Computational Statistics and Data Analysis 145 (2020) 106912

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A new approach to varying-coefficient additive models with longitudinal covariates

Xiaoke Zhang[a,*], Qixian Zhong[b], Jane-Ling Wang[c]

[a] Department of Statistics, George Washington University, Washington, DC 20052, USA
[b] Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China
[c] Department of Statistics, University of California, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 18 August 2019
Received in revised form 31 December 2019
Accepted 2 January 2020
Available online 15 January 2020

Keywords:
Additive model
B-splines
Dimension reduction
Empirical processes
Functional data analysis
Varying-coefficient model

ABSTRACT

The varying-coefficient additive model is a novel tool for analyzing functional data. The model generalizes both the varying-coefficient model and the additive model, and retains their merits as an effective dimension reduction model that is flexible yet easily interpretable. However, the original method only works for densely recorded functional response processes with time-invariant covariates. To broaden its applicability, the model is extended to allow for time-dependent covariates and a new fitting approach is proposed that can handle sparsely recorded functional response processes. Consistency and L^2 rate of convergence are developed for the proposed estimators of the unknown functions. A simple algorithm is developed that overcomes the computational difficulty caused by the non-convexity of the objective function. The proposed approach is illustrated through a simulation study and a real data application.

1. Introduction

In the analysis of functional and longitudinal data, the relationship between a time-varying response and a few covariates, each of which is either time-invariant or time-dependent, is of keen interest. Both varying-coefficient models (e.g., Brumback and Rice, 1998; Hoover et al., 1998; Fan and Zhang, 2000; Guo, 2002; Huang et al., 2002, 2004; Morris and Carroll, 2006; Şentürk and Nguyen, 2011) and additive models (e.g., Berhane and Tibshirani, 1998; Lin and Zhang, 1999; You and Zhou, 2007; Carroll et al., 2009; Xue et al., 2010; Scheipl et al., 2015; Luo et al., 2016; Luo and Qi, 2017; Qi and Luo, 2018) are widely used non-parametric modeling approaches to analyze functional and longitudinal data that enjoy flexibility and parsimony. Interested readers may refer to Fan and Zhang (2008), Park et al. (2015), Hastie and Tibshirani (1990), Wood (2017), and Stasinopoulos et al. (2017) for general introductions to varying-coefficient models and additive models. An intriguing question is how to choose between these two models in practice. In a recent article by Zhang and Wang (2015), it was shown that this dichotomy can be altogether bypassed by embedding both models into a larger model, the varying-coefficient additive model (VCAM), which includes both models as special cases. However, that work was specifically designed for densely observed functional response with vector covariates. In this article, we show how to extend the VCAM to more general settings that allow for sparsely observed functional responses (Yao et al., 2005; Li and Hsing, 2010), a.k.a. longitudinal data, and longitudinal covariates, in addition to vector covariates.

Denote the response by $X(t)$ which depends on time t, and the covariates by $\mathbf{Z}(t) = (Z_1(t), \dots, Z_d(t))$. Each covariate $Z_k, k = 1, \dots, d$, may also be time-invariant. Without loss of generality, we assume $t \in [a_0, b_0]$ and $Z_k \in [a_k, b_k]$,

[*] Corresponding author.
E-mail address: xkzhang@gwu.edu (X. Zhang).

https://doi.org/10.1016/j.csda.2020.106912
0167-9473/© 2020 Elsevier B.V. All rights reserved.

$k = 1, \ldots, d$. Let $m(t, \mathbf{Z}(t)) = E\{X(t) \mid \mathbf{Z}(t)\}$ be the regression function of interest. We consider the following VCAM for longitudinal data:

$$m(t, \mathbf{Z}(t)) = \beta_0(t) + \sum_{k=1}^{d} \beta_k(t)\phi_k(Z_k(t)), \tag{1}$$

where the β_k ($k = 1, \ldots, d$) are coefficient functions and the ϕ_k ($k = 1, \ldots, d$) are additive component functions, which are smooth (e.g., continuous or differentiable) and satisfy the following identifiability condition:

$$\frac{1}{b_0 - a_0}\int_{a_0}^{b_0} \beta_k(t)\,dt = 1, \quad \int_{a_k}^{b_k} \phi_k(z_k)\,dz_k = 0 \quad (k = 1, \ldots, d). \tag{2}$$

The proof of this condition is given in the supplementary material.

When the ϕ_k ($k = 1, \ldots, d$) are linear functions, VCAM (1) collapses to the varying-coefficient model for longitudinal or functional data:

$$m(t, \mathbf{Z}(t)) = \beta_0(t) + \sum_{k=1}^{d} \beta_k(t) Z_k(t), \tag{3}$$

with the abuse of notations. Therefore, the VCAM generalizes the varying-coefficient model in the sense that the effect of each covariate $Z_k(t)$ on the response may be non-linear through an unknown function ϕ_k.

Obviously, VCAM (1) also extends the classical additive model (Stone, 1985) since it is additive in $\mathbf{Z}(t)$ for any fixed time t. A special form of additive models that was considered by You and Zhou (2007) and Carroll et al. (2009) among others, is

$$m(t, \mathbf{Z}(t)) = \beta_0(t) + \sum_{k=1}^{d} \phi_k(Z_k(t)). \tag{4}$$

This additive model implicitly assumes that the influence of each covariate on the response only depends on its measured value, but not on the time when this covariate was measured, i.e., $\beta_k(t) = 1, k = 1, \ldots, d$ in VCAM (1). Thus, VCAM (1) is more flexible than both varying-coefficient models and additive models.

On the opposite spectrum, VCAM (1) may be regarded as a special case of the time-varying additive model (TVAM) in Zhang et al. (2013):

$$m(t, \mathbf{Z}(t)) = \mu_0(t) + \sum_{k=1}^{d} \mu_k(t, Z_k(t)), \tag{5}$$

where $\mu_k (k = 1, \ldots, d)$ are bivariate functions such that $\int_{a_k}^{b_k} \mu_k(t, z_k)\,dz_k = 0$. While TVAM (5) relaxes the aforementioned restrictive assumption of the additive model (4), the bivariate structure makes it less transparent to interpret the effect of each covariate. In contrast, the multiplicative form $\beta_k\phi_k$ ($k = 1, \ldots, d$) in VCAM (1) is an appealing feature, as it not only facilitates interpretability by separating the joint influence of covariates and time, but also involves only one-dimensional smoothing in contrast to TVAM (5), which requires two-dimensional smoothing. Thus, VCAM (1) is an effective dimension reduction model that enjoys flexibility, interpretability, and computational efficiency.

A special scenario of the VCAM, when $X(t)$ is densely observed functional data and the covariates $(Z_1, \ldots, Z_d)$ are time-independent covariates, was considered in Zhang and Wang (2015), who developed a very simple two-step estimation procedure for VCAM. Unfortunately, that approach does not work for sparsely sampled longitudinal responses or longitudinal covariates, whether densely or sparsely sampled. This provides the motivation for us to develop a new method to estimate the unknown components of VCAM (1) to address this deficit. The new method that is described in Section 2 differs substantially from the approach in Zhang and Wang (2015) and involves non-convex minimization of the objective function. A simple algorithm to address the non-convex minimization issue is proposed in Section 4 along with two methods of initial estimates for the algorithm that depend on the types of covariates.

The new estimation method necessitates a new theoretical framework, which is described in Section 3 and differs substantially from that of Zhang and Wang (2015). A key reason is that while a closed-form solution exists for the estimators in Zhang and Wang (2015), this is not the case for the estimators in Section 2. Rather, we employ nonparametric M-estimation approaches to derive the consistency and L^2 rate of convergence for the proposed estimators of unknown functions. Details of the proofs, which are based on the techniques of empirical processes, are available in the online supplement.

In the functional regression literature, the VCAM may be considered a functional concurrent model (e.g., Ramsay and Silverman, 2005; Zhang et al., 2013; Goldsmith and Schwartz, 2017; Hu et al., 2018), for which the influence of a covariate on the functional response is instantaneous. When all covariates are time-invariant, functional concurrent models are also called function-on-scalar regression models (e.g., Reiss et al., 2010; Goldsmith et al., 2015; Barber et al., 2017; Reimherr et al., 2019). Another type of approaches that aims at modeling the cumulative influence of the history of multiple functional covariates on the functional response has been considered in e.g., Scheipl et al. (2015, 2016) and Qi and Luo (2019). These approaches have theoretical and computational challenges because they involve an ill-posed problem of inverting a covariance operator but they are useful alternatives to the VCAM. For recent developments and future directions of functional regression, interested readers may refer to the survey papers by Morris (2015), Greven and Scheipl (2017), Paganoni and Sangalli (2017), and Reiss et al. (2017).

