij} - \sum_{p=0}^{d} \sum_{l=1}^{m_p} X_{ij}^{(p)} B_{pl}(t_{ij}; q_p)\alpha_{pl} \Big)^2 + \sum_{p=0}^{d} \lambda_p \boldsymbol{\alpha}_p' \mathbf{D}_{k_p}' \mathbf{D}_{k_p} \boldsymbol{\alpha}_p \\ &= \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\alpha})' \mathbf{W}_i (\mathbf{Y}_i - \mathbf{U}_i \boldsymbol{\alpha}) + \boldsymbol{\alpha} \mathbf{Q}_\lambda \boldsymbol{\alpha}, \end{aligned}$$

where $k_p$ is the differencing order for the $p$th component, $\lambda_p$ are the smoothing parameters, and

$$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})'$$

$$\mathbf{B}(t) = \begin{pmatrix} B_{01}(t;q_0) & \ldots & B_{0m_0}(t;q_0) & 0 \ldots 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \ddots & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 \ldots 0 & B_{d1}(t;q_d) & \ldots & B_{dm_d}(t,q_d) \end{pmatrix}$$

$$\mathbf{U}'_{ij} = \mathbf{X}'_{ij}\mathbf{B}(t_{ij}) \in I\!R^{1\times \dim}$$

$$\mathbf{U}_i = (\mathbf{U}'_{i1}, \ldots, \mathbf{U}'_{iN_i})' \in I\!R^{N_i\times \dim}$$

$$\mathbf{W}_i = \operatorname{diag}\left(N_i^{-1}, \ldots, N_i^{-1}\right) \in I\!R^{N_i\times N_i} \quad \text{(a diagonal matrix with } N_i \text{ times } N_i^{-1} \text{ on the diagonal)}$$

$$\mathbf{Q}_\lambda = \operatorname{diag}(\lambda_0\mathbf{D}'_{k_0}\mathbf{D}_{k_0}, \ldots, \lambda_d\mathbf{D}'_{k_d}\mathbf{D}_{k_d}) \in I\!R^{\dim\times\dim} \quad \text{(a block diagonal matrix with the matrices } \lambda_p\mathbf{D}'_{k_p}\mathbf{D}_{k_p} \text{ on the diagonal).}$$

If $\sum_{i=1}^n \mathbf{U}'_i\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda$ is invertible (see Lemma 1), then $S(\boldsymbol{\alpha})$ has a unique minimizer:

$$\widehat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^n \mathbf{U}'_i\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda\right)^{-1} \sum_{i=1}^n \mathbf{U}'_i\mathbf{W}_i\mathbf{Y}_i, \tag{2.3}$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}'_0, \ldots, \widehat{\boldsymbol{\alpha}}'_d)'$ and $\widehat{\boldsymbol{\alpha}}_p = (\widehat{\alpha}_{p1}, \ldots, \widehat{\alpha}_{pm_p})'$ for $p = 0, \ldots, d$.

The P-spline estimate of $\boldsymbol{\beta}(t)$ is then

$$\widehat{\boldsymbol{\beta}}(t) = \mathbf{B}(t)\widehat{\boldsymbol{\alpha}} = (\widehat{\beta}_0(t), \ldots, \widehat{\beta}_d(t))', \quad \text{with} \quad \widehat{\beta}_p(t) = \sum_{l=1}^{m_p} B_{pl}(t;q_p)\widehat{\alpha}_{pl}.$$

The existence of the P-spline estimator relies on the fact that $\sum_{i=1}^n \mathbf{U}'_i\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda$ is invertible, which is given in the next lemma.

*Lemma 1.* The matrix $\sum_{i=1}^n \mathbf{U}'_i\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda$ is invertible, except on an event with probability tending to zero, if $m_{\max}^{3/2}\,\lambda_{\max}\,n^{-1} = o(1)$, where $\lambda_{\max} = \max_{0\leqslant p\leqslant d}\lambda_p$.

The proof of this lemma is deferred to the Appendix in the online supplementary materials. Denote by $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_n)' \in I\!R^{N\times\dim}$ and $\mathbf{W} = \operatorname{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n) \in I\!R^{N\times N}$. From the results in the proof of Lemma 1, we have the following approximation for $(\mathbf{U}'\mathbf{W}\mathbf{U} + \mathbf{Q}_\lambda)^{-1}$ and $\widehat{\boldsymbol{\alpha}}$ if $m_{\max}^{3/2}\lambda_{\max}n^{-1} \to 0$ and $m_{\max}n^{-1} \to$ constant:

$$(\mathbf{U}'\mathbf{W}\mathbf{U} + \mathbf{Q}_\lambda)^{-1} = (\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} - (\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} + o_P\left(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\right)(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}$$

and

$$\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_{\text{reg}} - \left( (\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} - o_P\left(\frac{m_{\max}^{5/2}\lambda_{\max}}{n^2}\right)\mathbf{1}_{\dim\times\dim}\right) \cdot \sum_{i=1}^{n} \mathbf{U}_i'\mathbf{W}_i\mathbf{Y}_i, \tag{2.4}$$

where $\widehat{\boldsymbol{\alpha}}_{\text{reg}}$ is the regular B-spline estimator (i.e., (2.3) with $\lambda_0 = \cdots = \lambda_d = 0$).

We will use this approximation (2.4) in Section 2.3 to prove the consistency of the P-spline estimator, exploiting the consistency of the regular B-spline estimator.

### 2.3 CONSISTENCY

We prove the consistency of the P-spline estimator in varying-coefficient models when the number of knots increases with the number of individuals $n$. In this approach, $\beta_p(t)$ is not a spline function itself, but can be approximated by a spline function. Conversely, if $\beta_p$ is a spline function, it can be represented exactly in a B-spline basis with a fixed number of knots. This last approach is considered in Lu, Zhang, and Zhu (2008).

The proof of our consistency result is based on the consistency of the regular B-spline estimator in varying-coefficient models (Huang, Wu, and Zhou 2004) and approximation (2.4).

We will use the following assumptions:

*Assumption 1.*

(1) The observation times $t_{ij}$, $j = 1, \ldots, N_i$, $i = 1, \ldots, n$, are chosen independently according to a distribution function $F_T$ on $\mathcal{T}$. Moreover, they are independent of the response and the covariate process $\{(Y_i(t), \mathbf{X}_i(t))\}$, $i = 1, \ldots, n$. The distribution function $F_T$ has a Lebesgue density $f_T(t)$ that is bounded away from zero and infinity uniformly over all $t \in \mathcal{T}$, that is, there exist positive constants $M_3$ and $M_4$ such that $M_3 \leqslant f_T(t) \leqslant M_4$ for $t \in \mathcal{T}$.

(2) The eigenvalues $\eta_0(t), \ldots, \eta_d(t)$ of $\Sigma(t) = \mathrm{E}(\mathbf{X}(t)\mathbf{X}(t)')$ are bounded away from zero and infinity uniformly over all $t \in \mathcal{T}$, that is, there exist positive constants $M_5$ and $M_6$ such that $M_5 \leqslant \eta_0(t) \leqslant \cdots \leqslant \eta_d(t) \leqslant M_6$ for $t \in \mathcal{T}$.

(3) There exists a positive constant $M_7$ such that $|X_p(t)| \leqslant M_7$ for $t \in \mathcal{T}$ and $p = 0, \ldots, d$.

(4) There exists a positive constant $M_8$ such that $\mathrm{E}(\varepsilon(t)^2) \leqslant M_8 < \infty$ for $t \in \mathcal{T}$.

(5) $\limsup_n(\frac{\max_p m_p}{\min_p m_p}) < \infty$.

(6) The process $\varepsilon(t)$ can be decomposed as the sum of two independent stochastic processes, $\varepsilon^{(1)}(t)$ and $\varepsilon^{(2)}(t)$, where $\varepsilon^{(1)}(t)$ is an arbitrary mean zero process and $\varepsilon^{(2)}(t)$ is a process of measurement errors that are independent at different time points and have mean zero and constant variance $\sigma^2$.

(7) $\frac{K_{\max}^{3/2}\lambda_{\max}}{n} = o(1)$ and $\frac{K_{\max}}{n} = O(1)$, where $K_{\max} = \max_{0\leqslant p\leqslant d} K_p$.

(8) $\max_i N_i < \infty$.

Assumption 1(1) guarantees that the observation times are randomly scattered. The results still hold when the observation times are deterministic, using the following assumption (for details see Huang, Wu, and Zhou 2004):

*Assumption 2.* There exist positive constants $M_9$ and $M_{10}$ such that

$$M_9||g||_{L_2}^2 \leqslant \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j g(t_{ij})^2 \leqslant M_{10}||g||_{L_2}^2, \quad g \in \mathbb{G}_p,\ p = 0,\ldots,d,$$

with $||g||_{L_2} = \sqrt{\int_{\mathcal{T}} g(t)^2\,dt}$ the $L_2$-norm of a square integrable function $g(t)$ on $\mathcal{T}$.

A sufficient condition for Assumption 2 to hold is (see Huang, Wu, and Zhou 2004) that

$$\sup_{t\in\mathcal{T}} |F_n(t) - F_T(t)| = o\left(\frac{1}{m_{\max}}\right)$$

for some distribution function $F_T$ with a Lebesgue density $f_T(t)$ that is bounded away from zero and infinity uniformly over $t \in \mathcal{T}$, where

$$F_n(t) = \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j 1_{]-\infty,t]}(t_{ij}),$$

with $1_A(t)$ as the indicator function on the set $A$ (i.e., $1_A(t) = 1$ if $t \in A$ and 0 else).

The assumption that the $t_{ij}$ are independent of each other can also be relaxed by replacing Assumption 1(1) by the requirement that Assumption 2 holds, with probability tending to 1.

Note that the Assumptions 1(1)–1(6) (which come from the consistency and asymptotic normality of the regular B-spline estimator) is natural and has also been used in Wang, Li, and Huang (2008). Moreover, Assumption 1(7) is a sufficient condition for $(\mathbf{U}'\mathbf{W}\mathbf{U} + \mathbf{Q}_\lambda)^{-1}$ to exist.

Before formulating the consistency theorem, we introduce some notation:

(1) Let $\text{dist}(\beta_p, \mathbb{G}_p) = \inf_{g\in\mathbb{G}_p} \sup_{t\in\mathcal{T}} |\beta_p(t) - g(t)|$ be the $L_\infty$ distance between $\beta_p(\cdot)$ and $\mathbb{G}_p$.

