Statistica Sinica **24** (2014), 147-171
doi:http://dx.doi.org/10.5705/ss.2011.168

GENERALIZED VARYING COEFFICIENT MODELS: A SMOOTH VARIABLE SELECTION TECHNIQUE

Anneleen Verhasselt

Universiteit Hasselt

Abstract: We consider nonparametric smoothing and variable selection in generalized varying coefficient models. Generalized varying coefficient models are commonly used for analyzing the time-dependent effects of covariates on responses that are not necessary continuous, for example counts or categories. 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), by using an approximation of the link function. The combination of P-splines with nonnegative garrote (which is a variable selection method) leads to good estimation and variable selection. The method is illustrated with a simulation study and a data example.

Key words and phrases: generalized varying coefficient models, longitudinal data, nonparametric smoothing, P-splines, variable selection.

1. Introduction

Varying coefficient models (Hastie and Tibshirani (1993)) are an extension of classical linear regression models. They allow the regression coefficients to vary in a smooth way with another variable (for example time). We study varying coefficient models where the response, covariates, and regression coefficients are allowed to vary with t:

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

where $Y(t)$ is the response at time t ($\in \mathcal{T} = [0, T]$), $\mathbf{X}(t) = (X^{(0)}(t), \ldots, X^{(d)}(t))'$ the covariate vector at time t with $X^{(0)}(t) \equiv 1$, $\boldsymbol{\beta}(t) = (\beta_0(t), \ldots, \beta_d(t))'$ the vector of coefficients at time t, $\beta_0(t)$ is the baseline effect and $\varepsilon(t)$ a mean zero stochastic process at time t. These models are especially useful for modeling longitudinal data, they are flexible and easy to interpret. They have been used to discover dynamic patters in data in several scientific areas. See Fan and Zhang (2008) (and references therein) for an overview.

We consider samples of n independent subjects or individuals each measured repeatedly over a time period. Let $(t_{ij}, Y_{ij}, \mathbf{X}_{ij})$ be the jth measurement for subject i of $(t, Y(t), \mathbf{X}(t))$, 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 $\mathbf{X}_{ij} = (X_{ij}^{(0)}, \ldots, X_{ij}^{(d)})'$. The total number of observations is denoted by $N = \sum_{i=1}^{n} N_i$.

In some applications data are not continuous, for example counts or categories. Allowing for this type of data one should extend varying coefficient models in the same way as normal models were extended to generalized linear models (McCullagh and Nelder (1995)). A generalized varying coefficient model framework (see for example Cai, Fan, and Li (2000) and Şentürk and Müller (2008)) takes

$$\eta(\mathbf{X}(t)) = \beta_0(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_d(t)X^{(d)}(t) = \sum_{p=0}^{d} X^{(p)}(t)\beta_p(t),$$

where $\eta(\mathbf{X}(t))$ is linked to the mean, $\mu(\mathbf{X}(t)) = \mathrm{E}\,(Y(t)|\mathbf{X}(t))$, through the link function $g(\cdot)$: $\eta(\mathbf{X}(t)) = g(\mu(\mathbf{X}(t)))$. The density function of the random variable Y at time t is to belong to the exponential family:

$$f(Y;\theta,\phi) = \exp\Big(\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\Big), \tag{1.2}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions, ϕ is a scale parameter, and θ is the canonical parameter. In this context, Y and θ depend on $\mathbf{X}(t)$. Furthermore $\mathrm{E}\,(Y(\mathbf{X}(t))) = \mu(\mathbf{X}(t)) = \frac{db}{d\theta}\big|_{\theta(\mathbf{X}(t))}$ and $\mathrm{Var}\,(Y(\mathbf{X}(t))) = \frac{d^2b}{d\theta^2}\big|_{\theta(\mathbf{X}(t))}a(\phi)$. We assume that $a(\phi)$ is bounded. A link function $g(\cdot)$ is called a canonical link if $\theta(\mathbf{X}(t)) = g(\mu(\mathbf{X}(t)))$.

Consider an example that illustrates the kind of situations we consider. The data come from a study on the short term effect of concentrations of pollutants in ambient air on hospital admission for cardiovascular and respiratory diseases in Hong Kong. In 1994 and 1995 the daily concentration (in $\mu g/m^3$) of nitrogen dioxide (NO_2), sulphur dioxide (SO_2), particulates $< 10\mu m$ in aerodynamic diameter (i.e. dust) and ozone (O_3), temperature (in Celsius), humidity (in percentage), and hospital admissions for cardiovascular and respiratory diseases were measured in Hong Kong. The conditional distribution of the number of hospital admissions given the level of pollutants, temperature, and humidity can be modeled as a Poisson distribution. The mean $\mu(\mathbf{X})$ of the hospital admissions is linked to the linear predictor $\eta(\mathbf{X})$ with the canonical log-link: $\eta(\mathbf{X}) = \log(\mu(\mathbf{X}))$ (see also Table 1 in Section 4).

We study the problem of smoothing and variable selection in this generalized varying coefficient model setup; we want to estimate the regression coefficients $\beta_p(t)$ nonparametrically and select the relevant ones.

In the context of varying coefficient models, several nonparametric smoothing techniques, such as local polynomials (Fan and Zhang (1999, 2008)), regression

splines (Huang, Wu, and Zhou (2004)) and P-splines (Antoniadis, Gijbels, and Verhasselt (2012b)), have been proposed for estimating the coefficient functions $\boldsymbol{\beta}(t)$. Local polynomials have been used frequently for estimation in generalized varying coefficient models (see for example Fan and Zhang (2008), Zhang (2011), and Zhang and Peng (2010)). P-splines have been considered in generalized varying coefficient models by Eilers and Marx (2002) and Marx (2010) without any theoretical foundation. This paper gives the theoretical foundation, proving the estimation consistency of P-splines in generalized varying coefficient models. This result extends the consistency result of Antoniadis, Gijbels, and Verhasselt (2012b) for varying coefficient models to the broader class of generalized varying coefficient models.

We combine P-splines with the nonnegative garrote variable selection technique for estimating and selecting the relevant coefficient functions $\beta_p(t)$. As such, this generalizes the results of Antoniadis, Gijbels, and Verhasselt (2012b) to this generalized setup. The nonnegative garrote was originally proposed for variable selection in linear regression models by Breiman (1995), but it has been used in additive models by Antoniadis, Gijbels, and Verhasselt (2012a), Cantoni, Flemming, and Ronchetti (2000), and Yuan (2007), in generalized additive models by Marra and Wood (2011), and in varying coefficient models by Antoniadis, Gijbels, and Verhasselt (2012b). The estimation and variable selection consistency is proved for the nonnegative garrote, combined with P-splines in this generalized varying coefficient model framework. As such we give theoretical support for the P-spline estimation and nonnegative variable selection techniques in the broad class of generalized varying coefficient models. Key ingredients for the consistency are recent results in Antoniadis, Gijbels, and Verhasselt (2012b) for varying coefficient models (1.1), and the idea of Gijbels and Verhasselt (2010) for approximating $g^{-1}(\cdot)$. By using an approximation of $g^{-1}(\cdot)$, we approximate the optimization problem in the generalized context to a problem similar to the normal context.

The paper is organized as follows. In Section 2 we introduce P-splines in the generalized varying coefficient model context and show their consistency. The nonnegative garrote and its variable selection consistency are discussed in Section 3. We evaluate the performance of the method in Section 4 with simulations and on data. The details of the proofs are deferred to the Appendix.

2. P-spline Estimator

P-splines were first introduced by Eilers and Marx (1996) in the univariate nonparametric smoothing context. Since regular regression with B-splines tends to overfit, they proposed to add a difference penalty on the coefficients of adjacent B-splines, in the same sense as smoothing splines. This leads to the regression

P-splines technique. P-splines have been used as a tool in many different areas, see for example Ruppert, Wand, and Carroll (2003).

The extension of P-splines to generalized varying coefficient models is the following: suppose that for each $p = 0, \ldots, d$, $\beta_p(t)$ can be approximated by a B-spline basis expansion $\beta_p(t) = \sum_{l=1}^{m_p} B_{pl}(t; q_p)\alpha_{pl}$ and

$$\eta(\mathbf{X}(t)) = \sum_{p=0}^{d} X^{(p)}(t) \sum_{l=1}^{m_p} B_{pl}(t; q_p)\alpha_{pl},$$

where $\{B_{pl}(\cdot; q_p) : l = 1, \ldots, K_p + q_p = m_p\}$ is the q_pth degree B-spline basis with $K_p + 1$ equidistant knots $\xi_{p0}, \ldots \xi_{pK_p}$ for the pth 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. Let $m_{\max} = \max_{0 \leqslant p \leqslant d} m_p$, the maximal size of the B-spline basis of the various components.

We use the canonical link

$$\theta(\mathbf{X}(t)) = \eta(\mathbf{X}(t)) = \sum_{p=0}^{d} X^{(p)}(t) \sum_{l=1}^{m_p} B_{pl}(t; q_p)\alpha_{pl}.$$

We then obtain the P-spline estimates of the regression coefficients α_{pl} by minimizing $S_1(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \ldots, \boldsymbol{\alpha}_d')' \in I\!R^{\dim \times 1}$, where $\boldsymbol{\alpha}_p = (\alpha_{p1}, \ldots, \alpha_{pm_p})'$ and $\dim = \sum_p m_p$:

$$\begin{aligned} S_1(\boldsymbol{\alpha}) &= -2 \log(L(\boldsymbol{\alpha}; (t_{ij}, Y_{ij}, \mathbf{X}_{ij}), j = 1, \ldots, N_i, i = 1, \ldots, n)) \\ &\quad + \sum_{p=0}^{d} \lambda_p \boldsymbol{\alpha}_p' \mathbf{D}_{k_p}' \mathbf{D}_{k_p} \boldsymbol{\alpha}_p \\ &= -2 \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \Big(\frac{Y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(Y_{ij}; \phi) \Big) + \sum_{p=0}^{d} \lambda_p \boldsymbol{\alpha}_p' \mathbf{D}_{k_p}' \mathbf{D}_{k_p} \boldsymbol{\alpha}_p, \end{aligned}$$

with $L(\boldsymbol{\alpha}; (t_{ij}, Y_{ij}, \mathbf{X}_{ij}), j = 1, \ldots, N_i, i = 1, \ldots, n)$ the likelihood function derived from (1.2), $\mathbf{D}_{k_p} \in I\!R^{(m_p - k_p) \times m_p}$ is the matrix representation of the k_pth order differencing operator Δ_{k_p}, $\lambda_p > 0$ (for $p = 0, \ldots, d$) the smoothing parameters, and $\theta_{ij} = \theta(\mathbf{X}_{ij})$.