The remainder of the article proceeds as follows. Section 2 describes the proposed estimation procedure, and Section 3 includes the asymptotic result of each function estimator. Computational issues, including the algorithm, initial estimates, and tuning parameters, are addressed in Section 4. The numerical performance of the proposed method is examined through a simulation experiment in Section 5 and a real data application in Section 6. Section 7 concludes the article.

After completing the original version of this article, we noticed a related article by Hu et al. (2019) which was recently accepted by *Statistica Sinica*. This article is an independent work with a different estimation procedure and distinct theoretical proofs. We compare the two methods in Section 5 and demonstrate the superiority of our method in estimation and prediction accuracy.

2. Methodology

VCAM (1) may be alternatively expressed as

$$X(t) \mid \mathbf{Z}(t) = \beta_0(t) + \sum_{k=1}^{d} \beta_k(t)\phi_k(Z_k(t)) + W(t),$$

where β_k and $\phi_k, k = 1, \ldots, d$, satisfy (2) and W is the stochastic component of X, which is independent of $\mathbf{Z}$ and $E\{W(t)\} = 0$.

Suppose that we collect data from n independent subjects, i.e., $\{(X_i, \mathbf{Z}_i) : i = 1, \ldots, n\}$ are independent copies of $(X, \mathbf{Z})$, where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id})$. For the ith subject, the measurements are taken sparsely at irregularly time points $T_{ij}, j = 1, \ldots, n_i$, and the responses may be contaminated with random errors $e_{ij}, j = 1, \ldots, n_i$, which are independent copies of a random variable e with zero mean and finite variance. Thus we observe $\{(Y_{ij}, T_{ij}, \mathbf{Z}_{ij}) : i = 1, \ldots, n; j = 1, \ldots, n_i\}$, where $\mathbf{Z}_{ij} = \mathbf{Z}_i(T_{ij}) = (Z_{i1}(T_{ij}), \ldots, Z_{id}(T_{ij})) = (Z_{ij1}, \ldots, Z_{ijd})$ and

$$Y_{ij} = X_i(T_{ij}) + e_{ij} = \beta_0(T_{ij}) + \sum_{k=1}^{d} \beta_k(T_{ij})\phi_k(Z_{ijk}) + W_i(T_{ij}) + e_{ij}. \tag{6}$$

Following standard literature for functional data, we further assume that $\{T_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$ are independent copies of a random variable T and $\{\mathbf{Z}_i : i = 1, \ldots, n\}$ are independent of $\{T_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$. Therefore, $\{\mathbf{Z}_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$ are identical copies of $\mathbf{Z}(T)$, and $\{Y_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$ are identical copies of a random variable Y defined by $Y = m(T, \mathbf{Z}(T)) + W(T) + e$.

Theoretically we identify ϕ_k and β_k by minimizing

$$E\left\{Y - \beta_0(T) - \sum_{k=1}^{d} \beta_k(T)\phi_k(Z_k(T))\right\}^2,$$

subject to the constraints (2). An empirical version is to minimize

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{Y_{ij} - \beta_0(T_{ij}) - \sum_{k=1}^{d} \beta_k(T_{ij})\phi_k(Z_{ijk})\right\}^2, \tag{7}$$

where $N = \sum_{i=1}^{n} n_i$. Another empirical criterion inspired by Huang et al. (2004) and Li and Hsing (2010) is to assign equal weights to all subjects,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\left\{Y_{ij} - \beta_0(T_{ij}) - \sum_{k=1}^{d} \beta_k(T_{ij})\phi_k(Z_{ijk})\right\}^2.$$

Since the criterion (7), which assigns equal weights to each measurement, is generally preferable for longitudinal data following the work of Zhang and Wang (2016), we adopt this scheme hereafter.

We further use B-splines to approximate β_k $(k = 0, \ldots, d)$ and ϕ_k $(k = 1, \ldots, d)$. For each β_k, we use $K_{k,C}$ interior knots and $J_{k,C} = K_{k,C} + p_{k,C}$ B-spline basis functions of order $p_{k,C} \geq 1$, denoted by $\{B_{kl} : l = 1, \ldots, J_{k,C}\}$ as defined in Chapter IX of de Boor (2001). The subscript “C” refers to coefficient functions. Therefore, β_k is approximated by $\sum_{l=1}^{J_{k,C}} f_{kl}B_{kl}$ with a vector $\mathbf{f}_k = (f_{k1}, \ldots, f_{kJ_{k,C}})$.

Similarly, to approximate each ϕ_k, we use $K_{k,A}$ interior knots and $J_{k,A} = K_{k,A} + p_{k,A}$ B-spline basis functions of order $p_{k,A} \geq 1$, denoted by $\{N_{ks} : s = 1, \ldots, J_{k,A}\}$. The subscript “A” corresponds to additive component functions. Therefore, ϕ_k is approximated by $\sum_{s=1}^{J_{k,A}} g_{ks}N_{ks}$ with coefficients $\mathbf{g}_k = (g_{k1}, \ldots, g_{k,J_{k,A}})$.

Denote $\mathbf{f}=(\mathbf{f}_0,\mathbf{f}_1,\dots,\mathbf{f}_d)$ and $\mathbf{g}=(\mathbf{g}_1,\dots,\mathbf{g}_d)$. Combining (7) and the constraints (2), the estimates of $\mathbf{f}$ and $\mathbf{g}$, and thus the estimates of β_k and ϕ_k, denoted by $\hat{\beta}_k$ and $\hat{\phi}_k$, are obtained by solving the optimization: minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[Y_{ij}-\sum_{l=1}^{J_{0,C}} f_{0l}B_{0l}(T_{ij})-\sum_{k=1}^{d}\left\{\sum_{l=1}^{J_{k,C}} f_{kl}B_{kl}(T_{ij})\right\}\left\{\sum_{s=1}^{J_{k,A}} g_{ks}N_{ks}(Z_{ijk})\right\}\right]^2, \tag{8}$$

subject to

$$\frac{1}{b_0-a_0}\int_{a_0}^{b_0}\left\{\sum_{l=1}^{J_{k,C}} f_{kl}B_{kl}(t)\right\}dt=1,\ \int_{a_k}^{b_k}\left\{\sum_{s=1}^{J_{k,A}} g_{kl}N_{ks}(z_k)\right\}dz_k=0,$$

for $k=1,\dots,d$. We denote the estimated regression function by $\hat{m}=\hat{\beta}_0+\sum_{k=1}^{d}\hat{\beta}_k\hat{\phi}_k$.

3. Theoretical properties

In this section we provide the asymptotic results for $\hat{\beta}_k$, $k=0,\dots,d$, and $\hat{\phi}_k$, $k=1,\dots,d$, obtained from (8).

For simplicity and without loss of generality, we assume $[a_k,b_k]=[0,1]$ for all $k=0,\dots,d$. For any two $q-$ variate square-integrable functions $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$ defined on $[0,1]^q$, where $\mathbf{x}=(x_1,\dots,x_q)$ and $1\le q\le d+1$, define their L^2 inner product by $\langle m_1,m_2\rangle_{L^2}=\int m_1(\mathbf{x})m_2(\mathbf{x})\,d\mathbf{x}$, and denote the corresponding L^2 norm by $\|\cdot\|_{L^2}$. Moreover, denote the sup-norm by $\|\cdot\|_\infty$ that is defined by $\|m_1\|_\infty=\sup_{\mathbf{x}\in[0,1]^q}|m_1(\mathbf{x})|$.

For any two sequences $a(n),b(n)>0$ depending on n, the notation $a(n)\preceq b(n)$ (or $b(n)\succeq a(n)$) represents $\limsup_{n\to\infty}a(n)/b(n)<\infty$, and $a(n)\asymp b(n)$ means that $a(n)\preceq b(n)$ and $a(n)\succeq b(n)$.

Let $\mathcal{B}_k=\operatorname{span}\{B_{kl}:l=1,\dots,J_{k,C}\}$ $(k=0,\dots,d)$ and $\mathcal{N}_k=\operatorname{span}\{N_{ks}:s=1,\dots,J_{k,A}\}$ $(k=1,\dots,d)$. Define the approximation errors by

$$\rho_A=\max_{1\le k\le d}\inf_{\gamma_k\in\mathcal{N}_k}\|\phi_k-\gamma_k\|_\infty,\quad\text{and}\quad\rho_C=\max_{0\le k\le d}\inf_{\alpha_k\in\mathcal{B}_k}\|\beta_k-\alpha_k\|_\infty.$$

We first show that the estimated regression function $\hat{m}$ is consistent.