(2) Let $\rho_n = \max_{0\leqslant p\leqslant d} \text{dist}(\beta_p, \mathbb{G}_p)$ be the approximation error due to spline approximation.

(3) Let $\widetilde{\beta}_p(t) = \text{E}(\widehat{\beta}_p(t))$ be the mean of $\widehat{\beta}_p(t)$ conditioning on $\mathcal{X} = \{(\mathbf{X}_{ij}, t_{ij}); i = 1,\ldots,n,\ j = 1,\ldots,N_i\}$.

Theorem 2 gives the consistency of the P-spline estimator, and Theorem 1, its existence. The proofs are given in the Appendix.

*Theorem 1.* Suppose Assumptions 1(1)–1(5) and 1(7) hold,

$$\lim_n \operatorname{dist}(\beta_p, \mathbb{G}_p) = 0 \quad \text{for } p = 0, \ldots, d$$

and

$$\lim_n (m_{\max}\ \log(m_{\max})\, n^{-1}) = 0.$$

Then, $\widehat{\beta}_p(t)$ $(p = 0, \ldots, d)$ are uniquely defined, with probability tending to 1. Moreover, $\widehat{\beta}_p(t)$ $(p = 0, \ldots, d)$ are consistent in the sense that $||\widehat{\beta}_p(t) - \beta_p(t)||_{L_2} = o_P(1)$.

*Theorem 2.* Suppose Assumptions 1(1)–1(5) and 1(7) hold. If

$$\lim_n (m_{\max}\ \log(m_{\max})\, n^{-1}) = 0,$$

then

(1) $||\widetilde{\beta}_p(t) - \beta_p(t)||_{L_2} = o_P(\rho_n + m_{\max}^{3/2}\ \lambda_{\max}\ n^{-1})$,

(2) $||\widetilde{\beta}_p(t) - \widehat{\beta}_p(t)||_{L_2}^2 = o_P(r_n^2)$, where $r_n^2 = \frac{1}{n} + \frac{m_{\max}}{n^2} \sum_i \frac{1}{N_i}$.

Consequently,

$$||\widehat{\beta}_p(t) - \beta_p(t)||_{L_2} = o_P(\max(r_n, \rho_n, m_{\max}^{3/2}\ \lambda_{\max}\ n^{-1})).$$

With a specific choice of the numbers of knots and the smoothing parameters, Theorem 2 leads to the following corollary. The notation $a_n \asymp b_n$ means that $a_n b_n^{-1}$ and $b_n a_n^{-1}$ are bounded.

*Corollary 1.* Suppose Assumptions 1(1)–1(5), 1(7), and 1(8) hold and that $\beta_p(t)$ $(p = 0, \ldots, d)$ have bounded $q$th-order derivatives. Let $\mathbb{G}_p$ be a space of splines of degree no less than $q - 1$ and with $K_p \asymp (\frac{1}{n^2} \sum_{i=1}^n \frac{1}{N_i})^{-1/(2q+1)}$, $\lambda_p = (\frac{1}{n^2} \sum_{i=1}^n \frac{1}{N_i})^{\gamma}$ for $p = 0, \ldots, d$, where $\gamma \leqslant \frac{q-1/2}{2q+1}$. Then,

$$||\widehat{\beta}_p(t) - \beta_p(t)||_{L_2} = O_P\left(\left(\frac{1}{n^2} \sum_{i=1}^n \frac{1}{N_i}\right)^{q/(2q+1)}\right).$$

The proof of this corollary is deferred to the Appendix. Note that when $q = 2$ and the number of observations for each individual is bounded, then $K_p \asymp n^{1/5}$ and $||\widehat{\beta}_p(t) - \beta_p(t)||_{L_2} = O_P(n^{-2/5})$. The rate of convergence in this corollary is the optimal rate for nonparametric regression with iid data under the same smoothness assumptions on the $\beta_p$ (see Stone 1982).

### 2.4 ASYMPTOTIC NORMALITY

From (2.3), it is easy to see that the variance-covariance matrix of $\widehat{\boldsymbol{\alpha}}$ conditioning on $\mathcal{X}$ is

$$\operatorname{cov}(\widehat{\boldsymbol{\alpha}}) = \Big(\sum_i \mathbf{U}_i'\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda\Big)^{-1}\Big(\sum_i \mathbf{U}_i'\mathbf{W}_i\mathbf{V}_i\mathbf{W}_i\mathbf{U}_i\Big)\Big(\sum_i \mathbf{U}_i'\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda\Big)^{-1},$$

where $\mathbf{V}_i = \operatorname{cov}(\mathbf{Y}_i)$. The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}(t)$ conditioning on $\mathcal{X}$ is then

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}(t)) = \mathbf{B}(t)'\Big(\sum_i \mathbf{U}_i'\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda\Big)^{-1}\Big(\sum_i \mathbf{U}_i'\mathbf{W}_i\mathbf{V}_i\mathbf{W}_i\mathbf{U}_i\Big) \cdot\Big(\sum_i \mathbf{U}_i'\mathbf{W}_i\mathbf{U}_i + \mathbf{Q}_\lambda\Big)^{-1}\mathbf{B}(t).$$

Note that the diagonal elements of this matrix are $\operatorname{var}(\widehat{\beta}_p(t))$ for $p = 0, \ldots, d$.

The following theorem gives the asymptotic normality of the P-spline estimates $\widehat{\beta}_p(t)$. For the proof, see the Appendix in the online supplementary materials.

*Theorem 3.* Suppose Assumptions 1(1)–1(7) hold, and $\lim_n \rho_n = 0$, $\lim_n(m_{\max} \log m_{\max}\, n^{-1}) = 0$, and $\lim_n(m_{\max}\ \max_i N_i\, n^{-1}) = 0$ hold. Then,

$$((\operatorname{cov}(\widehat{\boldsymbol{\beta}}(t)))^{-1})^{1/2}(\widehat{\boldsymbol{\beta}}(t) - \widetilde{\boldsymbol{\beta}}(t)) \xrightarrow{\mathrm{D}} \mathrm{N}(0, \mathbf{I}_{d+1})$$

and, in particular,

$$(\operatorname{var}(\widehat{\beta}_p(t)))^{-1/2}(\widehat{\beta}_p(t) - \widetilde{\beta}_p(t)) \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1) \ \text{ for } p = 0, \ldots, d.$$

## 3. NONNEGATIVE GARROTE

### 3.1 ORIGINAL NONNEGATIVE GARROTE

Breiman (1995) proposed the nonnegative garrote for subset regression in a classical multiple linear regression model. It starts from an initial estimator, namely the ordinary least-square estimator (OLS), and shrinks or puts some coefficients of the OLS equal to zero. The data are of the form $(Y_i, X_{i1}, \ldots, X_{id})$ for $i = 1, \ldots, n$, and the multiple linear regression model is

$$Y_i = \beta_0 + \sum_{p=1}^{d} \beta_j X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n.$$

The nonnegative garrote shrinkage factors $\hat{c}_p$ are found by solving the following optimization problem:

$$\begin{cases} \min_{c_1,\ldots,c_d} \sum_{i=1}^n \Big(Y_i - \hat{\beta}_0^{\mathrm{OLS}} - \sum_{p=1}^d c_p \hat{\beta}_p^{\mathrm{OLS}} X_{ip}\Big)^2 + \theta \sum_{p=1}^d c_p \\ \text{s.t. } 0 \leqslant c_p \ (p = 1, \ldots, d), \end{cases}$$

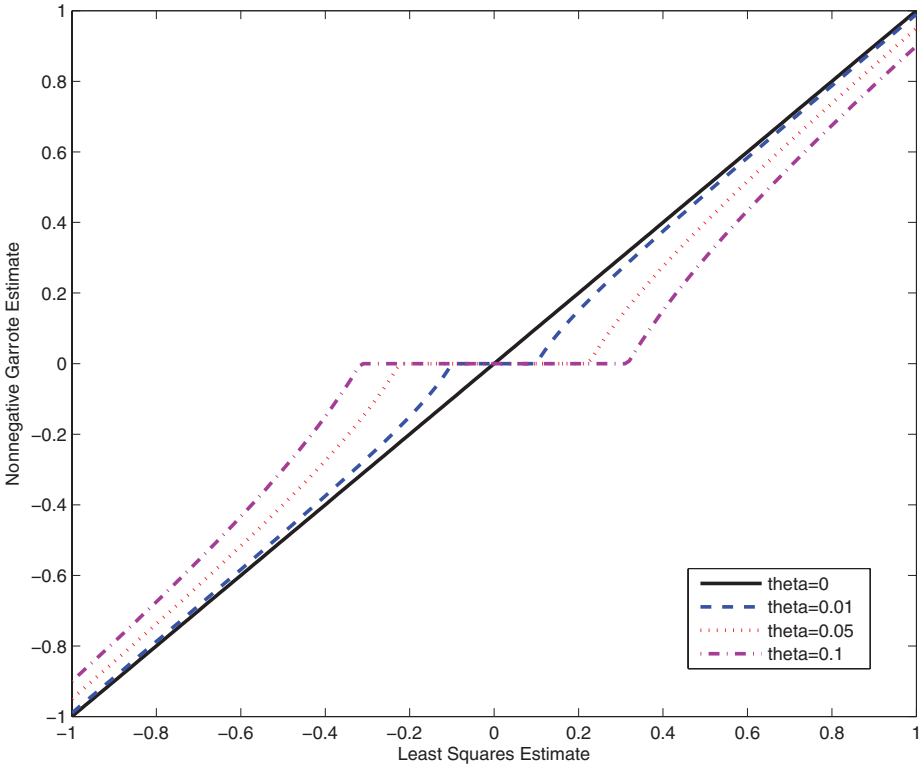



Figure 1. Shrinkage effect of the nonnegative garrote. The online version of this figure is in color.

where $\hat{\beta}_p^{\mathrm{OLS}}$ is the OLS of the regression coefficient of the $p$th component and $\theta > 0$ is a regularization parameter. The nonnegative garrote estimator of the regression coefficient is then

$$\hat{\beta}_p^{\mathrm{NNG}} = \hat{c}_p \hat{\beta}_p^{\mathrm{OLS}}.$$

In the special case that the design is orthogonal, that is, $\mathbf{X}'\mathbf{X} = \mathbf{I}_{n\times n}$, the nonnegative garrote estimates are the following:

$$\hat{c}_p = \left(1 - \frac{\theta}{2(\hat{\beta}_p^{\mathrm{OLS}})^2}\right)_+,$$

with $z_+ = \max(z, 0)$. This nonnegative garrote estimate is presented in Figure 1 for different values of $\theta$: the larger the $\theta$, the stronger the shrinkage effect.