This optimization problem is equivalent to minimizing $S_2(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$:

$$\begin{aligned} S_2(\boldsymbol{\alpha}) &= -2 \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \Big(Y_{ij}\theta_{ij} - b(\theta_{ij}) \Big) + a(\phi) \sum_{p=0}^{d} \lambda_p \boldsymbol{\alpha}_p' \mathbf{D}_{k_p}' \mathbf{D}_{k_p} \boldsymbol{\alpha}_p \\ &= -2 \sum_{i=1}^{n} \frac{1}{N_i} \Big(\mathbf{Y}_i' \mathbf{U}_i \boldsymbol{\alpha} - \mathbf{1}_{N_i}' b(\mathbf{U}_i \boldsymbol{\alpha}) \Big) + a(\phi) \boldsymbol{\alpha}' \mathbf{Q}_\lambda \boldsymbol{\alpha} \end{aligned}$$

$$= -2\Big(\mathbf{Y}'\mathbf{W}\mathbf{U}\boldsymbol{\alpha} - \mathbf{1}_N'\mathbf{W}b(\mathbf{U}\boldsymbol{\alpha})\Big) + a(\phi)\boldsymbol{\alpha}'\mathbf{Q}_\lambda\boldsymbol{\alpha},$$

where $\mathbf{1}_{N_i} = (1,\ldots,1)' \in I\!R^{N_i\times 1}$, $\mathbf{Y} = (\mathbf{Y}_1',\ldots,\mathbf{Y}_n')'$, $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{iN_i})'$, $\mathbf{U} = (\mathbf{U}_1',\ldots,\mathbf{U}_n')'$, $\mathbf{U}_i = (\mathbf{U}_{i1}',\ldots,\mathbf{U}_{iN_i}')'$, $\mathbf{U}_{ij}' = \mathbf{X}_{ij}'\mathbf{B}(t_{ij})$, $\mathbf{W} = \text{diag}(\mathbf{W}_1,\ldots,\mathbf{W}_n)$, $\mathbf{W}_i = \text{diag}\left(N_i^{-1},\ldots,N_i^{-1}\right) \in I\!R^{N_i\times N_i}$ a diagonal matrix with N_i on the diagonal, $\mathbf{Q}_\lambda = \text{diag}(\lambda_0\mathbf{D}_{k_0}'\mathbf{D}_{k_0},\ldots,\lambda_d\mathbf{D}_{k_d}'\mathbf{D}_{k_d})$ a block diagonal matrix with the matrices $\lambda_p\mathbf{D}_{k_p}'\mathbf{D}_{k_p}$ on the diagonal, and

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

Minimizing $S_2(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ leads to the system of equations

$$\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\mu}) = a(\phi)\mathbf{Q}_\lambda\boldsymbol{\alpha}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1',\ldots,\boldsymbol{\mu}_n')' \in I\!R^{N\times 1}$, $\boldsymbol{\mu}_i = (\mu_{i1},\ldots,\mu_{iN_i})' \in I\!R^{N_i\times 1}$, and $\mu_{ij} = g^{-1}(\eta(\mathbf{X}_{ij})) = g^{-1}(\theta_{ij})$. This last equation is difficult to solve, since $\boldsymbol{\alpha}$ appears in the left hand side in a nonlinear fashion (through $\boldsymbol{\mu}$). Therefore we use the idea of Gijbels and Verhasselt (2010) to approximate the mean in a linear way.

A first-order approximation of $g^{-1}(\theta)$ for small θ is

$$g^{-1}(\theta(\mathbf{X}(t))) \approx g^{-1}(0) + \frac{dg^{-1}}{d\theta}\Big|_0 \theta(\mathbf{X}(t)) = \zeta + \tau\theta(\mathbf{X}(t)), \tag{2.2}$$

where $0 < |\tau| < \infty$ and $|\zeta| < \infty$, if a Taylor series for $g^{-1}(\cdot)$ around 0 exists. This results in a first order approximation for $\boldsymbol{\mu}$,

$$\boldsymbol{\mu} \approx \boldsymbol{\zeta} + \tau\mathbf{U}\alpha,$$

where $\boldsymbol{\zeta} = \zeta\mathbf{1}_N$. The approximation (2.2) is valid if $g^{-1}(\cdot)$ is continuous differentiable and the second derivative of g^{-1}: $\frac{d^2g^{-1}}{d\theta^2}$ exists and if the remainder term $\frac{d^2g^{-1}}{d\theta^2}\Big|_\xi (\theta(\mathbf{X}(t)))^2$ (with $|\xi| < |\theta(\mathbf{X}(t))|$) is small with respect to $\frac{dg^{-1}}{d\theta}\Big|_0 \theta(\mathbf{X}(t))$.

Note that we could use a Taylor series for $g^{-1}(\theta(\mathbf{X}(t)))$ around a constant $c(t)$, allowing for a different approximation of the mean at different time points, but we restrict our attention to (2.2).

Using (2.2), (2.1) can be approximated by

$$\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta}) = (\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda)\boldsymbol{\alpha}. \tag{2.3}$$

Note that if $\lambda_0 = \ldots = \lambda_d = 0$, this system of equations corresponds to the system of equations in (normal) varying coefficient models (1.1), with response $\tau^{-1}(Y(t) - \zeta)$ (see Antoniadis, Gijbels, and Verhasselt (2012b)).

If $\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda$ is invertible, then (2.3) has a unique solution

$$\widehat{\boldsymbol{\alpha}} = (\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda)^{-1}\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta}), \tag{2.4}$$

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 $\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda$ is invertible.

Lemma 1. *The matrix $\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda$ 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 is deferred the Appendix. From the proof we have an approximation for $(\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda)^{-1}$ and $\widehat{\boldsymbol{\alpha}}$ if $m_{\max}^{3/2}\lambda_{\max}n^{-1} \to 0$ and $m_{\max}n^{-1} \to$ constant:

$$\begin{aligned}
(\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda)^{-1} &= \tau^{-1}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} - a(\phi)\tau^{-2}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} \\
&\quad + o_P\Big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\Big)\,\tau^{-1}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}, \\
\widehat{\boldsymbol{\alpha}} &= \tau^{-1}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta}) - a(\phi)\tau^{-2}\Big((\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} \\
&\quad - o_P\Big(\frac{m_{\max}^{5/2}\lambda_{\max}}{n^2}\Big)\mathbf{1}_{\dim\times\dim}\Big)\cdot\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta}) \\
&= \widehat{\boldsymbol{\alpha}}_{\text{reg}} - \Big(a(\phi)\tau^{-2}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} - o_P\Big(\frac{m_{\max}^{5/2}\lambda_{\max}}{n^2}\Big)\mathbf{1}_{\dim\times\dim}\Big) \\
&\quad \cdot\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta}),
\end{aligned} \tag{2.5}$$

where $\mathbf{1}_{\dim\times\dim} \in I\!R^{\dim\times\dim}$ is a matrix consisting of ones and $\widehat{\boldsymbol{\alpha}}_{\text{reg}}$ is the regular B-spline estimator in the varying coefficient model context, the solution of (2.3) with $\lambda_0 = \ldots = \lambda_d = 0$, that corresponds to the response $\tau^{-1}(Y(t) - \zeta)$.

2.1. Consistency

We prove the consistency of the P-spline estimator in generalized 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 β_p is a spline function it can be represented exactly in a B-spline basis with a fixed number of knots. A detailed study on the influence of the smoothing parameter, the degree of the B-splines,

the differencing order, and the number of knots is carried out in Gijbels and Verhasselt (2010) for P-spline estimation in generalized linear models. 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 (2.5). We need some assumptions on the design and the B-spline basis:

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 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 satisfies* $M_3 \leqslant f_T(t) \leqslant M_4$ *for* $t \in \mathcal{T}$ *and positive constants* M_3 *and* M_4.
2. *The eigenvalues* $\eta_0(t), \ldots, \eta_d(t)$ *of* $\Sigma(t) = \mathrm{E}\,(\mathbf{X}(t)\mathbf{X}(t)')$ *satisfy* $M_5 \leqslant \eta_0(t) \leqslant \cdots \leqslant \eta_d(t) \leqslant M_6$ *for* $t \in \mathcal{T}$ *and positive constants* M_5 *and* M_6.
3. *There exist 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 exist a positive constant* M_8 *such that* $\mathrm{Var}\,(Y(t)|\mathbf{X}(t)) \leqslant M_8 < \infty$ *for* $t \in \mathcal{T}$.
5. $\limsup_n \left(\max_p m_p / \min_p m_p\right) < \infty$.
6. $K_{\max}^{3/2}\lambda_{\max}/n = o(1)$ *and* $K_{\max}/n = O(1)$, *where* $K_{\max} = \max_{0 \leqslant p \leqslant d} K_p$.
7. $\max_i N_i < \infty$.

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

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

where $||g||_{L_2} = \sqrt{\int_{\mathcal{T}} g(t)^2\, dt}$ *is the* L_2*-norm, assumed finite.*

The independence of the t_{ij} can be relaxed by replacing Assumption 1.1 by the requirement that Assumption 2 holds with probability tending to 1. A sufficient condition for Assumption 2 to hold is (see Huang, Wu, and Zhou (2004))

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

for some distribution function F_T with 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 $1_A(t)$ the indicator function of the set A.