Theorem 1 (*Consistency*)**.** *Under Assumptions 1–6 in the appendix, if $\rho_A,\rho_C\to0$ as $n\to\infty$,*

$$\|\hat{m}-m\|_{L^2}=o_p(1),\quad \textit{as } n\to\infty.$$

Theorem 1 implies the consistency of each $\hat{\beta}_k$ and $\hat{\phi}_k$ as in the following corollary.

Corollary 1 (*Consistency*)**.** *Under Assumptions 1–6 in the appendix, if $\rho_A,\rho_C\to0$ as $n\to\infty$,*

$$\|\hat{\beta}_k-\beta_k\|_{L^2}=o_p(1)\ (k=0,\dots,d);$$
$$\|\hat{\phi}_k-\phi_k\|_{L^2}=o_p(1)\ (k=1,\dots,d).$$

We next establish the L^2 rates of convergence for the function estimates.

Theorem 2 (*L^2 Convergence*)**.** *Under Assumptions 1–7 in the appendix, if $\rho_A,\rho_C\to0$ as $n\to\infty$,*

$$\|\hat{m}-m\|_{L^2}=O_p\left\{\left(\frac{K_A+K_C}{n}\right)^{1/2}+\rho_A+\rho_C\right\},\quad \textit{as } n\to\infty.$$

Corollary 2 (*L^2 Convergence*)**.** *Under Assumptions 1–7 in the appendix, if $\rho_A,\rho_C\to0$ as $n\to\infty$,*

$$\|\hat{\beta}_k-\beta_k\|_{L^2}=O_p\left\{\left(\frac{K_A+K_C}{n}\right)^{1/2}+\rho_A+\rho_C\right\}\ (k=0,\dots,d);$$
$$\|\hat{\phi}_k-\phi_k\|_{L^2}=O_p\left\{\left(\frac{K_A+K_C}{n}\right)^{1/2}+\rho_A+\rho_C\right\}\ (k=1,\dots,d).$$

Theorem 2 and Corollary 2 indicate that each $\hat{\beta}_k$ and $\hat{\phi}_k$, together with $\hat{m}$, achieves the rate of convergence that is comparable with that of one-dimensional nonparametric smoothers. This may not be surprising considering the multiplicative form $\beta_k\phi_k$ $(k=1,\dots,d)$ assumed in VCAM (1). In particular, if $p_{k,C}=4$ $(k=0,\dots,d)$ and $p_{k,A}=4$ $(k=1,\dots,d)$, i.e., $\mathcal{B}_k$ $(k=0,\dots,d)$ and $\mathcal{N}_k$ $(k=1,\dots,d)$ are all cubic spline spaces, and all β_k $(k=0,\dots,d)$ and ϕ_k $(k=1,\dots,d)$ have bounded second derivatives, then $\rho_C\preceq K_C^{-2}$ and $\rho_A\preceq K_A^{-2}$ by Assumption 5 and Theorem 6.27 of Schumaker (2007). Additionally if $K_A\asymp K_C\asymp n^{1/5}$, the rate of convergence becomes optimal, which is $n^{2/5}$ (Stone, 1985).

4. Computational issues

4.1. Algorithm

The minimization in (8) is non-convex with respect to $(\mathbf{f}, \mathbf{g})$, which poses computational challenges to standard optimization methods. Fortunately, due to the special multiplicative structure of VCAM (1), we propose Algorithm 1 to circumvent this challenge.

Algorithm 1: The algorithm for fitting VCAM (1).

```
Obtain initial estimates $\hat\beta_k^{\text{ini}}$ $(k = 0, \ldots, d)$ and $\hat\phi_k^{\text{ini}}$ $(k = 1, \ldots, d)$.
Outer loop: Repeat until convergence.
    For $k = 1$ to $k = d$
    Inner loop: Repeat until convergence for each $k$.
        Find the latest residuals: $R_{ijk} = Y_{ij} - \hat\beta_0(T_{ij}) - \sum_{l\neq k} \hat\beta_l(T_{ij})\hat\phi_l(Z_{ijl})$.
        $\check\beta_k = \text{Smooth}(R_{ijk}, \hat\phi_k(Z_{ijk}), T_{ij})$; $\hat\beta_k = \check\beta_k / \int_{a_0}^{b_0} \check\beta_k$.
        $\check\phi_k = \text{Smooth}(R_{ijk}, \hat\beta_k(T_{ij}), Z_{ijk})$; $\hat\phi_k = \check\phi_k - \int_{a_k}^{b_k} \check\phi_k$.
    Output (inner loop): Return $\hat\beta_k$ and $\hat\phi_k$ at convergence.
    $\hat\beta_0 = \text{Smooth}(Y_{ij} - \sum_{k=1}^{d} \hat\beta_k(T_{ij})\hat\phi_k(Z_{ijk}), 1, T_{ij})$.
Output (outer loop): Return $\hat\beta_k$ $(k = 0, \ldots, d)$ and $\hat\phi_k$ $(k = 1, \ldots, d)$ at convergence.
```

In Algorithm 1, the generic notation "Smooth(R, U, V)" represents the B-spline estimator of V when fitting a multiplicative model $R \approx UV$ with response R and known covariate U. For the convergence criteria of the inner and outer loops respectively, define two mean residual squares as follows:

$$\text{MRS} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{Y_{ij} - \hat\beta_0(T_{ij}) - \sum_{k=1}^{d}\hat\beta_k(T_{ij})\hat\phi_k(Z_{ijk})\right\}^2,$$

$$\text{MRS}_k = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{R_{ijk} - \hat\beta_k(T_{ij})\hat\phi_k(Z_{ijk})\right\}^2.$$

The outer loop converges if the decrease of MRS is smaller than ϵ_1, while the inner loop for each $k = 1, \ldots, d$ converges if the decrease of MRS_k is smaller than ϵ_2. Here $\epsilon_1 > 0$ and $\epsilon_2 > 0$ are pre-specified small thresholds. In both Sections 5 and 6 we adopt $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = 10^{-2}$.

Remark 1. 1. Algorithm 1 is a modified backfitting algorithm (Breiman and Friedman, 1985). The *outer loop* is essentially the backfitting algorithm which updates each product $\beta_k\phi_k$ $(k = 1, \ldots, d)$ alternately, while the *inner loop* updates β_k and ϕ_k iteratively for each $k = 1, \ldots, d$. Due to the multiplicative form of $\beta_k\phi_k$, each iteration in the *inner loop* involves solving a convex (least squares) optimization, so the computation is fairly simple here. The main challenge of the non-convex optimization in (8) is to find good initial values and this will be discussed in Section 4.2.
2. Hu et al. (2019) used a different estimation procedure to fit VCAM (1) with three steps: first fit a two-dimensional additive model where each product $\eta_k(t, z_k) = \beta_k(t)\phi_k(z_k)$ $(k = 1, \ldots, d)$ is regarded as a general bivariate function, then find an initial estimate for each β_k, using the fact that $\beta_k(t)/\beta_k(t_0) = \eta_k(t, z_{k0})/\eta_k(t_0, z_{k0})$ if $\eta_k(t_0, z_{k0}) \neq 0$, and use this initial estimate to get the estimate $\hat\phi_k$, and finally update the initial estimates of β_k with the estimates $\hat\phi_k$ $(k = 1, \ldots, d)$ to get the final estimate of β_k. Different from Algorithm 1, the three-step estimation by Hu et al. (2019) does not involve backfitting or iterations and their theoretical proofs are developed for the three-step estimators.

4.2. Initialization

The initial estimates $\hat\beta_k^{\text{ini}}$ $(k = 0, \ldots, d)$ and $\hat\phi_k^{\text{ini}}$ $(k = 1, \ldots, d)$ are crucial for the convergence of Algorithm 1 and the performance of the final estimates. Depending on the types of covariates $\mathbf{Z}$, we propose two methods for initialization.

Initialization 1 (Time-Invariant Covariates). When all covariates are time-invariant, i.e., $Z_{ijk} = Z_{ik}$ for all $k = 1, \ldots, d$, we use a similar two-step estimation procedure in Zhang and Wang (2015) to obtain the initial estimates, which however requires densely recorded longitudinal responses to approximate the integrals $(b_0 - a_0)^{-1}\int Y(t)\,dt$. A regular additive model for independent data emerges after one integrates both sides of (1) since $(b_0 - a_0)^{-1}\int \beta_k(t)\,dt = 1$ $(k = 1, \ldots, d)$. Initial estimates $\hat\phi_k^{\text{ini}}$ $(k = 1, \ldots, d)$ can then be obtained by fitting such an additive model and $\hat\beta_k^{\text{ini}}$ $(k = 0, \ldots, d)$ are subsequently obtained by fitting a varying-coefficient model where $\hat\phi_k^{\text{ini}}(Z_{ik})$ $(k = 1, \ldots, d)$ are considered as covariates. This two-step procedure produces good estimates for dense longitudinal responses and does not require updating the estimates (Zhang and Wang, 2015). In our setting where the longitudinal response may be sparse, the integral $(b_0 - a_0)^{-1} \int Y(t)\,dt$ cannot be well approximated. Hence, this previous approach does not work to produce the final estimate but can be used to provide initial estimates. This approach to initializing Algorithm 1 turns out to be quite reliable in our experience.