### 3.2 NONNEGATIVE GARROTE IN VARYING-COEFFICIENT MODELS

In the varying-coefficient model setup, we define the nonnegative garrote shrinkage factors $\widehat{\mathbf{c}} = (\widehat{c}_0, \ldots, \widehat{c}_d)'$ as the solution of the following optimization problem:

$$\begin{cases} \min_{c_0,\ldots,c_d} \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{d} \left(Y_{ij} - \sum_{p=0}^{d} X_{ij}^{(p)} c_p \widehat{\beta}_p^{\mathrm{init}}(t_{ij})\right)^2 + \theta \sum_{p=0}^{d} c_p \\ \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d), \end{cases} \tag{3.1}$$

where $\widehat{\beta}_p^{\text{init}}(\cdot)$ is an initial estimator for the regression coefficient function $\beta_p(\cdot)$ and $\theta > 0$ is a regularization parameter. In the asymptotic study, we let $\theta$ depend on $N$, that is, $\theta = \theta_N$.

We will use the P-spline estimator, as introduced in Section 2.2, as an initial estimator. We first write (3.1) in matrix notation:

$$\begin{cases} \min_{\mathbf{c}} ||\widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\mathbf{c}||_2^2 + \theta \sum_{p=0}^{d} c_p \\ \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d), \end{cases}$$

where

$$\begin{aligned} \mathbf{Y} &= (\mathbf{Y}_1', \ldots \mathbf{Y}_n')' \in I\!R^{N\times 1} \\ \mathbf{W} &= \operatorname{diag}(\mathbf{W}_i)_{i=1,\ldots,n} \in I\!R^{N\times N} \\ \widetilde{\mathbf{Y}} &= \mathbf{W}^{1/2}\mathbf{Y} \\ \mathbf{z}_i^{(p)} &= \big(X_{i1}^{(p)}, \ldots, X_{iN_i}^{(p)}\big) \operatorname{diag}\big(\widehat{\beta}_p^{\text{init}}(t_{ij})\big)_{j=1,\ldots,N_i} \in I\!R^{1\times N_i} \\ \mathbf{Z}_p &= \big(\mathbf{z}_1^{(p)}, \ldots, \mathbf{z}_n^{(p)}\big)' \in I\!R^{N\times 1} \\ \mathbf{Z} &= (\mathbf{Z}_0, \ldots, \mathbf{Z}_d) \\ \widetilde{\mathbf{Z}} &= \mathbf{W}^{1/2}\mathbf{Z} \\ \mathbf{c} &= (c_0, \ldots, c_d)'. \end{aligned}$$

Note that the P-spline estimator

$$\widehat{f}_p(t) = X^{(p)}(t)\widehat{\beta}_p(t)$$

for the $p$th component $f_p(t) = X^{(p)}(t)\beta_p(t)$ is consistent (see the proof of Theorem 2). Let

$$\kappa_n = \max\big(\rho_n, r_n, m_{\max}^{3/2}\lambda_{\max}n^{-1}\big).$$

The nonnegative garrote estimator with the P-spline estimator as the initial estimator for $\beta_p(t)$ is

$$\widehat{f}_p^{\text{NNG}}(t) = \widehat{c}_p\widehat{f}_p(t).$$

It follows from Yuan (2007) [see also Antoniadis, Gijbels, and Verhasselt (forthcoming)] that $\widehat{f}_p^{\text{NNG}}(t)$ is estimation- and variable selection-consistent.

*Theorem 4.* If the assumptions of Theorems 1 and 2 hold and $\frac{\theta}{N} \to 0$ such that $\kappa_n = o(\frac{\theta}{N})$, then

(1) $P(\widehat{f}_p^{\text{NNG}}(t) = 0) \to 1$ for any $p$ such that $\beta_p(t) = 0$ for all $t \in \mathcal{T}$,

(2) $\sup_p \mathrm{E}(\widehat{f}_p^{\text{NNG}}(t) - f_p(t))^2 = O_P(\frac{\theta^2}{N^2})$ (where the expectation is with respect to $\mathcal{X}$) for all $t \in \mathcal{T}$.

## 4. ADDITIVE P-SPLINE SELECTION OPERATOR

The APSO (additive P-spline selection operator ) was introduced by Antoniadis, Gijbels, and Verhasselt (forthcoming) as an extension of P-splines to variable selection in additive models. The idea of this APSO is to encourage smoothness by using a P-spline penalty on differences of coefficients and selection by using a penalty on the regression coefficients. The APSO is inspired by the regularization technique Cosso (Lin and Zhang 2006) for additive models. From the simulations in Antoniadis, Gijbels, and Verhasselt (forthcoming), it is clear that nonnegative garrote (with P-splines) and the APSO outperform Cosso when the covariates are correlated.

In the varying-coefficient model context, we define the APSO regression coefficients $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_0', \ldots, \widehat{\boldsymbol{\alpha}}_d')'$ (where $\widehat{\boldsymbol{\alpha}}_p = (\widehat{\alpha}_{p1}, \ldots, \widehat{\alpha}_{pm_p})'$ for $p = 0, \ldots, d$) as the minimizer of

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \Big( Y_{ij} - \sum_{p=0}^{d} \sum_{l=1}^{m_p} X_{ij}^{(p)} \alpha_{pl} B_{pl}(t_{ij}) \Big)^2 + \lambda \sum_{p=0}^{d} \big( ||\boldsymbol{\alpha}_p||_2^2 + \mu ||\mathbf{D}_{k_p} \boldsymbol{\alpha}_p||_2^2 \big)^{1/2}$$

$$= \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \Big( Y_{ij} - \sum_{p=0}^{d} \sum_{l=1}^{m_p} X_{ij}^{(p)} \alpha_{pl} B_{pl}(t_{ij}) \Big)^2 + \lambda \sum_{p=0}^{d} \big( \boldsymbol{\alpha}_p' \mathbf{K}_p^{\mu} \boldsymbol{\alpha}_p \big)^{1/2}, \tag{4.1}$$

where $\mathbf{K}_p^{\mu} = \mathbf{I}_{m_p} + \mu \mathbf{D}_{k_p}' \mathbf{D}_{k_p}$.

Let $\mathbf{K}_p^{\mu} = \mathbf{V}_p' \mathbf{V}_p$ be the Cholesky decomposition of $\mathbf{K}_p^{\mu}$. We use a reparameterization $\widetilde{\boldsymbol{\alpha}}_p = \mathbf{V}_p \boldsymbol{\alpha}_p$ and use the same matrix notation as before [with $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_n)'$] to write (4.1) in matrix notation:

$$S(\widetilde{\boldsymbol{\alpha}}) = ||\widetilde{\mathbf{Y}} - \widetilde{\mathbf{C}} \widetilde{\boldsymbol{\alpha}}||_2^2 + \lambda \sum_{p=0}^{d} ||\widetilde{\boldsymbol{\alpha}}_p||_2, \tag{4.2}$$

where $\mathbf{C} = \mathbf{U}\mathbf{V}^{-1}$ (where $\mathbf{V}^{-1} = \operatorname{diag}(\mathbf{V}_p^{-1})_{p=0,\ldots,d}$) and $\widetilde{\mathbf{C}} = \mathbf{W}^{1/2}\mathbf{C}$. Note that we wrote the APSO optimization problem (4.1) as the grouped Lasso problem (4.2) (e.g., see Bach 2008 and Yuan and Lin 2006). The consistency of the APSO is, therefore, a corollary of the consistency of the grouped Lasso problem (Bach 2008).

## 5. NUMERICAL STUDY AND APPLICATIONS

We compare the performance of the nonnegative garrote with the P-splines (NGP) with that of the APSO (denoted later on as AgLasso). Moreover we also include comparisons with two other APSO-type methods, where the grouped Lasso penalty $\lambda \sum_{p=0}^{d} ||\widetilde{\boldsymbol{\alpha}}_p||_2$ of (4.2) is replaced, on the one hand, by a grouped Bridge penalty (AgBridge), $\lambda \sum_{p=0}^{d} m_p^{\gamma} ||\widetilde{\boldsymbol{\alpha}}_p||_1^{\gamma}$, and on the other hand, by a grouped Scad penalty (AgScad).

We used the R package `grpreg` to compute the APSO estimates with the three different penalties. In the practical implementation of AgBridge, we took $\gamma = 0.5$, and for AgScad, we used $a = 3.7$ as the Scad parameter. We also provide a comparison with the Scad-based method of Wang, Li, and Huang (2008), referred to as Scad in the sequel, using the R codes provided by the authors. For the implementation of the NGP method, we rely on the

Table 1. Evaluation criteria

| | |
|---|---|
| MS | Median number of selected components |
| ER | Mean of estimation error $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\sum_{p=0}^{d}(X^{(p)}(t_{ij})\beta_p(t_{ij}) - X^{(p)}(t_{ij})\hat{\beta}_p(t_{ij}))^2$ |
| MTZ | Median of zero components restricted to the true zero components |
| MFZ | Median of zero components restricted to the true nonzero components (false zeroes) |
| MTP | Median of the true positives |
| MFP | Median of the false positives |
| PercT | Percentage of replications that the exact true model was selected |
| AverR | Average of the number of relevant variables selected in the model |
| AverI | Average of the number of irrelevant variables selected in the model |
| time | Average of the computing time |



`lsqlin` procedure in the Optimization toolbox in MATLAB. The code for NGP is available from the authors upon request.

In the NGP and APSO (with the three different penalties), we use $q = 3$, $k = 2$, and $K = 7$ (i.e., 10 B-spline basis functions) for all components. The regularization parameter $\theta$ of the nonnegative garrote and $\lambda$ and $\mu$ of the APSO are chosen by BIC (Bayesian information criterion). The smoothing parameters $\lambda_0, \ldots, \lambda_d$ of the P-splines estimator are chosen by an EM (expectation–maximization) algorithm of Marx (2010). We compare the methods on the basis of the evaluation criteria described in Table 1, based on 500 simulated datasets.