Note that Assumptions 1.1−1.6 are natural and have been used in Wang, Li, and Huang (2008) and Antoniadis, Gijbels, and Verhasselt (2012b). Assumption 1.7 is a sufficient condition for $(\tau \mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda)^{-1}$ to exist.

We need some notation. 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_∞ distance between $\beta_p(\cdot)$ and $\mathbb{G}_p$. Let $\rho_n = \max_{0\leqslant p\leqslant d} \text{dist}(\beta_p, \mathbb{G}_p)$ the approximation error due to spline approximation. 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 deferred to the Appendix.

Theorem 1. *Suppose Assumptions* 1.1−1.6 *hold,* $\lim_n \text{dist}(\beta_p, \mathbb{G}_p) = 0$ *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* 1.1−1.6 *hold. If* $\lim_n(m_{\max}\ \log(m_{\max})\ n^{-1}) = 0$, *then*

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

Consequently

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

Theorem 2 has a corollary, here the notation $a_n \asymp b_n$ is used when $a_n b_n^{-1}$ and $b_n a_n^{-1}$ are bounded.

Corollary 1. *Suppose Assumptions* 1.1−1.7 *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* $K_p \asymp \left((1/n^2)\sum_{i=1}^n 1/N_i\right)^{-1/(2q+1)}$, $\lambda_p = \left((1/n^2)\sum_{i=1}^n 1/N_i\right)^{-\gamma}$ *for* $p = 0, \ldots, d$, *with* $\gamma \leqslant (q - 1/2)/(2q + 1)$. *Then*

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

The proof of this corollary is similar to the proof of Corollary 1 in Antoniadis, Gijbels, and Verhasselt (2012b), and is omitted here. 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 is the optimal rate for nonparametric regression with i.i.d. data under the same smoothness assumptions on the β_p (see Stone (1982)).

3. Nonnegative Garrote

The nonnegative garrote has been proposed by Breiman (1995) for subset regression in a classical multiple linear regression model. It starts with an initial estimator (the ordinary least squares estimator) and it shrinks or puts some coefficients of the ordinary least squares estimator equal to zero. In the varying coefficient model setup, the nonnegative garrote shrinkage factors $\widehat{\mathbf{c}} = (\widehat{c}_0, \ldots, \widehat{c}_d)'$ are defined as the solution of

$$\begin{cases} \min\limits_{\mathbf{c}} \Big[\sum\limits_{i=1}^{n} \dfrac{1}{N_i} \sum\limits_{j=1}^{N_i} \Big(Y_{ij} - \sum\limits_{p=0}^{d} X_{ij}^{(p)} c_p \widehat{\beta}_p^{\text{init}}(t_{ij}) \Big)^2 + \gamma \sum\limits_{p=0}^{d} c_p \Big], \\ \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d), \end{cases} \tag{3.1}$$

where $\mathbf{c} = (c_0, \ldots, c_d)'$, $\widehat{\beta}_p^{\text{init}}(\cdot)$ is the initial P-spline estimator for the regression coefficient function $\beta_p(\cdot)$, and $\gamma > 0$ is a regularization parameter. This optimization problem is equivalent to

$$\begin{cases} \min\limits_{\mathbf{c}} \Big[\sum\limits_{i=1}^{n} \dfrac{1}{N_i} \sum\limits_{j=1}^{N_i} \Big(-2Y_{ij} \sum\limits_{p=0}^{d} X_{ij}^{(p)} c_p \widehat{\beta}_p^{\text{init}}(t_{ij}) + \big(\sum\limits_{p=0}^{d} X_{ij}^{(p)} c_p \widehat{\beta}_p^{\text{init}}(t_{ij}) \big)^2 \Big) \\ \qquad + \gamma \sum\limits_{p=0}^{d} c_p \Big] \qquad \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d). \end{cases} \tag{3.2}$$

In the generalized varying coefficient model context, we replace the squared loss in (3.1) by -2 times the loglikelihood. The nonnegative garrote shrinkage factors $\widehat{\mathbf{c}} = (\widehat{c}_0, \ldots, \widehat{c}_d)'$ in this context are defined as the solution of

$$\begin{cases} \min\limits_{\mathbf{c}} \Big[-2 \sum\limits_{i=1}^{n} \dfrac{1}{N_i} \sum\limits_{j=1}^{N_i} (a(\phi))^{-1} \Big(Y_{ij} \sum\limits_{p=0}^{d} X_{ij}^{(p)} \widehat{\beta}_p^{\text{init}}(t_{ij}) c_p \\ \qquad - b\big(\sum_{p=0}^{d} X_{ij}^{(p)} \widehat{\beta}_p^{\text{init}}(t_{ij}) c_p \big) \Big) + \gamma \sum\limits_{p=0}^{d} c_p \Big] \qquad \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d). \end{cases} \tag{3.3}$$

In the asymptotic study we let γ depend on N, $\gamma = \gamma_N$. The nonnegative garrote estimate of the pth coefficient function is then $\widehat{\beta}_p^{\text{NNG}}(t) = \widehat{\beta}_p^{\text{init}}(t) \widehat{c}_p$.

Using the first order approximation of $g^{-1}(\theta)$ for small θ, we find a second order approximation of $b(\theta)$:

$$b(\theta) = b(0) + \zeta\theta + \frac{\tau}{2}\theta^2,$$

since $\frac{db}{d\theta}\big|_{\theta(\mathbf{X}(t))} = \mu(\mathbf{X}(t)) = g^{-1}(\theta(\mathbf{X}(t)))$. The use of this approximation in (3.3) gives an approximation for the nonnegative garrote optimization problem

for generalized varying coefficient models:

$$\begin{cases} \min\limits_{\mathbf{c}} \Big[\sum\limits_{i=1}^{n} \dfrac{1}{N_i} \sum\limits_{j=1}^{N_i} \dfrac{-2(Y_{ij}-\zeta) \sum\limits_{p=0}^{d} X_{ij}^{(p)} \widehat{\beta}_p^{\text{init}}(t_{ij}) c_p + \tau \big(\sum\limits_{p=0}^{d} X_{ij}^{(p)} \widehat{\beta}_p^{\text{init}}(t_{ij}) c_p \big)^2 - b(0)}{a(\phi)} \\ \qquad + \gamma \sum\limits_{p=0}^{d} c_p \Big] \qquad \text{s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d). \end{cases}$$

Since we minimize with respect to $\mathbf{c}$, this last optimization problem is equivalent to

$$\begin{cases} \min\limits_{\mathbf{c}} \Big[\sum\limits_{i=1}^{n} \dfrac{1}{N_i} \sum\limits_{j=1}^{N_i} \Big(-2\tau^{-1}(Y_{ij}-\zeta) \sum\limits_{p=0}^{d} X_{ij}^{(p)} \widehat{\beta}_p^{\text{init}}(t_{ij}) c_p \\ \qquad + \big(\sum\limits_{p=0}^{d} X_{ij}^{(p)} \widehat{\beta}_p^{\text{init}}(t_{ij}) c_p \big)^2 \Big) + \gamma a(\phi) \tau^{-1} \sum\limits_{p=0}^{d} c_p \Big] \text{ s.t. } 0 \leqslant c_p \ (p = 0, \ldots, d). \end{cases} \tag{3.4}$$

Note that this o corresponds to (3.2) for (normal) varying coefficient models (1.1) with response $\tau^{-1}(Y(t) - \zeta)$ and regularization parameter $\gamma a(\phi)\tau^{-1}$. The estimation and variable selection consistency of the nonnegative garrote with P-splines in generalized varying coefficient models is therefore an immediate corollary of the consistency of the nonnegative garrote in regular varying coefficient models (Theorem 4 of Antoniadis, Gijbels, and Verhasselt (2012b)).

Let $\kappa_n = \max(\rho_n, r_n, m_{\max}^{3/2} \lambda_{\max} n^{-1})$. The nonnegative garrote estimator with the P-spline estimator as initial estimator for $\beta_p(t)$ is $\widehat{f}_p^{\text{NNG}}(t) = \widehat{c}_p \widehat{f}_p^{\text{init}}(t)$, and its estimation and variable selection consistency are given as follows.

Theorem 3. *If the assumptions of Theorem* 1 *and Theorem* 2 *hold and* $\gamma a(\phi)\tau^{-1}/N \to 0$ *such that* $\kappa_n = o(\gamma a(\phi)\tau^{-1}/N)$, *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. $\sup_p \mathrm{E}\,(\widehat{f}_p^{\text{NNG}}(t) - f_p(t))^2 = O_P\big(\big(\gamma a(\phi)\tau^{-1}/N\big)^2\big)$ *(where the expectation is with respect to* $\mathcal{X}$*) for all* $t \in \mathcal{T}$.

This result is an immediate consequence of Theorem 4 in Antoniadis, Gijbels, and Verhasselt (2012b), using regularization parameter $\gamma a(\phi)\tau^{-1}$ and response $\tau^{-1}(Y(t) - \zeta)$.

Table 1. Some distributions belonging to the exponential family.

Distribution	Var (Y)	Canonical link $g(\mu)$	$b(\theta)$	$a(\phi)$	θ	ζ	τ
$\mathrm{N}(\mu, \sigma^2)$	σ^2	μ	$\theta^2/2$	ϕ (σ^2)	μ	0	1
Poisson(μ)	μ	$\log(\mu)$	e^θ	1	$\log(\mu)$	1	1
Bin(n, p)	$np(1-p)$	$\log(\frac{\mu}{n-\mu})$	$n\log(1+e^\theta)$	1	$\log(\frac{p}{1-p})$	$n/2$	$n/4$

4. Applications

4.1. Simulated data

We investigated the performance of the nonnegative garrote with P-splines on simulated data from a Poisson, Bernoulli and normal distribution. Moreover, if we use the normal distribution, we fall back on varying coefficient models (1.1). The corresponding link and other functions are given in Table 1. The parameters ζ and τ in this table are found by using a Taylor series of $g^{-1}(\cdot)$ at 0.