Explicitly, for the ith subject, we sort the pairs $\{(Y_{ij}, T_{ij}) : j = 1, \ldots, n_i\}$ in the ascending order of T_{ij} such that they are now rearranged as $\{(Y^*_{ij}, T^*_{ij}) : j = 1, \ldots, n_i\}$ where $T^*_{i1} \le \cdots \le T^*_{i,n_i}$ and $Y^*_{ij} = Y_{ik}$ when $T_{ik} = T^*_{ij}$. Next, we proceed with the following two steps:

$$\begin{aligned}
\text{Step 1}: \tilde{Y}_i = {} & \frac{1}{b_0 - a_0} \left\{ \frac{1}{2} \sum_{j=1}^{n_i - 1} (Y^*_{ij} + Y^*_{i,j+1})(T^*_{i,j+1} - T^*_{ij}) \right. \\
& \left. + Y^*_{i1}(T^*_{i1} - a_0) + Y^*_{i,n_i}(b_0 - T^*_{i,n_i}) \right\}, \\
(\check{\beta}_0, \hat{\mathbf{g}}^{\text{ini}}) = {} & \underset{\tilde{\beta}_0, \mathbf{g}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \tilde{Y}_i - \tilde{\beta}_0 - \sum_{k=1}^{d} \sum_{s=1}^{J_{k,A}} g_{ks} N_{ks}(Z_{ik}) \right\}^2, \\
\hat{\phi}^{\text{ini}}_k = {} & \sum_{s=1}^{J_{k,A}} \hat{g}^{\text{ini}}_{ks} N_{ks} - \frac{1}{b_k - a_k} \int_{a_k}^{b_k} \left\{ \sum_{s=1}^{J_{k,A}} \hat{g}^{\text{ini}}_{ks} N_{ks}(z_k) \right\} dz_k \ (k = 1, \ldots, d). \\
\text{Step 2}: \hat{\mathbf{f}}^{\text{ini}} = {} & \underset{\mathbf{f}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ Y_{ij} - \sum_{l=1}^{J_{0,C}} f_{0l} B_{0l}(T_{ij}) - \sum_{k=1}^{d} \hat{\phi}^{\text{ini}}_k(Z_{ik}) \sum_{l=1}^{J_{k,C}} f_{kl} B_{kl}(T_{ij}) \right\}^2, \\
\hat{\beta}^{\text{ini}}_0 = {} & \sum_{l=1}^{J_{0,C}} \hat{f}^{\text{ini}}_{0l} B_{0l}, \ \hat{\beta}^{\text{ini}}_k = \frac{\sum_{l=1}^{J_{k,C}} \hat{f}^{\text{ini}}_{kl} B_{kl}}{\left[\frac{1}{b_0 - a_0} \int_{a_0}^{b_0} \left\{ \sum_{l=1}^{J_{k,C}} \hat{f}^{\text{ini}}_{kl} B_{kl}(t) \right\} dt \right]} \ (k = 1, \ldots, d).
\end{aligned}$$

Initialization 2 (Longitudinal Covariates). If at least one of the covariates is time-dependent, the joint effect of ϕ_k and β_k for all $k = 1, \ldots, d$ cannot be separated by the intra-subject integration of Y_{ij}, so the previous initialization approach will not produce reliable initial estimates. Therefore, we propose an alternative two-step initialization method for this setting:

$$\begin{aligned}
\text{Step 1}: (\check{\mathbf{f}}_0, \hat{\mathbf{g}}^{\text{ini}}) = {} & \underset{\mathbf{f}_0, \mathbf{g}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ Y_{ij} - \sum_{l=1}^{J_{0,C}} f_{0l} B_{0l}(T_{ij}) - \sum_{k=1}^{d} \sum_{s=1}^{J_{k,A}} g_{ks} N_{ks}(Z_{ijk}) \right\}^2, \\
\hat{\phi}^{\text{ini}}_k = {} & \sum_{s=1}^{J_{k,A}} \hat{g}^{\text{ini}}_{ks} N_{ks} - \frac{1}{b_k - a_k} \int_{a_k}^{b_k} \left\{ \sum_{s=1}^{J_{k,A}} \hat{g}^{\text{ini}}_{ks} N_{ks}(z_k) \right\} dz_k \ (k = 1, \ldots, d). \\
\text{Step 2}: \hat{\mathbf{f}}^{\text{ini}} = {} & \underset{\mathbf{f}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ Y_{ij} - \sum_{l=1}^{J_{0,C}} f_{0l} B_{0l}(T_{ij}) - \sum_{k=1}^{d} \hat{\phi}^{\text{ini}}_k(Z_{ijk}) \sum_{l=1}^{J_{k,C}} f_{kl} B_{kl}(T_{ij}) \right\}^2, \\
\hat{\beta}^{\text{ini}}_0 = {} & \sum_{l=1}^{J_{0,C}} \hat{f}^{\text{ini}}_{0l} B_{0l}, \ \hat{\beta}^{\text{ini}}_k = \frac{\sum_{l=1}^{J_{k,C}} \hat{f}^{\text{ini}}_{kl} B_{kl}}{\left[\frac{1}{b_0 - a_0} \int_{a_0}^{b_0} \left\{ \sum_{l=1}^{J_{k,C}} \hat{f}^{\text{ini}}_{kl} B_{kl}(t) \right\} dt \right]} \ (k = 1, \ldots, d).
\end{aligned}$$

The idea of this method is simple: first obtain $\hat{\phi}^{\text{ini}}_k$ $(k = 1, \ldots, d)$ by fitting an additive model where we set $\beta_k = 1$ $(k = 1, \ldots, d)$, and then obtain $\hat{\beta}^{\text{ini}}_k$ $(k = 0, \ldots, d)$ by fitting a varying-coefficient model where $\hat{\phi}^{\text{ini}}_k(Z_{ijk})$ $(k = 1, \ldots, d)$ are regarded as known covariates.

One could also adopt an alternative initialization by essentially interchanging Steps 1 and 2. That is, first estimate β_k $(k = 0, \ldots, d)$ by fitting a varying-coefficient model with $\phi_k(z_k) = z_k - (a_k + b_k)/2$ $(k = 1, \ldots, d)$ so that the constraints (2) are met, and then update the estimates of ϕ_k $(k = 1, \ldots, d)$ with known $\hat{\beta}_k$ $(k = 0, \ldots, d)$. Simulations not presented in this article showed that its performance is typically worse.

4.3. Knots selection

The performance of the function estimates relies on how many interior knots are selected and where they are positioned. Automated selection of the locations of knots is a hard problem so a common practice is to use equally spaced knots or equal number of observations between knots. For illustration and simplicity, we use equally spaced knots in numerical implementations so Assumption 3 is automatically met. It thus remains to properly select the number of knots.

Leave-one-curve-out cross-validation was suggested by Rice and Silverman (1991) as a data-driven approach for tuning parameter selection in functional and longitudinal data analysis and has been widely adopted. However, this could be computationally costly, so we opt for V-fold cross-validation which has been shown to be effective (e.g., Jiang and Wang, 2010; Chen and Lei, 2015; Wong and Zhang, 2019).

Explicitly, let $\mathbf{K} = \{K_{k,C} : k = 0, \dots, d\} \cup \{K_{k,A} : k = 1, \dots, d\}$ be the number of interior knots for all functions to be estimated. We select $\hat{\mathbf{K}}$ such that

$$\hat{\mathbf{K}} = \operatorname*{argmin}_{\mathbf{K}} \sum_{v=1}^{V} \sum_{i \in \mathcal{J}_v} \frac{1}{|\mathcal{J}_v|} \sum_{j=1}^{n_i} \left\{ Y_{ij} - \hat{m}_{\mathbf{K}}^{(-v)}(T_{ij}, \mathbf{Z}_{ij}) \right\}^2,$$

where $\mathcal{J}_v$ represents the index set for the vth fold with its size $|\mathcal{J}_v|$, and $\hat{m}_{\mathbf{K}}^{(-v)}$ represents the estimated regression function with all observations, but excluding $\mathcal{J}_v$.

5. Simulation

In this section we study the numerical performance of the proposed method in terms of estimation and prediction accuracy with and without model misspecification. We also compare our method with the one by Hu et al. (2019).