The criteria MS and ER are also used in Huang, Horowitz, and Wei (2010), MTZ and MFZ in Fan and Li (2001), and PercT, AverR, and AverI in Wang, Li, and Huang (2008). The numbers in brackets in Tables 2 and 6 are the standard errors $(\cdot)$ or first and third quartiles $(\cdot, \cdot)$. To have an overall idea of the behavior of the methods, it is important to look at all criteria together. It is, for example, possible that for a method, MS $= 4$ (the true number of nonzero components in the first simulation example), but this might involve also selected components that are true zero components. Information on the occurrence of such events is to be found in MTZ and MFZ. In addition, we present the computing times on a compute node using two Intel Xeon E5540 processors and a 48-GB PC3-1066 SDRAM for the different methods.

### 5.1 First Simulation Example

We consider the following model, which has also been studied in Wang, Li, and Huang (2008):

$$Y(t_{ij}) = \sum_{p=0}^{23} X^{(p)}(t_{ij})\beta_p(t_{ij}) + s\,\varepsilon(t_{ij}) \quad \text{for } i = 1, \ldots, 200,\ \ j = 1, \ldots, N_i.$$

The first four variables $X^{(0)}(t)$, $X^{(1)}(t)$, $X^{(2)}(t)$, and $X^{(3)}(t)$ are the relevant variables. $X^{(0)}(t) \equiv 1$, $X^{(1)}(t)$ is uniform distributed on $[t/10, 2 + t/10]$ for any $t$, $X^{(2)}(t)$ conditioning on $X^{(1)}(t)$ is normal distributed with mean 0 and variance $\frac{1+X^{(1)}(t)}{2+X^{(1)}(t)}$, and $X^{(3)}(t)$ (independent of $X^{(1)}(t)$ and $X^{(2)}(t)$) is a Bernoulli random variable, with probability of success 0.6 (and thus it does not vary with $t$). The 20 irrelevant variables $X^{(p)}(t)$, $p = 4, \ldots, 23$ are

independent random realizations of a Gaussian process with mean zero and covariance structure $\operatorname{cov}(X^{(p)}(t), X^{(p)}(s)) = 4\exp^{-|t-s|}$.

The error $\varepsilon(t)$ is normal distributed with mean 0 and variance 1 for any $t$. The parameter $s$ controls the signal-to-noise ratio (SNR) and is taken to be, respectively, 1, 1.25, and 2, resulting in SNRs of, respectively, 6.25, 5, and 3.

The coefficient functions of the relevant variables are

$$\beta_0(t) = 15 + 20\sin\left(\frac{\pi t}{60}\right), \qquad \beta_1(t) = 2 - 3\cos\left(\frac{\pi(t-25)}{15}\right),$$
$$\beta_2(t) = 6 - 0.2t, \qquad \beta_3(t) = -4 + \frac{(20-t)^3}{2000},$$

and for the irrelevant variables $\beta_p(t) = 0$ $(p = 4, \ldots, 23)$. The observation time points $t_{ij}$ are the same for all subjects $(i = 1, \ldots, n)$: $\{1, \ldots, 30\}$.

Table 2 summarizes the results according to the criteria in Table 1. We can conclude that the nonnegative garrote and the APSO procedures outperform Scad in computing time.

Table 2. Simulation results for the first simulation example

| $s$ | Method / Optimal value | ER 0 | MS 4 | MTZ 20 | MFZ 0 | MTP 4 | MFP 0 | PercT 1 | AverR 3 | AverI 0 | Time 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AgLasso | 0.0380 | 4 | 20 | 0 | 4 | 0 | 0.978 | 3 | 0.0280 | 108.6698 |
| | | (0.0521) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0.1982) | (26.4597) |
| | AgBridge | 0.0075 | 4 | 20 | 0 | 4 | 0 | 0.978 | 3 | 0.0220 | 29.0388 |
| | | (0.0022) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0.1468) | (3.5160) |
| | AgScad | 0.0157 | 4 | 20 | 0 | 4 | 0 | 0.694 | 3 | 0.4260 | 37.9208 |
| | | (0.0034) | (4,5) | (19,20) | (0,0) | (4,4) | (0,1) | | (0) | (0.7468) | (4.5052) |
| | NGP | 0.0067 | 4 | 20 | 0 | 4 | 0 | 0.996 | 3 | 0.0040 | 6.0609 |
| | | (0.0022) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0.0632) | (0.3160) |
| | Scad | 0.0068 | 4 | 20 | 0 | 4 | 0 | 1 | 3 | 0 | 297.0027 |
| | | (0.0015) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0) | (47.6556) |
| 1.25 | AgLasso | 0.0351 | 4 | 20 | 0 | 4 | 0 | 0.982 | 3 | 0.0680 | 130.0369 |
| | | (0.0402) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0.7158) | (26.5147) |
| | AgBridge | 0.0111 | 4 | 20 | 0 | 4 | 0 | 0.858 | 3 | 0.1620 | 37.0656 |
| | | (0.0029) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0.4244) | (4.1698) |
| | AgScad | 0.0232 | 5 | 19 | 0 | 4 | 1 | 0.318 | 3 | 1.6780 | 42.3186 |
| | | (0.0057) | (4,7) | (17,20) | (0,0) | (4,4) | (0,3) | | (0) | (1.7586) | (4.7558) |
| | NGP | 0.0443 | 4 | 20 | 0 | 4 | 0 | 0.980 | 2.9980 | 0.0180 | 6.6155 |
| | | (0.7692) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0.0447) | (0.1331) | (0.7004) |
| | Scad | 0.0106 | 4 | 20 | 0 | 4 | 0 | 1 | 3 | 0 | 632.0616 |
| | | (0.0024) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0) | (1229.3704) |
| 2 | AgLasso | 0.0670 | 4 | 20 | 0 | 4 | 0 | 0.770 | 3 | 0.8780 | 125.4471 |
| | | (0.0674) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (2.4786) | (27.2323) |
| | AgBridge | 0.0345 | 7 | 17 | 0 | 4 | 3 | 0.064 | 3 | 3.4080 | 48.9450 |
| | | (0.0095) | (6,9) | (15,18) | (0,0) | (4,4) | (2,5) | | (0) | (2.2216) | (5.0557) |
| | AgScad | 0.0466 | 16 | 8 | 0 | 4 | 12 | 0.01 | 3 | 11.4240 | 47.5938 |
| | | (0.0117) | (12,19) | (5,12) | (0.0) | (4,4) | (8,15) | | (0) | (4.7368) | (5.6067) |
| | NGP | 0.1586 | 4 | 20 | 0 | 4 | 0 | 0.888 | 2.9920 | 0.1040 | 6.6008 |
| | | (1.5199) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0.0892) | (0.3056) | (0.7098) |
| | Scad | 0.0276 | 4 | 20 | 0 | 4 | 0 | 0.970 | 3 | 0.0300 | 629.2858 |
| | | (0.0067) | (4,4) | (20,20) | (0,0) | (4,4) | (0,0) | | (0) | (0.1708) | (1208.1391) |

NOTE: The numbers in brackets are the standard errors (for the continuous-valued criteria) or the first and third quartiles (for the discrete-valued criteria).

Table 3. First simulation example: appearance frequency of the variables

| $s$ | Method | $X^{(0)}$ | $X^{(1)}$ | $X^{(2)}$ | $X^{(3)}$ | $X^{(4)}$ | $X^{(5)}$ | $X^{(6)}$ | $X^{(7)}$ | $X^{(8)}$ | $X^{(9)}$ | $X^{(10)}$ | $X^{(11)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AgLasso | 500 | 500 | 500 | 500 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| | AgBridge | 500 | 500 | 500 | 500 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 |
| | AgScad | 500 | 500 | 500 | 500 | 8 | 10 | 12 | 13 | 7 | 12 | 9 | 12 |
| | NGP | 500 | 500 | 500 | 500 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | Scad | 500 | 500 | 500 | 500 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.25 | AgLasso | 500 | 500 | 500 | 500 | 3 | 4 | 0 | 4 | 2 | 2 | 1 | 2 |
| | AgBridge | 500 | 500 | 500 | 500 | 3 | 5 | 4 | 1 | 7 | 5 | 5 | 2 |
| | AgScad | 500 | 500 | 500 | 500 | 37 | 38 | 41 | 36 | 34 | 47 | 35 | 41 |
| | NGP | 500 | 500 | 500 | 499 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 2 |
| | Scad | 500 | 500 | 500 | 500 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | AgLasso | 500 | 500 | 500 | 500 | 14 | 24 | 16 | 23 | 21 | 26 | 23 | 22 |
| | AgBridge | 500 | 500 | 500 | 500 | 77 | 89 | 91 | 80 | 78 | 94 | 74 | 84 |
| | AgScad | 500 | 500 | 500 | 500 | 282 | 291 | 280 | 297 | 258 | 289 | 269 | 283 |
| | NGP | 500 | 500 | 500 | 496 | 2 | 4 | 2 | 2 | 2 | 4 | 1 | 5 |
| | Scad | 500 | 500 | 500 | 500 | 0 | 3 | 0 | 0 | 0 | 3 | 1 | 2 |

| $s$ | Method | $X^{(12)}$ | $X^{(13)}$ | $X^{(14)}$ | $X^{(15)}$ | $X^{(16)}$ | $X^{(17)}$ | $X^{(18)}$ | $X^{(19)}$ | $X^{(20)}$ | $X^{(21)}$ | $X^{(22)}$ | $X^{(23)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AgLasso | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| | AgBridge | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 1 |
| | AgScad | 18 | 17 | 8 | 9 | 9 | 6 | 11 | 15 | 11 | 4 | 8 | 14 |
| | NGP | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Scad | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.25 | AgLasso | 2 | 1 | 0 | 1 | 0 | 3 | 3 | 1 | 0 | 2 | 2 | 1 |
| | AgBridge | 2 | 8 | 5 | 1 | 3 | 4 | 8 | 5 | 3 | 1 | 3 | 6 |
| | AgScad | 46 | 52 | 32 | 41 | 46 | 48 | 38 | 47 | 44 | 39 | 47 | 50 |
| | NGP | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |
| | Scad | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | AgLasso | 23 | 25 | 16 | 22 | 18 | 28 | 20 | 24 | 18 | 25 | 22 | 29 |
| | AgBridge | 86 | 85 | 79 | 90 | 81 | 82 | 77 | 97 | 80 | 91 | 92 | 97 |
| | AgScad | 277 | 299 | 294 | 285 | 288 | 292 | 275 | 308 | 276 | 296 | 281 | 292 |
| | NGP | 4 | 1 | 2 | 5 | 0 | 4 | 1 | 6 | 5 | 1 | 0 | 1 |
| | Scad | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |

However, the highest accuracy is attained by Scad. When the SNR decreases, AgBridge and AgScad start including irrelevant variables, but for a high SNR ($=6.25$), all methods perform very well. This can also be concluded from Table 3, which presents the appearance frequency of the different variables in the 500 simulations. Table 4 presents the averages of the squared biases and the empirical variances of the estimates of the coefficient functions $\beta_p(t)$ (for $p = 0, \ldots, 3$), averaged over a grid of 300 equidistant points in $[1, 30]$. The numbers in brackets are the squared bias and empirical variances of the oracle estimates (i.e., the estimates obtained using a model with only the true relevant variables $X^{(0)}, \ldots, X^{(3)}$). Note that for all methods, the biases and empirical variances of the estimates are comparable with those for the oracle estimates. Further note that the variance of $\beta_0$ is remarkably larger than for the other components, because the range of this function is larger than that of the other coefficient functions. In general, the baseline coefficient has the highest bias, especially for AgBridge and AgScad.

Table 4. First simulation example: squared bias and variance for the estimated coefficient functions

| | | Squared bias (oracle) | | | | Variance (oracle) | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | Method | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| 1 | AgLasso | 0.0562 | 0.0009 | 0.0084 | 0.0025 | 31.4877 | 4.3208 | 2.8751 | 0.7203 |
| | | (0.0560) | (0.0009) | (0.0084) | (0.0025) | (31.5060) | (4.3238) | (2.8752) | (0.7206) |
| | AgBridge | 0.3311 | 0.0021 | 0.0331 | 0.0068 | 29.6735 | 4.1740 | 2.8864 | 0.6085 |
| | | (0.3352) | (0.0021) | (0.0330) | (0.0069) | (29.7601) | (4.1831) | (2.8869) | (0.6090) |
| | AgScad | 0.1711 | 0.0015 | 0.0127 | 0.0037 | 31.7366 | 4.4588 | 2.8849 | 0.7231 |
| | | (0.1570) | (0.0010) | (0.0119) | (0.0040) | (31.5323) | (4.4442) | (2.8826) | (0.7115) |
| | NGP | 0.0088 | 0.0029 | 0.0108 | 0.0117 | 33.9873 | 4.4701 | 2.8224 | 0.8489 |
| | | (0.0087) | (0.0054) | (0.0264) | (0.0343) | (33.9413) | (4.4639) | (2.8169) | (0.8475) |
| | Scad | 0.0108 | 0.0006 | 0.0002 | 0.0007 | 34.2092 | 4.6113 | 2.8315 | 0.8683 |
| | | (0.0108) | (0.0006) | (0.0002) | (0.0007) | (34.2092) | (4.6113) | (2.8315) | (0.8683) |
| 1.25 | AgLasso | 0.1856 | 0.0023 | 0.0255 | 0.0067 | 30.2397 | 4.2240 | 2.8926 | 0.6378 |
| | | (0.1847) | (0.0023) | (0.0255) | (0.0067) | (30.2626) | (4.2280) | (2.8926) | (0.6380) |
| | AgBridge | 0.5976 | 0.0052 | 0.0611 | 0.0119 | 28.4153 | 4.0569 | 2.8850 | 0.5426 |
| | | (0.6031) | (0.0052) | (0.0610) | (0.0121) | (28.5281) | (4.0617) | (2.8859) | (0.5428) |
| | AgScad | 0.2417 | 0.0022 | 0.0211 | 0.0056 | 30.9007 | 4.3899 | 2.8858 | 0.6913 |
| | | (0.2367) | (0.0019) | (0.0197) | (0.0066) | (30.8127) | (4.3998) | (2.8865) | (0.6773) |
| | NGP | 0.0109 | 0.0058 | 0.0262 | 0.0347 | 33.8607 | 4.4188 | 2.8144 | 0.8468 |
| | | (0.0108) | (0.0034) | (0.0111) | (0.0120) | (33.9114) | (4.4260) | (2.8199) | (0.8489) |
| | Scad | 0.0132 | 0.0009 | 0.0004 | 0.0011 | 34.2224 | 4.6136 | 2.8321 | 0.8721 |
| | | (0.0132) | (0.0009) | (0.0004) | (0.0011) | (34.2224) | (4.6136) | (2.8321) | (0.8721) |
| 2 | AgLasso | 0.1336 | 0.0032 | 0.0201 | 0.0072 | 30.7430 | 4.2744 | 2.8909 | 0.6752 |
| | | (0.1249) | (0.0032) | (0.0201) | (0.0072) | (30.7707) | (4.2784) | (2.8911) | (0.6758) |
| | AgBridge | 0.8222 | 0.0099 | 0.0853 | 0.0169 | 27.5664 | 3.9715 | 2.8848 | 0.5150 |
| | | (0.8393) | (0.0098) | (0.0852) | (0.0172) | (27.8247) | (3.9773) | (2.8876) | (0.5166) |
| | AgScad | 0.2806 | 0.0037 | 0.0261 | 0.0078 | 30.6066 | 4.3533 | 2.8902 | 0.6931 |
| | | (0.2873) | (0.0036) | (0.0245) | (0.0091) | (30.5670) | (4.3841) | (2.8913) | (0.6797) |
| | NGP | 0.0207 | 0.0151 | 0.0743 | 0.1044 | 33.5269 | 4.2834 | 2.7917 | 0.8446 |
| | | (0.0205) | (0.0126) | (0.0582) | (0.0818) | (33.5715) | (4.2893) | (2.7972) | (0.8467) |
| | Scad | 0.0238 | 0.0023 | 0.0010 | 0.0029 | 34.2882 | 4.6239 | 2.8354 | 0.8880 |
| | | (0.0237) | (0.0023) | (0.0010) | (0.0029) | (34.2877) | (4.6238) | (2.8354) | (0.8879) |

NOTE: The numbers in brackets are the results for the oracle estimates (i.e., the estimates when using a model with only the true variables).

Overall, AgLasso, NGP, and Scad perform the best in terms of estimation and variable selection properties. However, when taking the computing time into account, NGP clearly outperforms the other methods.

Graphs of the first six fitted coefficient functions for the simulated dataset with median ER are provided in Figure 2 for $s = 1.25$ (the graphs for $s = 1$ and $s = 2$ are similar and thus omitted here). The fitted coefficients are close to the true coefficients $\beta_p(t)$. Moreover, one can see that the P-splines fit for the zero components is nonzero, while the NGP fit for these components is zero. The mean response for this simulated dataset is presented in Figure 3, revealing a good performance of all methods.

### 5.2 SECOND SIMULATION EXAMPLE

In this section, we consider a model without irrelevant variables, but where some of the variables have a smaller influence. This will allow us to show that the proposed methods will not unnecessarily remove (possibly weak) components. The considered model is as

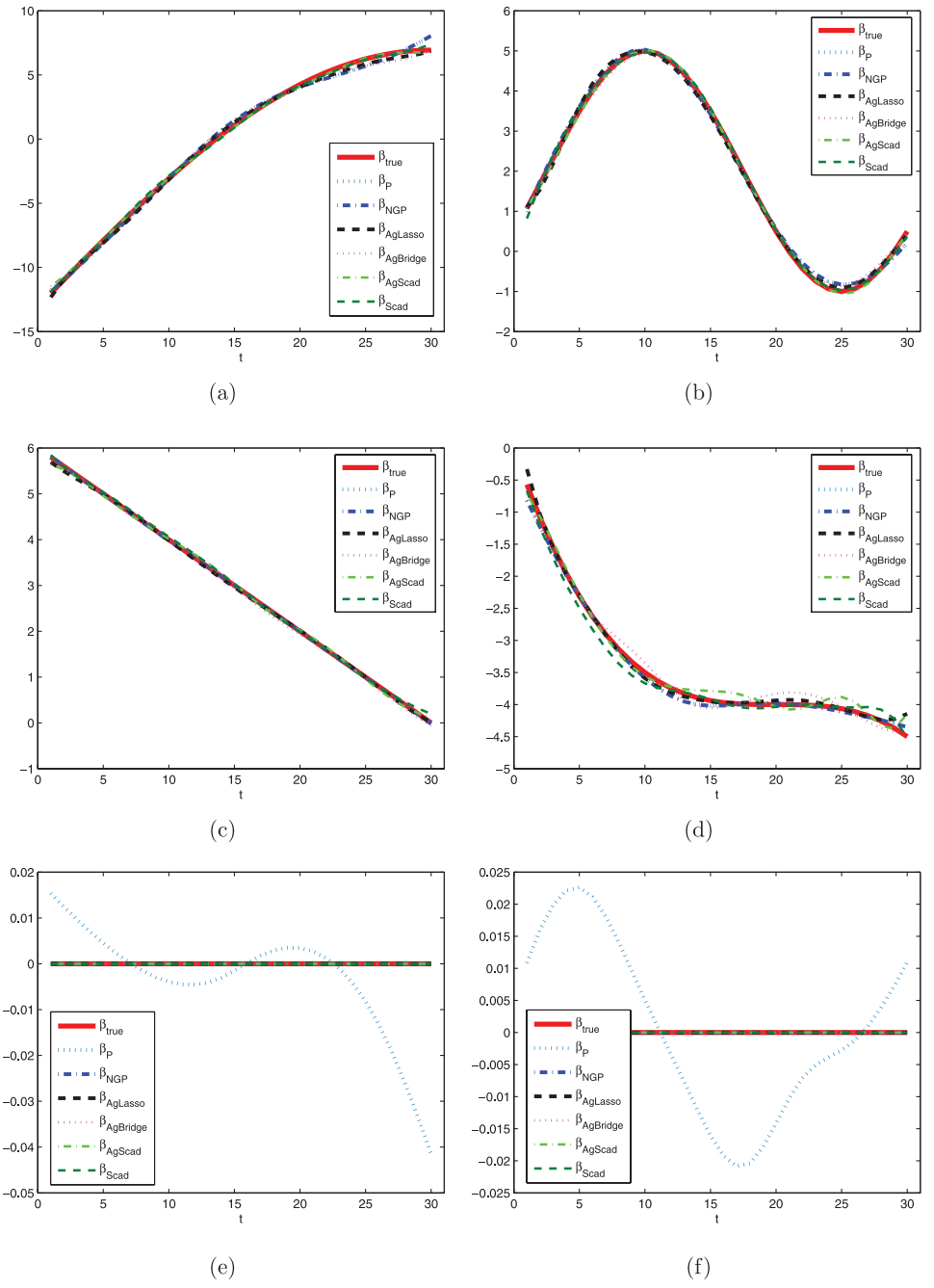



Figure 2. First simulation example: fitted coefficient for $s = 1.25$ of (a) $X^{(0)}(t)$, (b) $X^{(1)}(t)$, (c) $X^{(2)}(t)$, (d) $X^{(3)}(t)$, (e) $X^{(4)}(t)$, and (f) $X^{(5)}(t)$. Thick red solid line: true coefficient, thick light-blue dotted line: P-spline, thick dark-blue dashed-dotted line: NGP, thick black dashed line: AgLasso, thin purple dotted line: AgBridge, thin light-green dashed-dotted line: AgScad, and thin dark-green dashed line: Scad. The online version of this figure is in color.