The covariates were simulated in a similar fashion as in Antoniadis, Gijbels, and Verhasselt (2012b). The simulated data examples consist of 100 samples of size $n = 200$. We considered two settings with 4 nonzero regression coefficients out of 11 ($d = 10$) and 21 ($d = 20$). The first three variables $X^{(1)}(t), X^{(2)}(t)$ and $X^{(3)}(t)$ and the intercept $X^{(0)}(t)$ were the relevant variables. $X^{(0)}(t) \equiv 1$, $X^{(1)}(t)$ uniformly distributed on $[t/10, 2+t/10]$ for any t, $X^{(2)}(t)$, conditioned on $X^{(1)}(t)$, normally distributed with mean 0 and variance $(1 + X^{(1)}(t))/(2 + X^{(1)}(t))$, and $X^{(3)}(t)$, independent of $X^{(1)}(t)$ and $X^{(2)}(t)$, a Bernoulli random variable with probability of success 0.6 (and thus it did not vary with t). The irrelevant variables $X^{(p)}(t)$ were independent realizations of a Gaussian process with mean zero and $\mathrm{Cov}\,(X^{(p)}(t), X^{(p)}(s)) = 4\exp^{-|t-s|}$. The observation time points t_{ij} were the same for all subjects ($i = 1, \ldots, n$): $\{1, \ldots, 30\}$.

We used P-splines with 10 equidistant knots, degree 3, and differencing order 2 for the estimation of all regression coefficients. The smoothing parameters $\lambda_0, \ldots, \lambda_d$ were chosen with the EM algorithm of Marx (2010). The shrinkage parameter γ was found by minimizing BIC (Bayesian Information Criterion).

The selection performance of the nonnegative garrote with P-splines is evaluated on the basis of criteria described in Table 2 based on 100 simulations. These criteria were also used in Antoniadis, Gijbels, and Verhasselt (2012b). In addition a table with the appearance frequency of each variable in the 100 simulated data sets is given for the setting with ten covariates.

We give a graph of the fitted mean (blue dashed-dotted line) and the true mean (red solid line) for the simulated data set with median 'linear estimation error': $\sum_{i=1}^{n}\sum_{j=1}^{N_i}(\eta_{ij} - \hat{\eta}_{ij})^2$ in each simulation setup. In addition we present - for the simulation setups with $d = 10$ - the true (red solid line) and estimated

Table 2. Evaluation criteria.

MS	median number of selected coefficients
MTZ	median of zero coefficients restricted to the true zero coefficients $(\beta_4, \ldots, \beta_d)$
MFZ	median of zero coefficients restricted to the true non-zero coefficients $(\beta_0, \ldots, \beta_3)$
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 ($X^{(1)}, X^{(2)}$ and $X^{(3)}$) selected in the model
AverI	average of the number of irrelevant variables $(X^{(4)}, \ldots, X^{(d)})$ selected in the model

Table 3. Simulated examples. Appearance frequency of the variables for models with $d = 10$.

Distribution	$X^{(0)}$	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	$X^{(6)}$	$X^{(7)}$	$X^{(8)}$	$X^{(9)}$	$X^{(10)}$
Poisson$(\mu(t))$	100	100	100	100	0	0	0	0	0	0	0
Bin$(1, p(t))$	100	100	100	99	1	3	4	2	3	2	2
N$(\mu(t), 1)$	100	100	100	100	0	0	0	0	0	0	0
N$(\mu(t), 1.25^2)$	100	100	100	100	0	0	0	0	0	0	0
N$(\mu(t), 2^2)$	100	100	100	100	2	1	0	0	0	0	0

regression coefficient functions (blue dashed-dotted line) for the same simulated data set, as well as the estimates from the simulations corresponding to the first and third quartile of the 'linear estimation error'. Moreover we give a measure for the estimation error (AISE) in Table 4: AISE$= (1/R)\sum_{r=1}^{R}(1/n)\sum_{i=1}^{n}(\hat{\beta}_r(t_i) - \beta_r(t_i))^2$, where R is the number of selected components and $t_i = i/30$ for $i = 1, \ldots, 30$.

Poisson distribution

The linear predictor for the Poisson varying coefficient model is

$$\eta(\mathbf{X}(t)) = 5.5 + 0.1(\beta_0(t) + \beta_1(t)X^{(1)} + \beta_2(t)X^{(2)} + \ldots + \beta_d(t)X^{(d)}).$$

The coefficient functions of the relevant variables are

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

and for the irrelevant variables $\beta_p(t) = 0$ $(p = 4, \ldots, d)$.

From Figure 1 it is clear that in both settings, the estimated mean is close to the true mean. The first four estimated regression coefficients for the model with $d = 10$ are given in Figure 2. From both sets of figures we can conclude that the nonnegative garrote performs well as an estimation technique. The estimation of the baseline effect and the coefficient of $X^{(3)}$ is less good than the estimation of the other coefficients. Overall, the estimation of the mean response is good

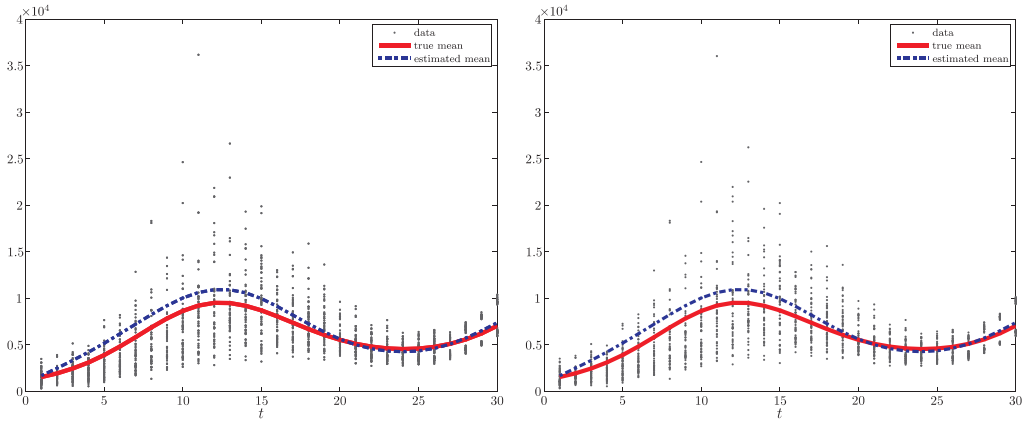


Figure 1. Simulated example Poisson($\mu(t)$). Data, true mean and fitted mean. Model with (a) $d = 10$ and (b) $d = 20$.

with almost no variability over the simulations. Note that the 'bad', 'median' and 'good' estimates of the regression coefficient functions are almost exactly the same.

Nonnegative garrote with P-splines performs well also as a selection technique, this is illustrated in Tables 4 and 3. The nonnegative garrote never removes true non-zero covariates (MFZ= 0) and in all simulated data sets all relevant covariates were included and all irrelevant covariates excluded (AverR= 3 and AverI= 0).

Bernoulli distribution

In the Bernoulli varying coefficient models the linear predictor is

$$\eta(\mathbf{X}(t)) = -3 + 0.1(\beta_0(t) + \beta_1(t)X^{(1)} + \beta_2(t)X^{(2)} + \ldots + \beta_d(t)X^{(d)}),$$

where the regression coefficients are the same as in the Poisson setup.

The estimation performance of the nonnegative garrote with P-splines is good in both settings (see Figure 3, the estimated probability of success is close to the true success probability, though there is more variability over the simulations than in the Poisson model (see Figure 4 and Table 4). The selection performance is also good, but in a few cases (less than 10%) an irrelevant variable is included in the model and once (for $d = 10$) the third covariate is excluded (see Tables 3 and 4).

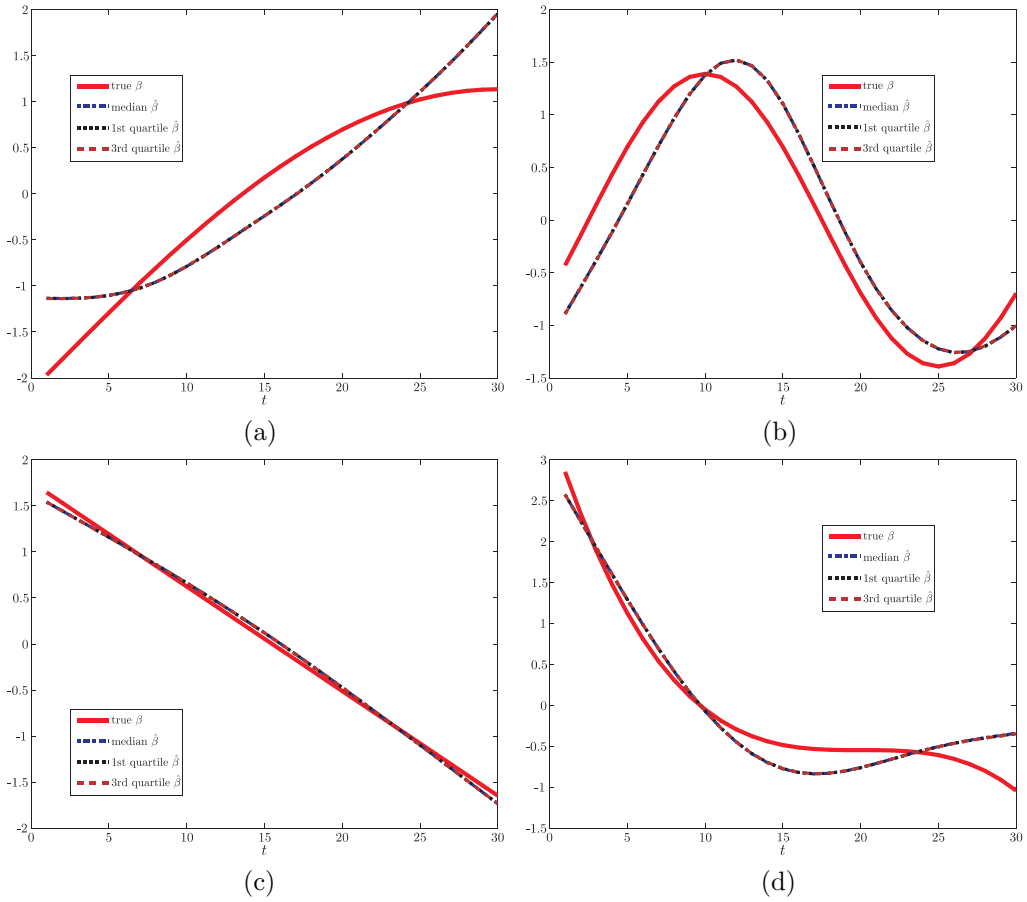


Figure 2. Simulated example Poisson($\mu(t)$). Regression coefficients for model with $d = 10$. Coefficient of (a) $X^{(0)}$, (b) $X^{(1)}$, (c) $X^{(2)}$ and (d) $X^{(3)}$.