We considered $d = 2$, $[a_0, b_0] = [0, 2]$ and $[a_k, b_k] = [0, 1]$ $(k = 1, 2)$. The true functions were constructed in terms of cubic B-spline basis functions, i.e., $p_{0,C} = p_{1,C} = p_{2,C} = p_{1,A} = p_{2,A} = 4$. We positioned equidistance knots with $K_{0,C} = 4$, $K_{1,C} = 1$, $K_{2,C} = 2$, $K_{1,A} = 3$, and $K_{2,A} = 2$. The true functions that satisfy the identifiability condition (2) are:

$$\beta_0 = B_{01} + 2B_{02} + 4B_{03} + 3B_{04} - 2B_{05} + 3B_{07} + 6B_{08};$$

$$\beta_1 = \frac{\tilde{\beta}_1}{0.5\int_0^2 \tilde{\beta}_1(t)dt} = \tilde{\beta}_1/2.25, \quad \text{where} \quad \tilde{\beta}_1 = 5B_{12} + 3B_{13} + B_{14};$$

$$\beta_2 = \frac{\tilde{\beta}_2}{0.5\int_0^2 \tilde{\beta}_1(t)dt} = \tilde{\beta}_2/2, \quad \text{where} \quad \tilde{\beta}_2 = 6B_{22} + 2B_{23} + 3B_{25};$$

$$\phi_1 = \tilde{\phi}_1 - \int_0^1 \tilde{\phi}_1(z)\,dz = \tilde{\phi}_1 - 1, \quad \text{where} \quad \tilde{\phi}_1 = -2N_{12} + 5N_{15};$$

$$\phi_2 = \tilde{\phi}_2 - \int_0^1 \tilde{\phi}_2(z)\,dz = \tilde{\phi}_2 - 1.5, \quad \text{where} \quad \tilde{\phi}_2 = 4N_{23} + 2N_{24}.$$

Here we generated the true functions using B-spline basis functions in order to assess the cross-validation method for knot selection as in Section 4.3. These true functions are more general than low-degree polynomials, e.g., linear or quadratic polynomials, which are typically used in simulations.

We had $Q = 300$ simulation runs where $n = 50, 100, 200$ curves per run were generated. For each simulated data, $\{n_i : i = 1, \dots, n\}$ were independently generated from a discrete uniform distribution on $\{2, \dots, 10\}$, and $\{T_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$ were independently generated from a continuous uniform distribution on $[0, 2]$. We considered two settings for the covariates:

Time-Invariant **Z**: For $Z_{ijk} = Z_{ik}$ $(k = 1, 2)$, we generated $\{(Z_{i1}, Z_{i2}) : i = 1, \dots, n\}$ independently from a Gaussian copula with correlation parameter 0.6 using a MATLAB function “copula()”. Marginally both Z_{i1} and Z_{i2} follow a continuous uniform distribution on $[0, 1]$.

Longitudinal **Z**: First we generated independent $\{(U_{i1}, U_{i2}) : i = 1, \dots, n\}$ from a Gaussian copula with correlation 0.6, and independently generated $\{(V_{i1}, V_{i2}) : i = 1, \dots, n\}$ from a Gaussian copula with correlation 0.5. Then the two covariates were

$$Z_{ij1} = 0.5U_{i1}(0.5T_{ij})^{1/2} + 0.5V_{i1}, \quad Z_{ij2} = 0.5U_{i2}(0.5T_{ij})^{1/3} + 0.5V_{i2}.$$

Marginally both $Z_{ij1}, Z_{ij2} \in [0, 1]$.

We generated the stochastic component $W_{ij} = \sum_{l=1}^{4} A_{il}\gamma_l(T_{ij})$, where $\gamma_1(t) = 2^{1/2}\cos(2\pi t)$, $\gamma_2(t) = 2^{1/2}\sin(2\pi t)$, $\gamma_3(t) = 2^{1/2}\cos(4\pi t)$, and $\gamma_4(t) = 2^{1/2}\sin(4\pi t)$, and $\{A_{il} : i = 1, \dots, n\}$ were independently sampled from $N(0, \lambda_l)$ with $\lambda_l = 1/(l + 1)^2, l = 1, \dots, 4$. The random errors $\{e_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$ were independently generated from $N(0, \sigma^2)$ with $\sigma = 0.1$ and 0.4 respectively.

To generate the response, we considered four different true models: the VCAM, the additive model (AM), the varying-coefficient model (VCM), and the time-varying additive model (TVAM) in Zhang et al. (2013):

$$\text{Model 1 (VCAM)} : Y_{ij} = \beta_0(T_{ij}) + \sum_{k=1}^{d} \beta_k(T_{ij})\phi_k(Z_{ijk}) + W_{ij} + e_{ij};$$

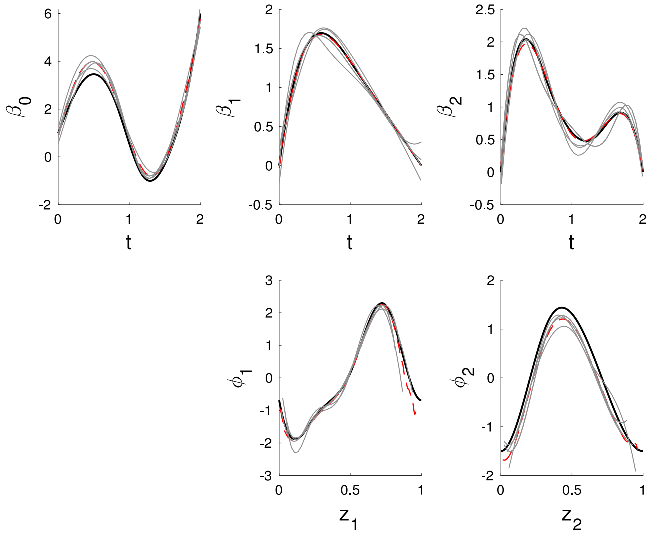

Fig. 1. Five randomly selected estimates out of 300 simulation runs and the mean function of all 300 estimates obtained by the proposed VCAM approach for the setting $n = 100$, $\sigma = 0.4$ when the true model is Model 1 with longitudinal covariates. Thick solid line: true function; dashed line: average function estimates; thin solid lines: five randomly selected estimates.

$$\text{Model 2 (AM)}: Y_{ij} = \beta_0(T_{ij}) + \sum_{k=1}^{d} \phi_k(Z_{ijk}) + W_{ij} + e_{ij};$$

$$\text{Model 3 (VCM)}: Y_{ij} = \tilde{\beta}_0(T_{ij}) + \sum_{k=1}^{d} \beta_k(T_{ij})(Z_{ijk} - 0.5) + W_{ij} + e_{ij};$$

$$\text{Model 4 (TVAM)}: Y_{ij} = \mu_0(T_{ij}) + \sum_{k=1}^{d} \mu_k(T_{ij}, Z_{ijk}) + W_{ij} + e_{ij},$$

where $\tilde{\beta}_0(t) = \beta_0(t)+0.5\beta_1(t)+0.5\beta_2(t)$, $\mu_0(t) = t^2-0.6\sin(2\pi t)-2+t\{(0.5t)^{1/2}/8+t/24+1/12\}+\cos(2\pi t)\{(0.5)^{1/3}/4+1/4\}$, $\mu_1(t, z_1) = t\{z_1^2 - (0.5t)^{1/2}/8 - t/24 - 1/12\}$, and $\mu_2(t, z_2) = \cos(2\pi t)\{z_2 - (0.5)^{1/3}/4 - 1/4\}$. Both Models 2 (AM) and 3 (VCM) are submodels of Model 1 (VCAM), while all of them are submodels of Model 4 (TVAM). All β_k and ϕ_k ($k = 1, \ldots, d$) in Models 1–3 satisfy the identifiability condition (2) (the covariates are centered in Model 3 such that $\int(z_k - 0.5)\,dz_k = 0$), so the function estimates obtained by fitting a VCAM are comparable with those obtained by fitting an AM or a VCM. so the additive component function estimates by fitting AM and VCAM respectively are comparable, while the coefficient function estimates by fitting VCM and VCAM respectively are comparable. This is why in Model 3 (VCM) the covariates are centered. Model 4 (TVAM) are not comparable with the other three models in terms of function estimation, so we only compare their prediction accuracy.