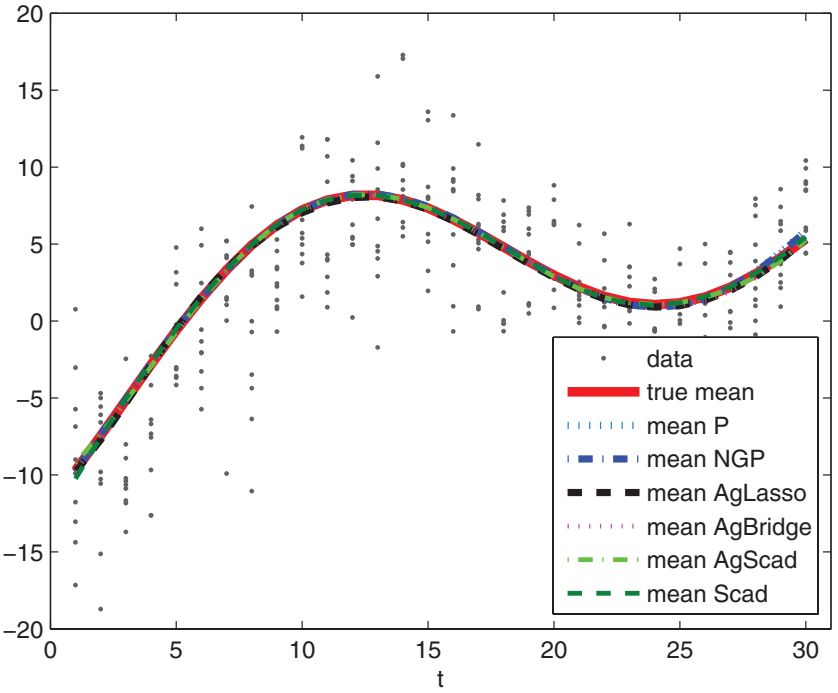



Figure 3. First simulation example: estimated mean for $s = 1.25$. Thick red solid line: true mean response, thick light-blue dotted line: P-spline, thick dark-blue dashed-dotted line: NGP, thick black dashed line: AgLasso, thin purple dotted line: AgBridge, thin light-green dashed-dotted line: AgScad, and thin dark-green dashed line: Scad. The online version of this figure is in color.

follows:

$$Y(t_{ij}) = \sum_{p=1}^{6} X^{(p)}(t_{ij})\beta_p(t_{ij}) + s\,\varepsilon(t_{ij}) \quad \text{for } i = 1, \ldots, 200,\ \ j = 1, \ldots, N_i,$$

where $X^{(1)}(t)$ is uniform distributed on $[t/10, 2 + t/10]$ for any $t$, $X^{(2)}(t)$, $X^{(3)}(t)$, $X^{(4)}(t)$, and $X^{(5)}(t)$ conditioning on $X^{(1)}(t)$ are independent and normal distributed with mean 0 and variance $\frac{1+X^{(1)}(t)}{2+X^{(1)}(t)}$, and $X^{(6)}(t)$ is normal distributed with mean $3e^{t/30}$ and variance 1. The error $\varepsilon(t)$ is normal distributed with mean 0 and variance 1 for any $t$. The parameter $s$ controls the SNR and is taken to be, respectively, 13, 16, and 27, resulting in the same SNRs as in the first simulation example (respectively, 6.25, 5, and 3). As in the first example, the observation time points $t_{ij}$ are the same for all subjects ($i = 1, \ldots, n$): $\{1, \ldots, 30\}$. The coefficient functions $\beta_p(t)$ of the variables are

$$\beta_1(t) = 15 + 20\sin\left(\frac{\pi t}{15}\right), \qquad \beta_2(t) = 15 + 20\cos\left(\frac{\pi t}{15}\right), \quad \beta_3(t) = 2 - 3\sin\left(\frac{\pi(t-25)}{15}\right),$$

$$\beta_4(t) = 2 - 3\cos\left(\frac{\pi(t-25)}{15}\right), \quad \beta_5(t) = 6 - 0.2t^2, \qquad \beta_6(t) = -4 + \frac{(20-t)^3}{2000}.$$

In this model, the influence of $X^{(3)}$, $X^{(4)}$, and $X^{(6)}$ is smaller than that of the other three variables; therefore, the deviation of the estimated coefficients to the true coefficients for these less influential variables is higher, as can be seen in Figure 4 for $s = 16$ (the fitted

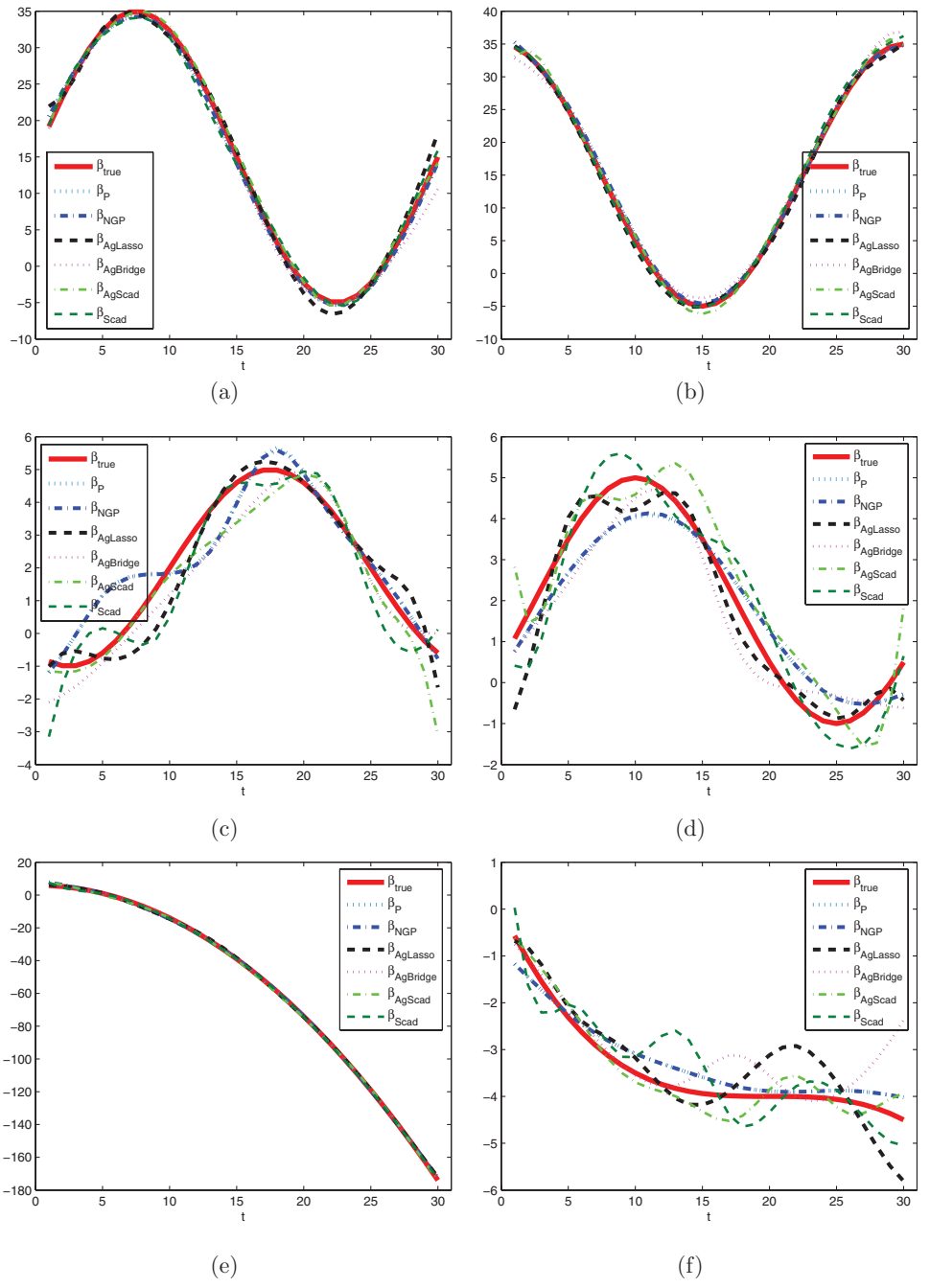



Figure 4. Second simulation example: fitted coefficient for $s = 16$ of (a) $X^{(1)}(t)$, (b) $X^{(2)}(t)$, (c) $X^{(3)}(t)$, (d) $X^{(4)}(t)$, (e) $X^{(5)}(t)$, and (f) $X^{(6)}(t)$. Thick red solid line: true coefficient, thick light-blue dotted line: P-spline, thick dark-blue dashed-dotted line: NGP, thick black dashed line: AgLasso, thin purple dotted line: AgBridge, thin light-green dashed-dotted line: AgScad, and thin dark-green dashed line: Scad. The online version of this figure is in color.