Normal distribution

The linear predictor in this normal varying coefficient model is given by

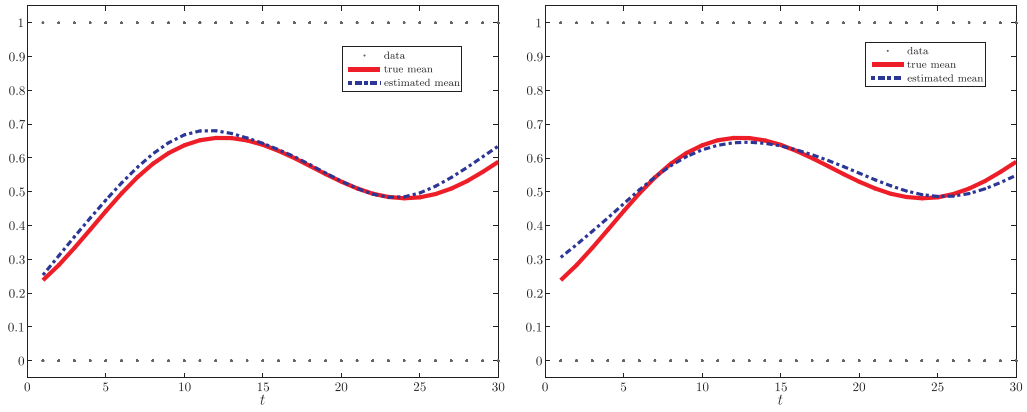
$$\eta(\mathbf{X}(t)) = \beta_0(t) + \beta_1(t)X^{(1)} + \beta_2(t)X^{(2)} + \ldots + \beta_d(t)X^{(d)},$$

where the regression coefficients are the same as before. We consider three variance levels for the normal distribution: 1, 1.25^2 and 2^2.

From Tables 3 and 4 we can see that when the variance increases a few irrelevant variables are included in the selected model. Especially in the harder setting with 20 covariates and when the variance of the error term is 4, more irrelevant covariates are included, though relevant covariates are never excluded from the selected model. The fitted mean coincides almost perfectly with the true mean in all settings (see Figure 5). The estimated coefficients are almost

Table 4. *Simulated examples. Evaluation criteria for the different models. The numbers in brackets are the standard deviation* $(\cdot)$ *or first and third quartiles* $(\cdot,\cdot)$.

Distribution	d	AISE	MS	MTZ	MFZ	MTP	MFP	PercT	AverR	AverI
optimal value	10/20	0	4	7/17	0	4	0	1	3	0
Poisson$(\mu(t))$	10	0.0847	4	7	0	4	0	1	3	0
		(0.0004)	(4,4)	(7,7)	(0,0)	(4,4)	(0,0)		(0)	(0)
Poisson$(\mu(t))$	20	0.0845	4	17	0	4	0	1	3	0
		(0.0004)	(4,4)	(17,17)	(0,0)	(4,4)	(0,0)		(0)	(0)
Bin$(1,p(t))$	10	0.1167	4	7	0	4	0	0.91	2.9900	0.1700
		(0.1268)	(4,4)	(7,7)	(0,0)	(4,4)	(0,0)		(0.1000)	(0.6039)
Bin$(1,p(t))$	20	0.1149	4	17	0	4	0	0.94	3	0.1400
		(0.1066)	(4,4)	(17,17)	(0,0)	(4,4)	(0,0)		(0)	(0.6034)
N$(\mu(t),1)$	10	0.0031	4	7	0	4	0	1	3	0
		(0.0005)	(4,4)	(7,7)	(0,0)	(4,4)	(0,0)		(0)	(0)
N$(\mu(t),1)$	20	0.0031	4	17	0	4	0	1	3	0
		(0.0006)	(4,4)	(17,17)	(0,0)	(4,4)	(0,0)		(0)	(0)
N$(\mu(t),1.25^2)$	10	0.0035	4	7	0	4	0	1	3	0
		(0.0008)	(4,4)	(7,7)	(0,0)	(4,4)	(0,0)		(0)	(0)
N$(\mu(t),1.25^2)$	20	0.0035	4	17	0	4	0	1	3	0
		(0.0008)	(4,4)	(17,17)	(0,0)	(4,4)	(0,0)		(0)	(0)
N$(\mu(t),2^2)$	10	0.0108	4	7	0	4	0	0.97	3	0.0300
		(0.0329)	(4,4)	(7,7)	(0,0)	(4,4)	(0,0)		(0)	(0.1714)
N$(\mu(t),2^2)$	20	0.0338	4	17	0	4	0	0.85	3	0.1500
		(0.0691)	(4,4)	(17,17)	(0,0)	(4,4)	(0,0)		(0)	(0.3589)



(a) (b)

Figure 3. Simulated example Bin$(1,p(t))$. Data, true mean and fitted mean. Model with (a) $d=10$ and (b) $d=20$.

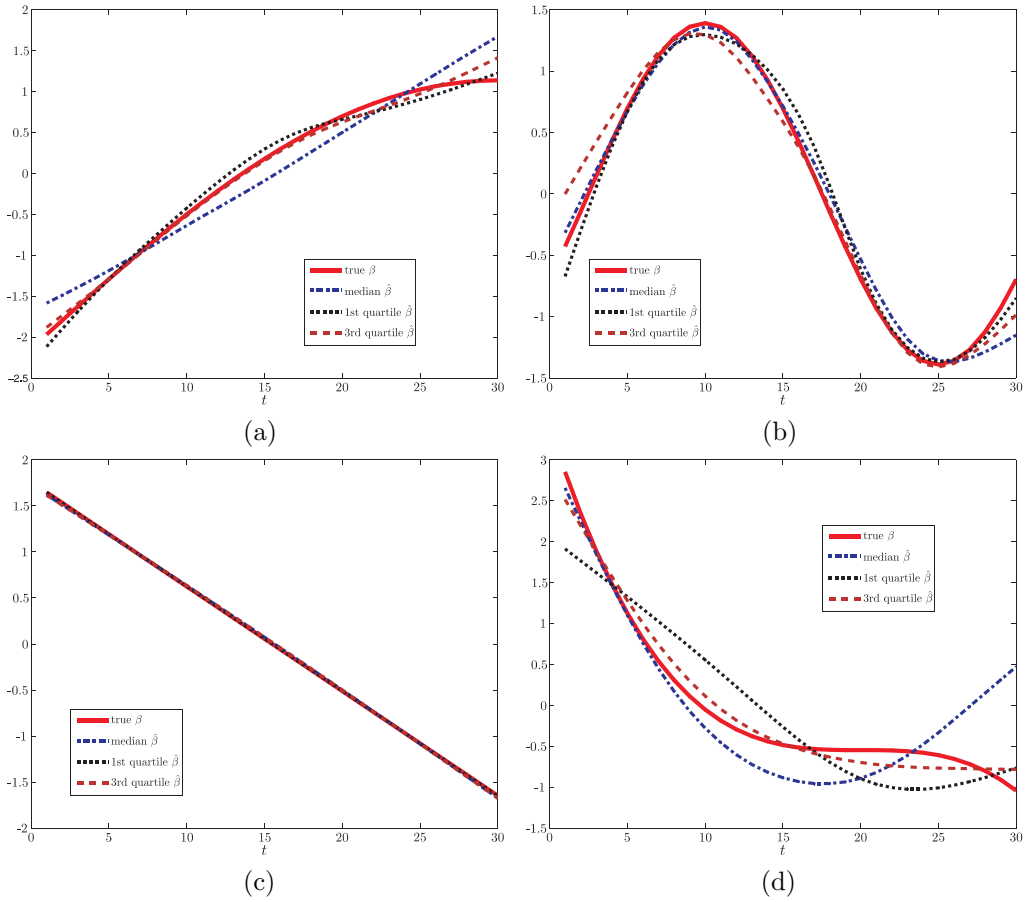


Figure 4. Simulated example Bin$(1, p(t))$. Regression coefficients for model with $d = 10$. Coefficient of (a) $X^{(0)}$, (b) $X^{(1)}$, (c) $X^{(2)}$ and (d) $X^{(3)}$.

exactly the same as the true regression coefficients (see Figures 6 and 7 for $d = 10$ and $\sigma = 1$ and 2 respectively).