For Models 1–3, we compared the performance in estimation accuracy of fitting the VCAM by Algorithm 1, the VCAM by Hu et al. (2019), hereafter denoted as "HYY", the AM by B-spline smoothing (Stone, 1985), and the VCM by B-spline smoothing (Huang et al., 2002). For each function, we used 80% of the simulated data in each simulation run to obtain the estimates and the integrated squared errors (ISE) to evaluate its estimation accuracy. For example, with $\hat{\beta}_0^{[q]}$ obtained from the qth simulation run ($q = 1, \ldots, Q$), we calculated its $\text{ISE}^{[q]} = \int |\hat{\beta}_0^{[q]}(t) - \beta_0(t)|^2\, dt/(b_0 - a_0)$ and the median and absolute mean deviation of $\{\text{ISE}^{[q]} : q = 1, \ldots, Q\}$. For the AM fitting, only the estimates of β_0 and the additive component functions ϕ_k ($k = 1, \ldots, d$) are attainable and comparable with those obtained from the two VCAM fittings, while for the VCM fitting, only the estimates of the coefficient functions, i.e., $\tilde{\beta}_0$ and β_k ($k = 1, \ldots, d$), are attainable and comparable with those obtained from the two VCAM fittings. Model 4 is not comparable to the other models in terms of estimation so we only use it for comparisons in prediction.

To further illustrate whether the proposed VCAM approach can capture the shape of each target function, we provide in Fig. 1 the average of the 300 function estimates for each target when the true model is Model 1 for the case $n = 100$ and $\sigma = 0.4$ with longitudinal covariates. On average the proposed estimates show very small biases and can capture the shapes of the targets. A random sample of five estimates also shown in Fig. 1 illustrates the satisfactory performance of each individual estimate. Additional figures for longitudinal covariates with $n = 50$ and for time-invariant covariates with $n = 50$ and 100 are presented in the supplement.

For Models 1–4, we compared the performance in prediction accuracy of the four fittings above together with the TVAM fitting by smooth backfitting (Zhang et al., 2013). We used the mean squared prediction error (MSPE) of the estimated

Table 1
Estimation accuracy (time-invariant covariates) for the four methods: the proposed VCAM, HHY, AM and VCM fittings. The values of median and median absolute value (in the parentheses) of integrated squared errors are reported. In each simulation run, only 80% of the data were used to obtain the function estimates, so the actual sample sizes are 40, 160 for $n = 50, 200$.

True models			Model 1 (VCAM)				Model 2 (AM)		Model 3 (VCM)	
	n	σ	Proposed	HHY	AM	VCM	Proposed	AM	Proposed	VCM
β_0	50	0.1	0.0651 (0.0346)	1.0844 (1.1964)	0.1543 (0.0854)	2.8209 (19.0569)	0.0827 (0.0419)	0.0807 (0.0352)	0.0714 (0.0427)	0.0486 (0.0247)
		0.4	0.0751 (0.0426)	0.9624 (1.1928)	0.1644 (0.0923)	2.9509 (32.7028)	0.1428 (0.0576)	0.1375 (0.0552)	0.0818 (0.0469)	0.0587 (0.0293)
	200	0.1	0.0189 (0.0094)	0.0611 (0.0310)	0.0472 (0.0213)	0.1720 (0.3856)	0.0250 (0.0107)	0.0266 (0.0108)	0.0159 (0.0095)	0.0128 (0.0066)
		0.4	0.0210 (0.0105)	0.0441 (0.0202)	0.0487 (0.0222)	0.1479 (1.7489)	0.0281 (0.0121)	0.0299 (0.0122)	0.0210 (0.0107)	0.0164 (0.0076)
β_1	50	0.1	0.0213 (0.0223)	0.1051 (0.3387)	– –	0.1676 (95.9368)	0.0207 (0.0253)	– –	0.7568 (4.5252)	0.6532 (0.6794)
		0.4	0.0257 (0.0186)	0.1112 (0.6597)	– –	0.2043 (0.7121)	0.0266 (0.0169)	– –	1.0940 (24.3383)	0.9235 (1.3383)
	200	0.1	0.0051 (0.0033)	0.0094 (0.0124)	– –	0.0447 (0.0338)	0.0046 (0.0032)	– –	0.1810 (0.1278)	0.1849 (0.1078)
		0.4	0.0058 (0.0040)	0.0117 (0.0133)	– –	0.0500 (0.0321)	0.0058 (0.0038)	– –	0.2315 (0.1962)	0.2267 (0.1598)
β_2	50	0.1	0.0388 (0.0223)	0.1198 (0.3387)	– –	2.3982 (95.9368)	0.0331 (0.0253)	– –	0.9350 (4.5252)	0.6209 (0.6794)
		0.4	0.0421 (0.0304)	0.1416 (0.9765)	– –	2.0351 (167.3636)	0.0442 (0.0342)	– –	0.9865 (8.8343)	0.7403 (0.9251)
	200	0.1	0.0105 (0.0057)	0.0142 (0.0136)	– –	0.6252 (12.1283)	0.0080 (0.0055)	– –	0.1783 (0.1579)	0.1750 (0.1345)
		0.4	0.0124 (0.0068)	0.0156 (0.0256)	– –	0.7186 (15.1527)	0.0094 (0.0057)	– –	0.2230 (0.1758)	0.2115 (0.1613)
ϕ_1	50	0.1	0.0105 (0.0099)	1.0273 (0.2854)	0.0302 (0.0217)	– –	0.0122 (0.0135)	0.0099 (0.0101)	0.0061 (0.0074)	– –
		0.4	0.0111 (0.0109)	0.9193 (0.2708)	0.0306 (0.0215)	– –	0.0138 (0.0128)	0.0122 (0.0122)	0.0085 (0.0079)	– –
	200	0.1	0.0027 (0.0021)	0.6191 (0.3413)	0.0069 (0.0042)	– –	0.0030 (0.0024)	0.0029 (0.0024)	0.0022 (0.0016)	– –
		0.4	0.0084 (0.0077)	0.3128 (0.0993)	0.0200 (0.0186)	– –	0.0132 (0.0128)	0.0122 (0.0113)	0.0111 (0.0103)	– –
ϕ_2	50	0.1	0.0086 (0.0058)	0.2800 (0.0874)	0.0188 (0.0161)	– –	0.0083 (0.0070)	0.0075 (0.0069)	0.0075 (0.0043)	– –
		0.4	0.0063 (0.0038)	0.3240 (0.1615)	0.0120 (0.0083)	– –	0.0066 (0.0038)	0.0070 (0.0034)	0.0054 (0.0051)	– –
	200	0.1	0.0022 (0.0014)	0.2889 (0.1874)	0.0063 (0.0039)	– –	0.0026 (0.0018)	0.0026 (0.0017)	0.0021 (0.0014)	– –
		0.4	0.0032 (0.0020)	0.3052 (0.1889)	0.0083 (0.0043)	– –	0.0034 (0.0024)	0.0033 (0.0025)	0.0027 (0.0022)	– –

regression function to evaluate the prediction accuracy of a model fitting. In the qth simulation run ($q = 1, \ldots, Q$), we obtained a regression function estimate $\hat{m}^{[q]}$ using 80% of the randomly selected subjects as the training set, then calculated $\text{MSPE}^{[q]} = \text{ave}\{\sum_{i^* \in \mathcal{E}} \sum_{j=1}^{n_{i^*}} \{\hat{m}^{[q]}(T_{i^*j}, \mathbf{Z}_{i^*j}) - m(T_{i^*j}, \mathbf{Z}_{i^*j})\}^2 / N_{\mathcal{E}}\}$ where m is the true regression function and $\mathcal{E}$ represents the index set of the remaining 20% subjects as the test set with $N_{\mathcal{E}} = \sum_{i^* \in \mathcal{E}} n_{i^*}$, and finally obtained the median and absolute mean deviation of $\{\text{MSPE}^{[q]} : q = 1, \ldots, Q\}$.

Tables 1 and 2, corresponding to time-invariant and longitudinal covariates respectively, give the estimation accuracy results, for sample size $n = 50$ and 200, of four competing methods: the proposed VCAM, HHY, and the AM and VCM fittings. Generally, all estimators perform better as the sample size n increases or the error standard deviation σ decreases. Compared with the AM and VCM fittings, the performance of the proposed VCAM method is always superior when the model is correctly specified, i.e., under Model 1 (VCAM), and is very competitive when the model is misspecified, i.e., under Model 2 (AM) or Model 3 (VCM). This is not completely surprising, since both AM and VCM are submodels of VCAM, but it is still encouraging news for the proposed VCAM method. When the model is correctly specified, i.e., under Model 1 (VCAM), the proposed method always outperforms the HHY method substantially; the function estimates obtained by the misspecified AM and VCM fittings are sometimes better than those by the HHY method.

Table 3 gives the prediction accuracy results of five methods: the proposed VCAM, HHY, AM, VCM, and TVAM fittings. When the model is correctly specified, i.e., under Model 1 (VCAM), the proposed VCAM method is substantially better than all the other methods. Under the true Model 2 (AM) or Model 3 (VCM), the proposed VCAM fitting is not as good as the correctly specified one under the AM and VCM setting respectively. However, under Model 4 (TVAM), the proposed

Table 2
Estimation accuracy (longitudinal covariates) for the four methods: the proposed VCAM, HHY, AM and VCM fittings. The values of median and median absolute value (in the parentheses) of integrated squared errors are reported. In each simulation run, only 80% of the data were used to obtain the function estimates, so the actual sample sizes are 40, 160 for $n = 50, 200$.