Table 5. Second simulation example: appearance frequency of the variables

| $s$ | Method | $X^{(1)}$ | $X^{(2)}$ | $X^{(3)}$ | $X^{(4)}$ | $X^{(5)}$ | $X^{(6)}$ |
|---|---|---|---|---|---|---|---|
| 13 | AgLasso | 500 | 500 | 500 | 500 | 500 | 500 |
| | AgBridge | 500 | 500 | 500 | 500 | 500 | 500 |
| | AgScad | 500 | 500 | 500 | 500 | 500 | 500 |
| | NGP | 500 | 500 | 500 | 499 | 500 | 500 |
| | Scad | 500 | 500 | 500 | 500 | 500 | 500 |
| 16 | AgLasso | 500 | 500 | 500 | 500 | 500 | 500 |
| | AgBridge | 500 | 500 | 500 | 500 | 500 | 500 |
| | AgScad | 500 | 500 | 500 | 500 | 500 | 500 |
| | NGP | 500 | 500 | 499 | 497 | 500 | 500 |
| | Scad | 500 | 500 | 500 | 500 | 500 | 500 |
| 27 | AgLasso | 500 | 500 | 500 | 500 | 500 | 500 |
| | AgBridge | 500 | 500 | 500 | 500 | 500 | 500 |
| | AgScad | 500 | 500 | 500 | 500 | 500 | 500 |
| | NGP | 500 | 500 | 489 | 472 | 500 | 500 |
| | Scad | 500 | 500 | 500 | 500 | 500 | 500 |

coefficients for $s = 13$ and 27 are comparable and thus omitted). The estimated mean response for all methods is very close to the true mean response; see Figure 5.

From Tables 5 and 6, we see that when the SNR increases, NGP eliminates, in a very few cases (less than 8%), covariates $X^{(3)}$ and $X^{(4)}$. The APSO-based methods and Scad

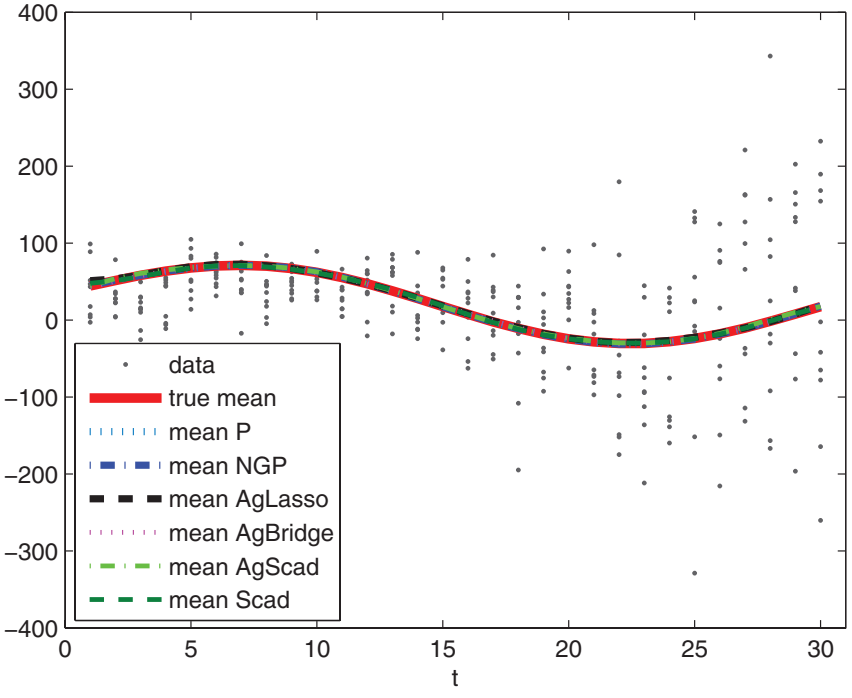



Figure 5. Second simulation example: estimated mean for $s = 16$. Thick red solid line: true mean response, thick light-blue dotted line: P-spline, thick dark-blue dashed-dotted line: NGP, thick black dashed line: AgLasso, thin purple dotted line: AgBridge, thin light-green dashed-dotted line: AgScad, and thin dark-green dashed line: Scad. The online version of this figure is in color.

Table 6. Simulation results for the second simulation example

| *s* | Method<br>Optimal value | ER<br>0 | MS<br>6 | MFZ<br>0 | MTP<br>6 | PercT<br>1 | AverR<br>6 | Time<br>0 |
|---|---|---|---|---|---|---|---|---|
| 13 | AgLasso | 2.5902 | 6 | 0 | 6 | 1 | 6 | 23.0686 |
| | | (1.3637) | (6,6) | (0,0) | (6,6) | | (0) | (9.9094) |
| | AgBridge | 3.2050 | 6 | 0 | 6 | 1 | 6 | 12.3299 |
| | | (3.3136) | (6,6) | (0,0) | (6,6) | | (0) | (2.2272) |
| | AgScad | 3.4222 | 6 | 0 | 6 | 1 | 6 | 7.2261 |
| | | (2.9105) | (6,6) | (0,0) | (6,6) | | (0) | (1.4687) |
| | NGP | 1.5345 | 6 | 0 | 6 | 0.998 | 5.9980 | 1.0342 |
| | | (1.1370) | (6,6) | (0,0) | (6,6) | | (0.0447) | (0.2449) |
| | Scad | 1.6956 | 6 | 0 | 6 | 1 | 6 | 166.1929 |
| | | (0.3146) | (6,6) | (0,0) | (6,6) | | (0) | (29.5009) |
| 16 | AgLasso | 3.9241 | 6 | 0 | 6 | 1 | 6 | 31.4087 |
| | | (2.0651) | (6,6) | (0,0) | (6,6) | | (0) | (13.5418) |
| | AgBridge | 4.7792 | 6 | 0 | 6 | 1 | 5.9760 | 12.8088 |
| | | (4.9705) | (6,6) | (0,0) | (6,6) | | (0) | (2.2971) |
| | AgScad | 5.0040 | 6 | 0 | 6 | 1 | 6 | 7.4198 |
| | | (4.5669) | (6,6) | (0,0) | (6,6) | | (0) | (1.4140) |
| | NGP | 2.2548 | 6 | 0 | 6 | 0.992 | 5.9920 | 1.0306 |
| | | (1.6330) | (6,6) | (0,0) | (6,6) | | (0.0892) | (0.2801) |
| | Scad | 2.5681 | 6 | 0 | 6 | 1 | 6 | 170.9046 |
| | | (0.4765) | (6,6) | (0,0) | (6,6) | | (0) | (29.5623) |
| 27 | AgLasso | 11.1857 | 6 | 0 | 6 | 1 | 6 | 37.3271 |
| | | (5.8929) | (6,6) | (0,0) | (6,6) | | (0) | (16.3712) |
| | AgBridge | 13.3599 | 6 | 0 | 6 | 1 | 6 | 13.1288 |
| | | (14.2027) | (6,6) | (0,0) | (6,6) | | (0) | (2.3198) |
| | AgScad | 13.7258 | 6 | 0 | 6 | 1 | 6 | 7.9092 |
| | | (13.0172) | (6,6) | (0,0) | (6,6) | | (0) | (1.6097) |
| | NGP | 5.8285 | 6 | 0 | 6 | 0.922 | 5.9220 | 1.7590 |
| | | (3.3040) | (6,6) | (0,0) | (6,6) | | (0.2684) | (4.5366) |
| | Scad | 7.3117 | 6 | 0 | 6 | 1 | 6 | 173.0339 |
| | | (1.3569) | (6,6) | (0,0) | (6,6) | | (0) | (27.2765) |

NOTE: The numbers in the brackets are the standard errors (for the continuous-valued criteria) or the first and third quartiles (for the discrete-valued criteria).



never remove a relevant variable, while the nonnegative garrote method always has the lowest estimation error and computing time. As in the first simulated example, Scad has a remarkably higher computing time than the other methods, without resulting in the lowest estimation error. Of all APSO-based methods, AgLasso seems to be the best one in terms of estimation error, but with a slightly higher computing time.

### 5.3 REAL-DATA EXAMPLES

*5.3.1 AIDS Data Example.* The data are a subset from the multicenter AIDS Cohort Study (MACS) and contain repeated measurements of physical examinations, laboratory results, CD4 cell counts, and CD4 percentages of homosexual men who became HIV-positive between 1984 and 1991. All individuals were scheduled to have their measurements done twice a year, but due to missed visits and the fact that HIV infections occurred randomly during the study, there were unequal numbers of repeated measurements and

different measurement times for each individual. Details of the design, methods, and medical implications of the MACS can be found in Kaslow et al. (1987).

We consider the 283 homosexual men who became HIV-positive and try to evaluate the effects of cigarette smoking, preinfection CD4 cell percentage, and age at HIV infection on the mean CD4 percentage after infection. The number of repeated measurements ranged from 1 to 14, with a median of 6 and a mean of 6.57. The number of distinct time points was 59. Denote by $t_{ij}$ the time in years of the $j$th measurement on the $i$th individual after HIV infection, and by $Y_{ij}$, the $i$th individual's CD4 percentage at time $t_{ij}$.

The covariates are $X_i^{(1)}$, the smoking status of the $i$th individual (1 or 0 if the individual ever or never smoked cigarettes); $X_i^{(2)}$, the centered age at HIV infection for the $i$th individual; and $X_i^{(3)}$, the centered preinfection CD4 percentage. The varying-coefficient model for $Y_{ij}$ we consider is

$$Y_{ij} = \beta_0(t_{ij}) + \sum_{p=1}^{3} X_i^{(p)} \beta_p(t_{ij}) + \varepsilon_{ij},$$

where $\beta_0(t)$, the baseline CD4 percentage, represents the mean CD4 percentage $t$ years after the HIV infection for a nonsmoker, with average pre-infection CD4 percentage 42.6841% and average age at infection 34.3642 years. The time-varying effects of cigarette smoking, age at HIV infection, and preinfection CD4 percentage on the postinfection CD4 percentage at time $t$ are given by $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$, respectively.

We apply nonnegative garrote with P-splines, AgLasso, AgBridge, and Scad on this dataset. We do not include the results of AgScad, since already from the first simulation example, it is clear that this method does not give good results. The `gMCP` option with the Scad parameter ($a = 3.7$) of the `grpreg` function in R is not stable enough and does not give reasonable results.