4.2. Hong Kong environmental data

We consider the data set introduced in Section 1. We are interested in estimating the number of hospital admissions for cardiovascular and respiratory diseases on every Friday from January 1, 1994 to December 31, 1995, and in determining which pollutants are most influencing the hospital admissions. For each Friday, the concentration of nitrogen dioxide ($X^{(1)}$), sulphur dioxide ($X^{(2)}$), dust ($X^{(3)}$) and ozone ($X^{(4)}$), temperature ($X^{(5)}$), and humidity ($X^{(6)}$) as well as the total number of admissions for cardiovascular and respiratory diseases are measured. This data set is also considered in Cai, Fan, and Li (2000). However,

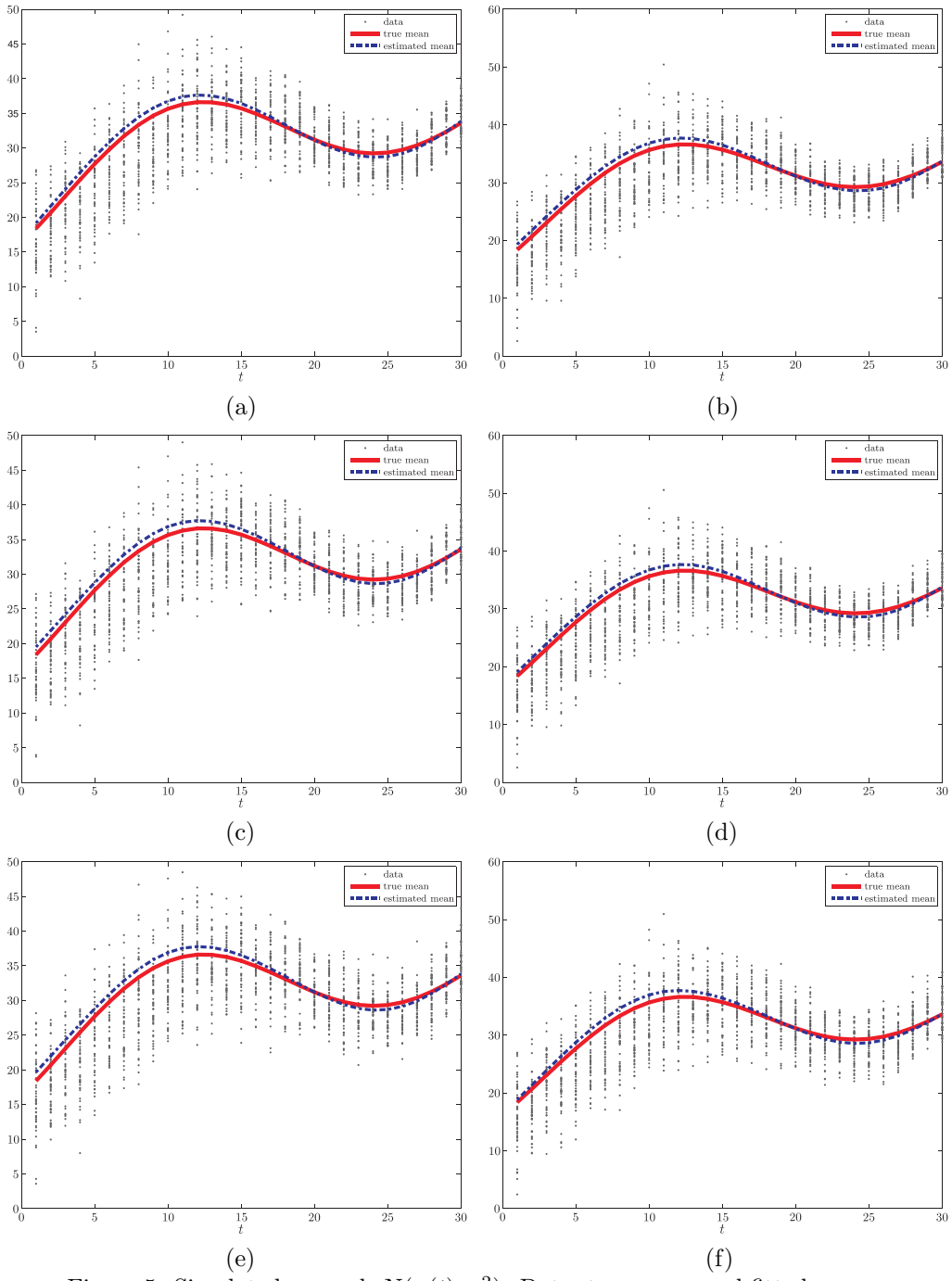


Figure 5. Simulated example $\mathrm{N}(\mu(t), \sigma^2)$. Data, true mean and fitted mean. Model with (a) $d = 10$, $\sigma = 1$; (b) $d = 20$, $\sigma = 1$;(c) $d = 10$, $\sigma = 1.25$; (d) $d = 20$, $\sigma = 1.25$; (e) $d = 10$, $\sigma = 2$ and (f) $d = 20$, $\sigma = 2$.

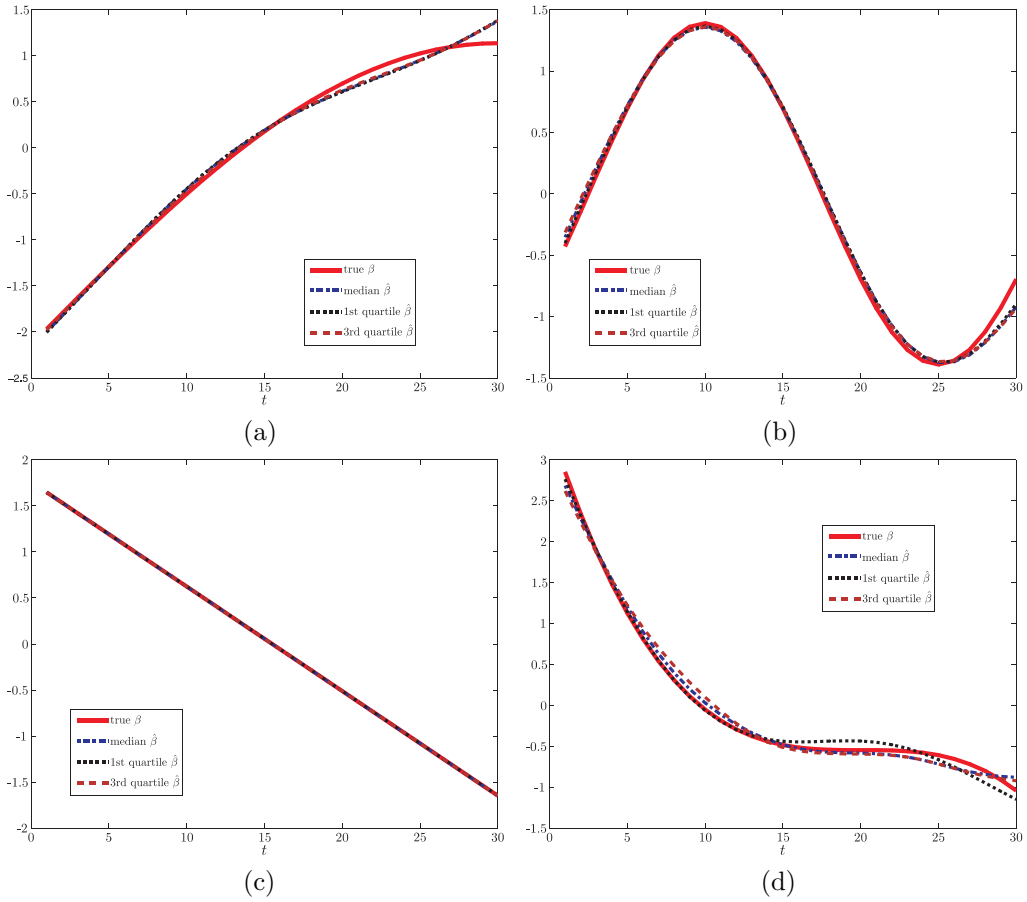


Figure 6. Simulated example $\mathrm{N}(\mu(t), \sigma^2)$. Regression coefficients for model with $d = 10$ and $\sigma = 1$. Coefficient of (a) $X^{(0)}$, (b) $X^{(1)}$, (c) $X^{(2)}$ and (d) $X^{(3)}$.

they consider only the concentration of nitrogen dioxide, sulphur dioxide, and dust as covariates.

P-splines with 10 equidistant knots, degree 3 and differencing order 2 were used for the estimation of all regression coefficients. The smoothing parameters $\lambda_0, \ldots, \lambda_6$ and the shrinkage parameter γ were chosen as in the simulated data examples.

The nonnegative garrote procedure selects all covariates, except the concentration of nitrogen dioxide. The fitted regression coefficients are given in Figure 8 and the logarithm of the fitted mean is presented in Figure 9. The fitted mean follows the data cloud very well. Globally there is an increasing trend in the number of hospital admissions over time with a peak around 65 weeks. This peak is also prominent in the baseline coefficient $\beta_0(t)$. These results coincide