True models			Model 1 (VCAM)				Model 2 (AM)		Model 3 (VCM)	
	n	σ	Proposed	HHY	AM	VCM	Proposed	AM	Proposed	VCM
β_0	50	0.1	0.1173 (0.0742)	1.3010 (85.8510)	0.1926 (0.1002)	24.5585 (8.4675)	0.2405 (0.1152)	0.1911 (0.0745)	0.1090 (0.1556)	0.1538 (0.1839)
		0.4	0.1321 (0.0980)	0.8483 (3.9681)	0.2467 (0.1104)	23.6962 (8.5085)	0.2409 (0.1131)	0.1859 (0.0712)	0.1122 (0.1703)	0.1567 (0.9286)
	200	0.1	0.0318 (0.0170)	0.1577 (0.2755)	0.1912 (0.0560)	0.6291 (3.0003)	0.0357 (0.0166)	0.0201 (0.0164)	0.0256 (0.0149)	0.0313 (0.0196)
		0.4	0.0412 (0.0202)	0.1869 (0.1918)	0.1989 (0.0515)	0.7069 (9.8924)	0.0406 (0.0205)	0.0245 (0.0131)	0.0303 (0.0162)	0.0393 (0.0341)
β_1	50	0.1	0.0296 (0.0260)	0.2625 (1.4618)	– –	0.0255 (0.0719)	0.0300 (0.0245)	– –	1.4647 (5.3743)	1.4053 (1.6408)
		0.4	0.0370 (0.0322)	0.1957 (0.9095)	– –	0.0641 (0.0663)	0.0313 (0.0242)	– –	1.4292 (5.9967)	1.3733 (1.7807)
	200	0.1	0.0085 (0.0056)	0.0263 (0.1137)	– –	0.0280 (0.0213)	0.0079 (0.0054)	– –	0.3724 (0.3899)	0.3944 (0.3055)
		0.4	0.0098 (0.0073)	0.0273 (0.0447)	– –	0.0272 (0.0218)	0.0094 (0.0072)	– –	0.5057 (0.5030)	0.4905 (0.4014)
β_2	50	0.1	0.0649 (0.0549)	0.3400 (1.6324)	– –	4.0624 (80.8300)	0.0552 (0.0451)	– –	1.4515 (8.5422)	1.2324 (1.5135)
		0.4	0.0862 (0.0789)	0.2989 (0.9516)	– –	5.2395 (109.2635)	0.0676 (0.0583)	– –	1.9443 (9.3415)	1.3985 (1.9495)
	200	0.1	0.0202 (0.0102)	0.0363 (0.0798)	– –	4.2656 (3.003)	0.0151 (0.0092)	– –	0.4648 (0.4321)	0.3917 (0.3095)
		0.4	0.0211 (0.0129)	0.0340 (0.0273)	– –	4.1240 (47.6214)	0.0159 (0.0115)	– –	0.4718 (0.4295)	0.4372 (0.3081)
ϕ_1	50	0.1	0.0090 (0.0082)	0.6753 (0.2430)	0.0255 (0.0256)	– –	0.0087 (0.0066)	0.0090 (0.0063)	0.0077 (0.0071)	– –
		0.4	0.0134 (0.0104)	0.5814 (0.1730)	0.0301 (0.0260)	– –	0.0121 (0.0079)	0.0117 (0.0073)	0.0084 (0.0079)	– –
	200	0.1	0.0074 (0.0053)	0.8133 (0.3581)	0.0358 (0.0213)	– –	0.0059 (0.0036)	0.0054 (0.0039)	0.0038 (0.0057)	– –
		0.4	0.0091 (0.0060)	0.7934 (0.3326)	0.0383 (0.0241)	– –	0.0070 (0.0053)	0.0074 (0.0049)	0.0048 (0.0053)	– –
ϕ_2	50	0.1	0.0045 (0.0044)	0.1172 (0.0472)	0.0107 (0.0118)	– –	0.0048 (0.0052)	0.0045 (0.0046)	0.0053 (0.0058)	– –
		0.4	0.0052 (0.0053)	0.0887 (0.0274)	0.0126 (0.0116)	– –	0.0081 (0.0074)	0.0073 (0.0069)	0.0062 (0.0073)	– –
	200	0.1	0.0043 (0.0030)	0.1663 (0.0689)	0.0199 (0.0133)	– –	0.0038 (0.0027)	0.0034 (0.0025)	0.0029 (0.0036)	– –
		0.4	0.0043 (0.0038)	0.1685 0.0643	0.0208 (0.0136)	– –	0.0042 (0.0033)	0.0042 (0.0030)	0.0046 (0.0050)	– –

VCAM fitting is comparable with, and sometimes better than, the TVAM fitting. The estimation and prediction results for $n = 100$ are given in the supplementary material and similar comparisons can be observed.

Next we demonstrate the performance of the cross-validation method to select the number of knots. For Model 1 (VCAM), a few summary statistics for the selected number of knots by cross-validation for the proposed method are given in Table 4. The number of knots selected by cross-validation for each function is very close to the true one for both time-invariant and longitudinal covariates.

In conclusion, this simulation study demonstrates the appealing performance of the proposed VCAM approach in both estimation and prediction with and without model misspecification, and the satisfactory performance of cross-validated knots selection.

6. Data application

We applied our method to a dataset from the US National Longitudinal Survey of Youth to study the wage trajectories of high school dropouts. The dataset is an illustrative example in Ch.6 of Singer and Willett (2003) and is available from the website of UCLA Statistical Consulting Group (https://stats.idre.ucla.edu/other/examples/alda/). This dataset contains 6,402 observations of hourly wages from $n = 888$ male subjects who left high school before graduation. The number of observations for each subject n_i varies from 1 to 13. TIME refers to the time (in years) associated with observed wages since entry into the labor force. We considered two covariates that might be related to wage, the highest grade of school

Table 3
Prediction accuracy for the five methods: the proposed VCAM, HHY, AM, VCM and TVAM fittings. The values of median and median absolute value (in the parentheses) of mean squared prediction errors are reported. In each simulation run, 80% of the data were used as the training set to obtain the regression function estimates and the remaining 20% were used as the test set for validation.

True models		Model 1 (VCAM)					Model 2 (AM)		Model 3 (VCM)		Model 4 (TVAM)	
Time-Invariant **Z**												
n	σ	Proposed	HHY	AM	VCM	TVAM	Proposed	AM	Proposed	VCM	Proposed	TVAM
50	0.1	0.1085 (0.0611)	2.5122 (1.3892)	0.8337 (0.3200)	2.3055 (1.1042)	0.6002 (0.3109)	0.1034 (0.0537)	0.0555 (0.0321)	0.1095 (0.0695)	0.0493 (0.0281)	0.2159 (0.1487)	0.1496 (0.0562)
	0.4	0.1232 (0.0683)	2.5034 (1.6080)	0.8220 (0.3934)	2.3918 (1.5195)	0.6274 (0.3917)	0.1253 (0.0612)	0.0668 (0.0388)	0.1313 (0.0837)	0.0633 (0.0329)	0.2570 (0.2576)	0.1769 (0.0626)
200	0.1	0.0277 (0.0087)	1.5303 (0.5433)	0.8171 (0.1277)	2.3642 (0.4717)	0.2244 (0.0441)	0.0258 (0.0082)	0.0170 (0.0059)	0.0246 (0.0080)	0.0177 (0.0058)	0.0413 (0.0164)	0.0557 (0.0184)
	0.4	0.0339 (0.0101)	1.6121 (0.5817)	0.8062 (0.1149)	2.4257 (0.4283)	0.2260 (0.0444)	0.0316 (0.0090)	0.0210 (0.0072)	0.0317 (0.0113)	0.0216 (0.0077)	0.0519 (0.0220)	0.0617 (0.0199)
Longitudinal **Z**												
n	σ	Proposed	HHY	AM	VCM	TVAM	Proposed	AM	Proposed	VCM	Proposed	TVAM
50	0.1	0.0930 (0.0443)	1.5771 (29.7635)	0.4344 (0.1750)	11.5085 (5.5338)	0.3312 (0.1352)	0.0820 (0.0404)	0.1033 (0.0250)	0.0907 (0.0554)	0.0435 (0.0952)	0.2002 (0.6109)	0.1346 (0.0664)
	0.4	0.1063 (0.0494)	1.5771 (29.7635)	0.4132 (0.1561)	0.5319 (0.2492)	0.3205 (0.1423)	0.0540 (0.0117)	0.0367 (0.0128)	0.1149 (0.0531)	0.0995 (0.2332)	0.2430 (0.6472)	0.1472 (0.0632)
200	0.1	0.0297 (0.0087)	1.0034 (0.4959)	0.5236 (0.0842)	0.5888 (0.1183)	0.1230 (0.0261)	0.0278 (0.0080)	0.0172 (0.0096)	0.0282 (0.0100)	0.0235 (0.0102)	0.0489 (0.0263)	0.0575 (0.0249)
	0.4	0.0334 (0.0102)	1.0071 (0.4317)	0.4912 (0.0770)	0.5838 (0.1026)	0.1306 (0.0234)	0.0310 (0.0100)	0.0210 (0.0083)	0.0323 (0.0103)	0.0285 (0.0162)	0.0558 (0.0271)	0.0603 (0.0297)

Table 4
Summary statistics for the number of knots selected by five-fold cross-validation for the proposed VCAM fitting. All values reported are based on 300 simulation runs and the 80% training data in each run, so the actual sample sizes are 40, 80, 160 for $n = 50, 100, 200$ respectively. Med: median; Std: standard deviation.