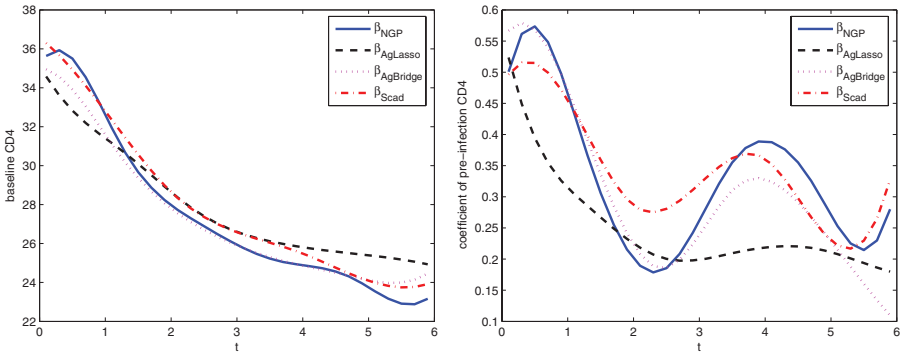
For the P-spline-based methods, we use B-splines of degree 3, with $K = 5$ and differencing order 2 for all components. The Scad method of Wang, Li, and Huang (2008) selected $K = 3$. The regularization parameters are chosen via the same procedures as in the simulation study. In Table 7, the residual sum of squares (RSS), that is,

$$\text{RSS} = ||\mathbf{Y} - \hat{\mathbf{Y}}||_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left( Y_{ij} - \sum_{p=0}^{d} X_{ij}^{(p)} \widehat{\beta}_p(t_{ij}) \right)^2,$$

and the number of selected components (NS) are given.

Table 7. AIDS data: summary parameters

| Method | NS | RSS | RSS/$N$ |
|---|---|---|---|
| NGP | 2 | 192197 | 105.777 |
| AgLasso | 2 | 194931 | 107.282 |
| AgBridge | 2 | 192674 | 106.040 |
| Scad | 2 | 191396 | 105.336 |




(a) (b)

Figure 6. AIDS data: fitted (a) baseline effect and (b) coefficient of preinfection CD4. Dark-blue solid line: NGP, black dashed line: AgLasso, purple dotted line: AgBridge, and red dashed-dotted line: Scad. The online version of this figure is in color.

All methods select the baseline CD4 percentage and the preinfection CD4 percentage as the important components and have a comparable RSS. The estimated regression coefficients are presented in Figure 6, and a graph of the evolution of the CD4 percentage over time for three different initial values of the preinfection CD4 percentage is given in Figure 7. This figure allows to see how the CD4 percentage evolves over time for

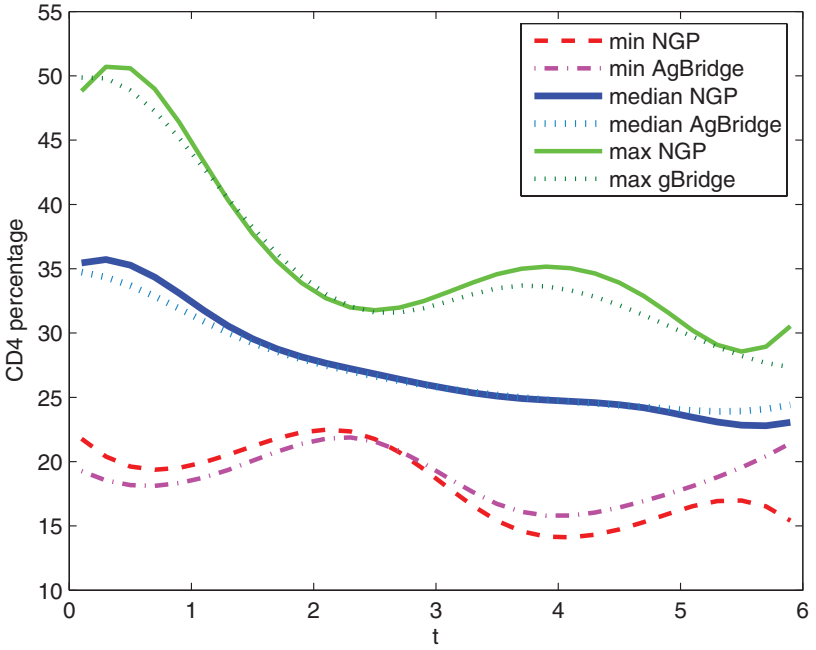



Figure 7. AIDS data: fitted CD4 percentage for person with minimum (−27.6841)-, median (−0.3841)-, and maximum (26.3159)-centered preinfection CD4 percentage. Thin red dashed line: NGP, thin purple dashed-dotted line: AgBridge for person with minimum-centered preinfection CD4 percentage; thick dark blue solid line: NGP, thick light-blue dotted line: AgBridge for person with median-centered preinfection CD4 percentage; thin light-green solid line: NGP, thin dark-green dotted line: AgBridge for person with maximum-centered preinfection CD4 percentage. The online version of this figure is in color.

three patients with different initial conditions distinguished via their centered preinfection CD4 percentage. For example, a patient with a centered preinfection CD4 percentage of −0.3841 is estimated to have a CD4 percentage of about 25 after 6 years, while for a patient with a centered preinfection CD4 of 26.3159, this is estimated to be about 30%.

*5.3.2 KTB Data Example.* These are data from the German Continental Deep Drilling Program (Kontinentales Tiefbohrprogramm der Bundesrepublik Deutschland, KTB), containing measurements on the upper 10 km of the continental crust. Our data analysis involves 22 variables, and the "depth" into the crust plays the role of "time." The application of the discussed variable selection methods to this dataset and the findings from these are given in the supplementary materials.

## 6. CONCLUSION

From the simulation studies, it is clear that the nonnegative garrote method with P-splines and the APSO method (i.e., AgLasso) perform very well. It is to be mentioned that the former method is superior to all others in terms of computational costs. The fact that AgBridge and, especially, AgScad perform worse than the other methods in the first example (in the sense that they include more irrelevant variables) is in our opinion rather due to an instability in the `grpreg` procedure of `R`, than due to the method itself.

The Scad procedure of Wang, Li, and Huang (2008) performs a bit better than the other methods, in the sense that it adds less irrelevant variables, but at a cost of a much higher computing time. This extra computing time is in our opinion not worth the effort, since the other methods also perform well and in a much lower time.

In the simulation study, we have taken the number of knots ($K$), the order of the difference operator ($k$), and the degree of the P-splines ($q$) as fixed and given. In fact, there is a strong interplay between these parameters, as was revealed in detail by the study in Gijbels and Verhasselt (2010). One could choose the number of knot points in a data-driven way, but our experiences show that it is somewhat redundant to aim at choosing all parameters in a data-driven way, due to their interrelationships.

## SUPPLEMENTARY MATERIALS

**AIDS data example:** Computer codes for running this example for all methods discussed.

**KTB data example:** The variable selection methods are applied to these data, providing information on the influential variables (out of the 22 variables) and on the kind of impact they have.

**Appendix:** Contains the proofs of all theoretical results presented in Sections 2 and 3.

## ACKNOWLEDGMENTS

The authors are grateful to Professor Brian Marx for providing the KTB data, and to the authors of Wang, Li, and Huang (2008) for providing their software code. The authors thank the editor, the associate editor, and two reviewers for their detailed reading of the article and their valuable comments, which led to a considerable improvement of the article. Part of the research was carried out while the third author was affiliated at the Katholieke Universiteit Leuven. This research is supported by the Belgian IAP (Interuniversity Attraction Pole)

research network P6/03 (Belgian Science Policy) and the research fund of the KU Leuven (GOA/07/04-project). Funding by the Flemish Science Foundation (FWO-project G.0328.08N) is also gratefully acknowledged.

*[Received September 2011. Revised December 2010.]*

## REFERENCES

Antoniadis, A., Gijbels, I., and Verhasselt, A. (2012), "Variable Selection in Additive Models Using P-Splines," *Technometrics* (tentatively accepted). [639,647]

Bach, F. R. (2008), "Consistency of the Group Lasso and Multiple Kernel Learning," *Journal of Machine Learning Research*, 9, 1179–1225. [639,648]

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 51, 373–384. [639,645]

Eilers, P., and Marx, B. (1996), "Flexible Smoothing With B-Splines and Penalties," *Statistical Science*, 11, 89–102. [639]

Fan, J., and Li, R. (2001), "Variable selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [639,649]

GeoForschungsZentrum . "Kontinentales Tiefbohrprogramm der Bundesrepublik Deutschland [German Continental Deep Drilling Program]," Available at *http://www.icdp-online.org/sites/ktb/welcome.html* [xxxx]

Gijbels, I., and Verhasselt, A. (2010), "P-Splines Regression Smoothing and Difference Type of Penalty," *Statistics and Computing*, 20, 499–511. [660]

Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society*, Series B, 55, 757–796. [638]

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L-P (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809–822. [639]

Huang, J., Horowitz, J., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [649]

Huang, J. Z., Wu, C. O., and Zhou, L. (2004), "Polynomial Spline Estimation and Inference for Varying Coefficient Models With Longitudinal Data," *Statistica Sinica*, 14, 763–788. [639,642,643]

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo C. R. (1987), "The Multicenter AIDS Cohort Study: Rationale, Organization and Selected Characteristics of the Participants," *American Journal of Epidemiology*, 126, 310–318. [658]

Lin, B. Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 32, 2272–2297. [648]

Lu, Y., Zhang, R., and Zhu, L. (2008), "Penalized Spline Estimation for Varying Coefficient Models," *Communications in Statistics—Theory and Methods*, 37, 2249–2261. [639,640,642]

Marx, B. (2010), "P-Spline Varying Coefficient Models for Complex Data," in *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, eds. T. Kneiband G. Tutz, New York: Springer, pp. 19–44. [649]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1348–1360. [644]

Wang, L., Li, H., and Huang, J. Z. (2008), "Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, 103, 1556–1569. [639,643,648,649,658,660]

Yuan, M. (2007), "Nonnegative Garrote Component Selection in Functional ANOVA Models," in Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (Vol. 2), pp. 660–666. [639,647]

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. [639,648]