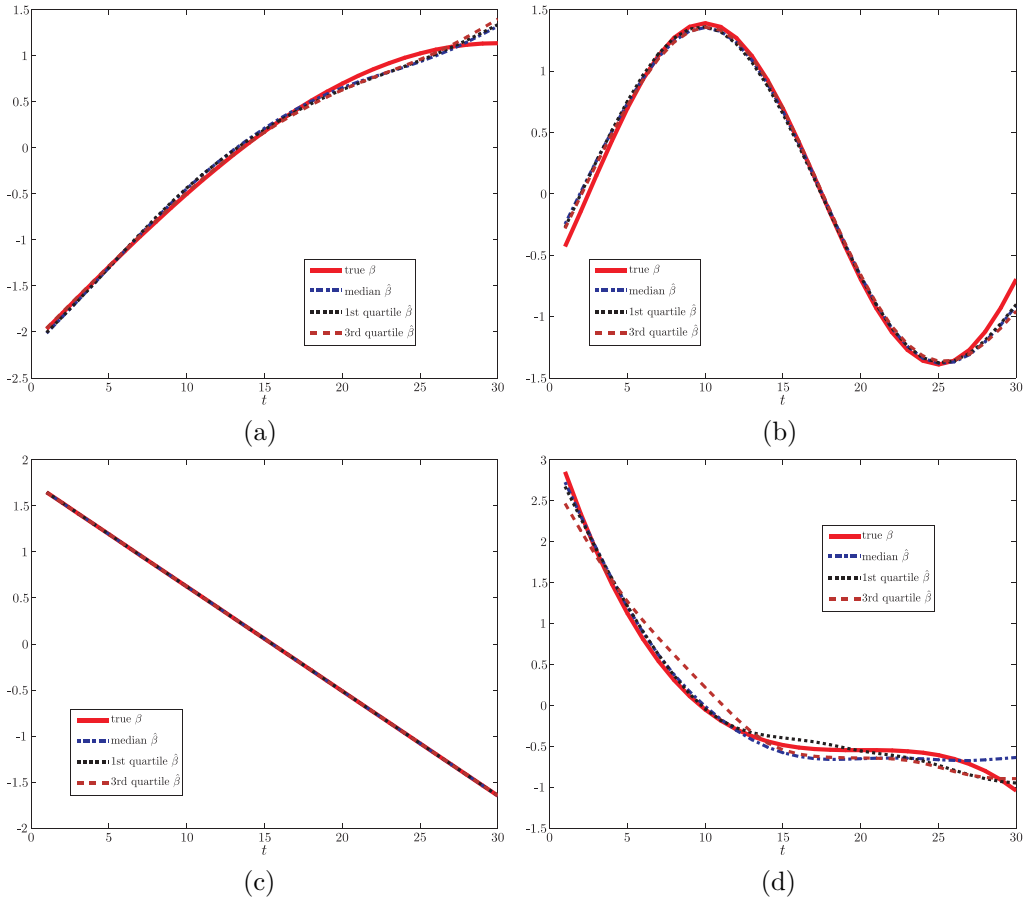


Figure 7. Simulated example $\mathrm{N}(\mu(t), \sigma^2)$. Regression coefficients for model with $d = 10$ and $\sigma = 2$. Coefficient of (a) $X^{(0)}$, (b) $X^{(1)}$, (c) $X^{(2)}$ and (d) $X^{(3)}$.

with those of Cai, Fan, and Li (2000). Nevertheless in their analysis the concentration of nitrogen dioxide is relevant. However, if we take the same covariates as they do, the concentration of nitrogen dioxide is included in our model. In fact none of the covariates is removed from the model when regressing on the same covariates. The dust level has a globally decreasing effect over time on the number of hospital admissions, while the effect of ozone and SO_2 seem to be higher in the last year.

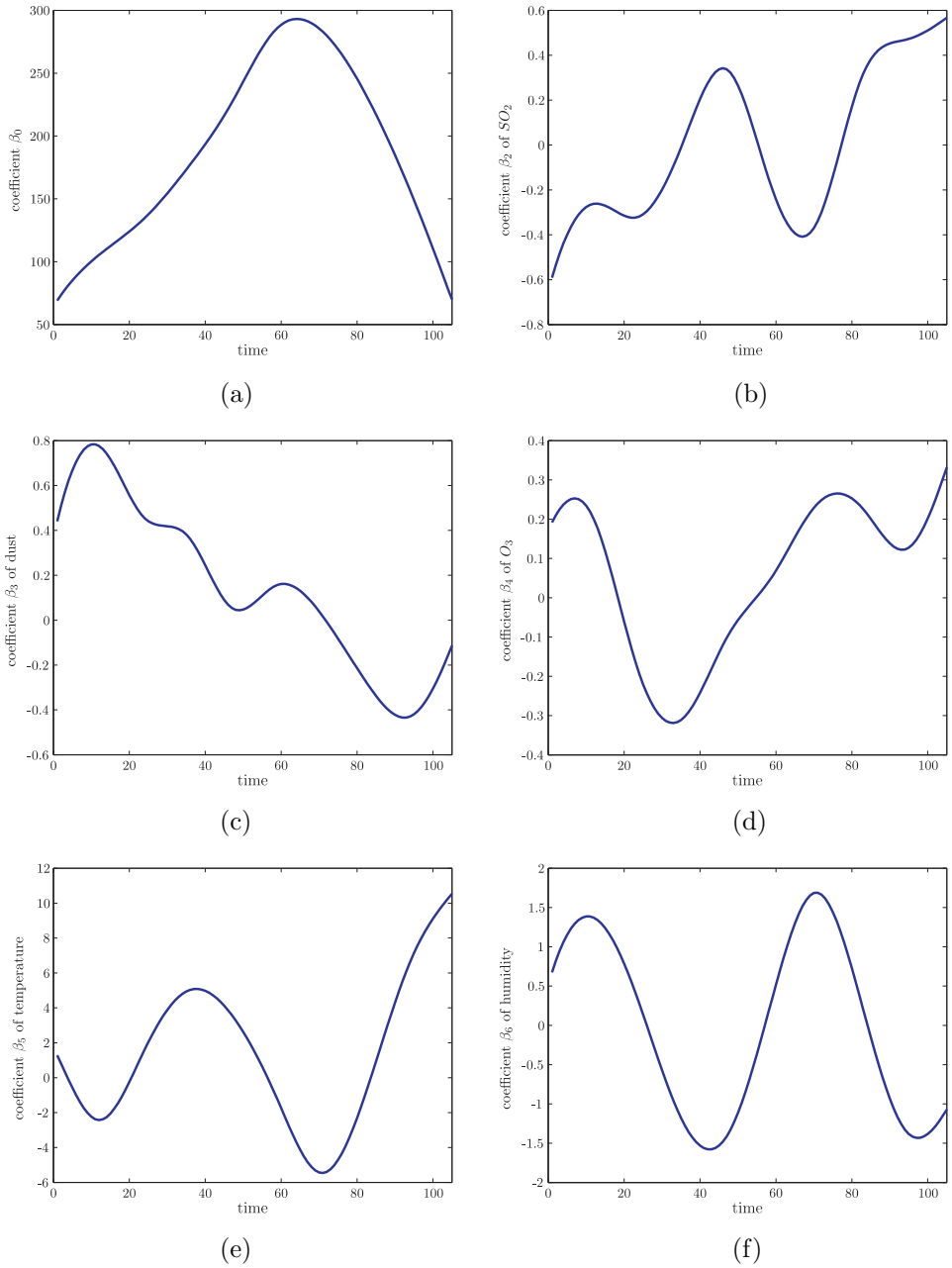


Figure 8. Hong Kong environmental data. Regression coefficients of (a) $X^{(0)}$, (b) SO_2, (c) dust, (d) O_3, (e) temperature and (f) humidity.

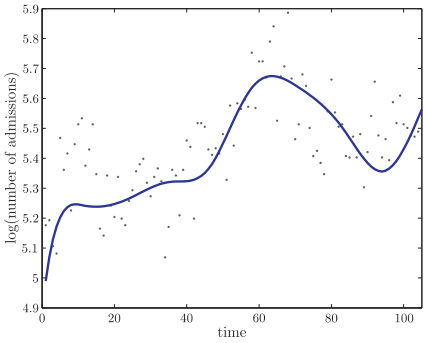


Figure 9. Hong Kong environmental data. Logarithm of the number of hospital admissions.

Acknowledgement

The author is very grateful to Professor Jianqing Fan for providing the Hong Kong environmental data.

This research was supported by the IAP Research Network P6/03 of the Belgian State (Belgian Science Policy) and the Research Fund of the KULeuven (PDM-short term).

Appendix

A.1. Proof of Lemma 1

We need some notation. If $A = (A_{ij})$ is an $m \times n$ real valued matrix, a_n and b_n sequences of positive numbers, then the ∞-norm of A is $\|A\|_\infty = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} |A_{ij}|$, for $1 < \nu < \infty$, the ν-norm of A is $\|A\|_\nu = (\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^\nu)^{1/\nu}$, and $a_n \lesssim b_n$ means that a_n/b_n is bounded and $a_n \asymp b_n$ means that a_n/b_n and b_n/a_n are bounded.

Proof. From Lemma A.3 of Huang, Wu, and Zhou (2004) there exist positive constants M_1 and M_2 such that, except on an event whose probability tends to zero, all the eigenvalues of $(m_{\max}/n)\mathbf{U}'\mathbf{W}\mathbf{U}$ fall between M_1 and M_2, and consequently $\mathbf{U}'\mathbf{W}\mathbf{U}$ is invertible. Therefore $||(m_{\max}/n)\mathbf{U}'\mathbf{W}\mathbf{U}||_2 \asymp 1$ and $||((m_{\max}/n)\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}||_2 \asymp 1$. Moreover from Gijbels and Verhasselt (2010) (Proposition 1 and the proof of Theorem 1) we know that

$$(\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda)^{-1} = \sum_{j=0}^{\infty} (-1)^j (\tau\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} \Big(a(\phi)\mathbf{Q}_\lambda (\tau\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} \Big)^j$$

and the series converges if $||a(\phi)\mathbf{Q}_\lambda(\tau\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}||_2 < 1$.

We now show that $||a(\phi)\mathbf{Q}_\lambda(\tau\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}||_2 = O(m_{\max}^{3/2}\lambda_{\max}/n)$ except on an event with probability tending to zero:

$$\begin{aligned}||a(\phi)\mathbf{Q}_\lambda(\tau\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}||_2 &\leqslant \Big|\frac{a(\phi)}{\tau}\Big|||\mathbf{Q}_\lambda||_2||(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}||_2 \lesssim \frac{m_{\max}}{n}\sqrt{\sum_{p=0}^{d} m_p||\mathbf{Q}_\lambda||_\infty^2}\\ &\leqslant \frac{m_{\max}}{n}\sqrt{d\ m_{\max}}||\mathbf{Q}_\lambda||_\infty \leqslant \frac{m_{\max}^{3/2}}{n}\sqrt{d}\lambda_{\max}\max_{0\leqslant p\leqslant d} 4^{k_p}\\ &= O\Big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\Big),\end{aligned}$$

since $||\lambda_p\mathbf{D}_{k_p}'\mathbf{D}_{k_p}||_\infty = \lambda_p 4^{k_p}$ (see Gijbels and Verhasselt (2010), proof of Theorem 1) with k_p fixed, $a(\phi) < \infty$, and $0 < |\tau| < \infty$.

A.2. Proofs of Theorems 1 and 2

Proof of Theorem 1. The existence of the P-spline estimates of the coefficients $\boldsymbol{\alpha}$ and thus also the existence of the P-spline estimates $\widehat{\beta}_p(t)$ $(p = 0, \ldots, d)$ follows from (2.4) and Lemma 1. The consistency of $\widehat{\beta}_p(t)$ $(p = 0, \ldots, d)$ is a consequence of Theorem 2.

Let $\widetilde{Y}_{ij} = \mu_{ij} - \zeta$, $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{iN_i})'$, $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1, \ldots, \widetilde{\mathbf{Y}}_n)'$ and

$$\widetilde{\boldsymbol{\alpha}} = \big(\tau\mathbf{U}'\mathbf{W}\mathbf{U} + a(\phi)\mathbf{Q}_\lambda\big)^{-1}\mathbf{U}\mathbf{W}\widetilde{\mathbf{Y}}.$$

Then $\mathrm{E}\,(\widehat{\boldsymbol{\alpha}}) = \widetilde{\boldsymbol{\alpha}}$ and $\mathrm{E}\,(\widehat{\boldsymbol{\beta}}(t)) = \widetilde{\boldsymbol{\beta}}(t) = \mathbf{B}(t)\widetilde{\boldsymbol{\alpha}}$ for $t \in \mathcal{T}$, where the expectation is taken conditioning on $\mathcal{X}$. Let $\widehat{\boldsymbol{\alpha}}_{\mathrm{reg}}$ be the regular B-spline estimator, (2.4) with $\lambda_0 = \ldots = \lambda_d = 0$, and write $\mathrm{E}\,(\widehat{\boldsymbol{\alpha}}_{\mathrm{reg}}) = \widetilde{\boldsymbol{\alpha}}_{\mathrm{reg}}$.

Proof of Theorem 2. First note that from the properties of B-spline functions (see for example Lemma A.1. in Huang, Wu, and Zhou (2004)) we know that $||\beta_p(t)||_{L_2}^2 \asymp ||\boldsymbol{\alpha}_p||_2^2/m_p$ for $p = 0, \ldots, d$.

We first find the rate of $||\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}||_{L_2}^2$, with $||\boldsymbol{\beta}||_{L_2} = \sqrt{\sum_{p=0}^{d}||\beta_p||_{L_2}^2}$, based on the rates of the regular B-spline estimator of Huang, Wu, and Zhou (2004).

From (2.5) we have that

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} = \Big(&\tau^{-1}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1} - a(\phi)\tau^{-2}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\\ &+o_P\Big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\Big)\tau^{-1}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\Big)\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta} - \widetilde{\mathbf{Y}})\\ = \widehat{\boldsymbol{\alpha}}_{\mathrm{reg}} - \widetilde{\boldsymbol{\alpha}}_{\mathrm{reg}} &- a(\phi)\tau^{-2}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta} - \widetilde{\mathbf{Y}})\\ &+\tau^{-1}o_P\Big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\Big)(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}(\mathbf{Y} - \boldsymbol{\zeta} - \widetilde{\mathbf{Y}}).\end{aligned}$$

Consequently

$$||\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}||_2 \leqslant ||\widehat{\boldsymbol{\alpha}}_{\text{reg}}-\widetilde{\boldsymbol{\alpha}}_{\text{reg}}||_2+\Big(|a(\phi)\tau^{-1}|||(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}||_2||\mathbf{Q}_\lambda||_2+o_P\big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\big)\Big)$$
$$\cdot\tau^{-1}|||(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}(\mathbf{Y}-\boldsymbol{\zeta}-\widetilde{\mathbf{Y}})||_2.$$

From the proof of Lemma 1 we know that $|a(\phi)\tau^{-1}|||(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}||_2||\mathbf{Q}_\lambda||_2 = O_P(m_{\max}^{3/2}\lambda_{\max}/n)$, Lemma A.4 and A.5 of Huang, Wu, and Zhou (2004) give that (by Assumption 1.4 and the fact that $a(\phi)<\infty$ and $0<|\tau|<\infty$),

$$||\tau^{-1}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}(\mathbf{Y}-\boldsymbol{\zeta}-\widetilde{\mathbf{Y}})||_2^2 = O_P\Big(\frac{m_{\max}^2}{n^2}\sum_i\big(\frac{1}{N_i}+\frac{1}{m_{\max}}\big(1-\frac{1}{N_i}\big)\big)\Big)$$
$$||\widehat{\boldsymbol{\alpha}}_{\text{reg}}-\widetilde{\boldsymbol{\alpha}}_{\text{reg}}||_2^2 = O_P\Big(\frac{m_{\max}^2}{n^2}\sum_i\big(\frac{1}{N_i}+\frac{1}{m_{\max}}\big(1-\frac{1}{N_i}\big)\big)\Big).$$

Therefore

$$||\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}||_2^2 = O_P\Big(\frac{m_{\max}^2}{n^2}\sum_i\big(\frac{1}{N_i}+\frac{1}{m_{\max}}\big(1-\frac{1}{N_i}\big)\big)\big(1+\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\big)^2\Big)$$
$$= O_P\Big(m_{\max}r_n^2\big(1+\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\big)^2\Big) = O_P\big(m_{\max}r_n^2\big),$$
$$||\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}||_{L_2}^2 = O_P\Big(r_n^2\big(1+\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\big)^2\Big) = O_P(r_n^2),$$

since Assumption 1.5 and 1.6 hold.

Finally we prove the rate of $||\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}||_{L_2}^2$. First note that

$$\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\alpha}}_{\text{reg}} - a(\phi)\tau^{-2}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{Q}_\lambda(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}\widetilde{\mathbf{Y}}$$
$$+o_P\Big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\Big)\tau^{-1}(\mathbf{U}'\mathbf{W}\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}\widetilde{\mathbf{Y}}$$
$$= \Big(1-O_P\big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\big)\Big)\widetilde{\boldsymbol{\alpha}}_{\text{reg}},$$

consequently

$$\widetilde{\boldsymbol{\beta}} = \mathbf{B}(t)\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\beta}}_{\text{reg}}\Big(1-O_P\big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\big)\Big)$$
$$||\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}||_{L_2} \leqslant ||\widetilde{\boldsymbol{\beta}}_{\text{reg}}-\boldsymbol{\beta}||_{L_2} + O_P\Big(\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\Big)||\widetilde{\boldsymbol{\beta}}_{\text{reg}}||_{L_2},$$

where $\widetilde{\boldsymbol{\beta}}_{\text{reg}} = \mathbf{B}(t)\widetilde{\boldsymbol{\alpha}}_{\text{reg}}$.

From Theorem 2 of Huang, Wu, and Zhou (2004) we know that $||\widetilde{\boldsymbol{\beta}}_{\text{reg}} - \boldsymbol{\beta}||_{L_2} = O_P(\rho_n)$. Since a spline $\beta_p(\cdot)$ is a continuous function on $\mathcal{T}=[0,T]$, $||\widetilde{\boldsymbol{\beta}}_{\text{reg}}||_{L_2}$ is bounded and therefore $||\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}||_{L_2} = O_P\Big(\rho_n + m_{\max}^{3/2}\lambda_{\max}/n\Big)$.

References

Antoniadis, A., Gijbels, I. and Verhasselt, A. (2012a). Variable selection in additive models using P-splines. *Technometrics* **54**, 425-438.

Antoniadis, A., Gijbels, I. and Verhasselt, A. (2012b). Variable selection in varying coefficient models using P-splines. *J. Comput. Graph. Statist.* **21**, 638-661.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **51**, 373-384.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Cantoni, E., Flemming, J. and Ronchetti, E. (2000). Variable selection in additive models by nonnegative garrote. *Statistical Modelling* **11**, 165-180.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11**, 89-102.

Eilers, P. and Marx, B. (2002). Generalized Linear Additive Smooth Structures. *J. Comput. Graph. Statist.* **11**, 758-783.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models, *Ann. Statist.* **27**, 1491-1518.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models, *Statistics and Its Interface* **1**, 179-195.

Gijbels, I. and Verhasselt, A. (2010). Regularization and P-splines in generalized linear models. *J. Nonparametr. Stat.* **22**, 271-295.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55**, 757-796.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14**, 763-788.

Marra, G. and Wood, S. (2011). Practical variable selection in generalized additive models. *Comput. Statist. Data Anal.* **55**, 2372-2387.

Marx, B. (2010). P-spline varying coefficient models for complex data. In *Statistical Modelling and Regression Structures*. Festschrift in Honour of Ludwig Fahrmeir (Edied by T. Kneib and G.Tutz). Springer, New York.

McCullagh, P. and Nelder, J. A. (1995). *Generalized Linear Models*. Chapman and Hall, New York.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Şentürk, D. and Müller, H. G. (2008). Generalized varying coefficient models for longitudinal data. *Biometrika* **95**, 653-666.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1348-1360.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1569.

Yuan, M. (2007). Nonnegative Garrote Component Selection in Functional ANOVA models. *Proceedings of the Eleventh Internat. Conf. on Artificial Intelligence and Statist.* **2**, 660-666.

Zhang, W. (2011). Identification of the constant components in generalised semivarying coefficient models by cross-validation. *Statist. Sinica* **21**, 1913-1929.

Zhang, W. and Peng, H. (2010). Simultaneous confidence band and hypothesis test in generalised varying-coefficient models. *J. Multivariate Anal.* **101**, 1656-1680.

Hasselt University Interuniversity Institute for Biostatistics and statistical Bioinformatics, CenStat Agoralaan building D 3590 - Diepenbeek, Belgium.
E-mail: anneleen.verhasselt@uhasselt.be

(Received July 2011; accepted November 2012)