True number		β_0			β_1			β_2			ϕ_1			ϕ_2		
		4			1			2			3			2		
Time-Invariant **Z**																
n	σ	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std
50	0.1	3.00	3	0.89	1.89	2	0.82	1.86	2	0.82	3.18	3	0.73	2.04	2	0.99
	0.4	2.87	3	0.85	1.88	2	0.81	1.87	2	0.76	3.03	3	0.87	2.01	2	0.99
100	0.1	3.19	3	0.88	1.79	2	0.80	2.13	2	0.76	3.28	3	0.45	2.41	2	0.96
	0.4	3.07	3	0.85	1.84	2	0.81	1.91	2	0.76	3.30	3	0.48	2.37	2	0.97
200	0.1	3.35	4	0.90	1.86	2	0.81	2.06	2	0.72	3.14	3	0.35	2.70	3	0.76
	0.4	3.37	4	0.88	1.81	2	0.81	2.11	2	0.74	3.14	3	0.35	2.66	3	0.89
Longitudinal **Z**																
n	σ	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std	Mean	Med	Std
50	0.1	2.82	3	0.92	1.78	2	0.78	1.84	2	0.79	2.34	2	1.11	1.81	2	0.89
	0.4	2.78	3	0.94	1.84	2	0.81	1.82	2	0.77	2.26	2	1.08	1.89	2	1.02
100	0.1	3.00	3	0.91	1.84	2	0.81	1.93	2	0.78	2.79	3	0.97	2.16	2	1.02
	0.4	2.95	3	0.91	1.86	2	0.82	1.94	2	0.80	2.67	3	1.03	2.15	2	1.08
200	0.1	3.42	4	0.87	1.92	2	0.83	1.96	2	0.76	3.25	3	0.71	2.38	2	1.03
	0.4	3.4167	4	0.85	1.84	2	0.81	1.98	2	0.79	3.19	3	0.72	2.36	2	1.01

completed and the local unemployment rate for the year of survey, denoted by HGC and UER respectively. HGC is time-invariant while UER is longitudinal. To be consistent with the analysis in Singer and Willett (2003), we took the natural logarithm of wages as the response, denoted by LNW, and fitted the following VCAM:

$$\mathrm{LNW}_{ij} \approx \beta_0\left(\mathrm{TIME}_{ij}\right) + \beta_1\left(\mathrm{TIME}_{ij}\right)\phi_1\left(\mathrm{HGC}_i\right) + \beta_2\left(\mathrm{TIME}_{ij}\right)\phi_2\left(\mathrm{UER}_{ij}\right).$$

We performed Algorithm 1 with the same specifications as in Section 5. The numbers of knots selected by the five-fold cross-validation are $K_{0,C} = 2$, $K_{1,C} = 1$, $K_{2,C} = 2$, $K_{1,A} = 3$, and $K_{2,A} = 1$.

The function estimates are shown in Fig. 2. The $\hat{\beta}_0$ curve indicates that overall wages increase over time, which conforms with common sense and the results in Murnane et al. (1999) and Singer and Willett (2003). In the plot for either ϕ_1 or ϕ_2, the reference line for the constant zero is not entirely contained in the region between the confidence band. This suggests that both ϕ_1 and ϕ_2 may be significantly different from zero at the 5% level, which implies that HGC and UER are both significant covariates. Similarly, both β_1 and β_2 may be significantly different from being a constant (one in this case) at the 5% level. This suggests time-varying effects of the covariates and that a conventional additive model, such as (4), has a lack of fit. Since the confidence band for ϕ_2 does not cover a linear function that integrates to

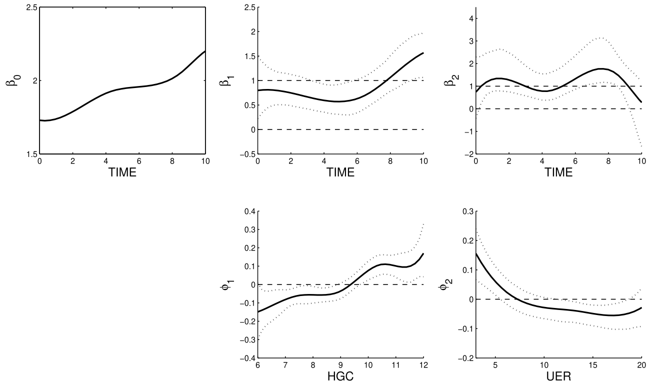

Fig. 2. Results of fitting VCAM for the wage data. Solid line: function estimate; dotted line: upper or lower bound of 95% pointwise bootstrap confidence band; dashed line: reference for a constant, either zero or one.

zero, the conventional varying coefficient model (3) may not be a good fit for this data either. These conclusions would be stronger if they were based on simultaneous confidence bands, of which development requires substantial theoretical work and is beyond the scope of this article.

The overall increasing pattern of $\hat{\phi}_1$, together with the positivity of $\hat{\beta}_1$, indicates that with the same years of working experience, people who completed higher grades before dropping out earned higher wages on average. This may not be surprising at a first glance as the benefit of education to wages is well known or at least expected. However, we show that even among those who dropped out of high school, more education is associated with higher wages. Moreover, the product $\hat{\beta}_1\hat{\phi}_1$ reveals that the effect of HGC on wages is time-dynamic since β_1 is not a constant function: the influence of HGC on wages is initially steady since the entry of the labor force, but it amplifies dramatically after six years (cf. the plot for β_1 in Fig. 2). In particular, for dropouts with HGC ≤ 9 and those with HGC ≥ 10, the difference in their income grows rapidly with labor force experience (cf. the plots for ϕ_1 and β_1 in Fig. 2).

The observation that $\hat{\beta}_2$ is positive and $\hat{\phi}_2$ is a decreasing function demonstrates that with the same years since the entry of labor force, on average people in areas with a higher local unemployment rate earn less than those in areas with a lower unemployment rate. This association is referred to as the "wage curve" in economics (e.g., Blanchflower and Oswald, 1994). Since β_2 is not a constant function, UER has a time-varying influence on wages, i.e., the effect of the unemployment rate on hourly wages depends on not only the value of UER but also the time when UER is observed.

This application illustrates the applicability and advantages of VCAM (1) for longitudinal data. Apart from easy interpretations, the VCAM is able to help identify the time-dynamic association between HGC, UER and LNW, which neither the varying-coefficient model (3) nor the additive model (4) can adequately capture. Fitting a VCAM also reveals interesting interactive effects between time and covariates as discussed above, which were ignored in previous studies based on this dataset, e.g., by Murnane et al. (1999) and Singer and Willett (2003).

7. Discussion

In this article we consider the VCAM for functional data where the data may be sparsely measured and/or longitudinal covariates are present. Compared to Zhang and Wang (2015) who handled dense functional data with vector covariates, we propose a new estimation procedure, and develop the consistency and L^2 rate of convergence for each estimated function. To tackle the non-convex minimization issue, we take advantage of the multiplicative structure of the VCAM and provide a modified backfitting algorithm. The appealing numerical performance in both estimation and prediction, subject to minor model misspecification or not, is demonstrated by a simulation study. The desirable interpretability of the VCAM is also illustrated through a real data application. While we focus on continuous covariates in this article, the proposed approach can easily be modified to handle discrete covariates by choosing a pre-specified additive component function for each discrete covariate. Likewise, the proposed approach can also be modified to accommodate any pre-specified additive component or coefficient function.

This article, together with Zhang and Wang (2015), provides the foundation for several future research topics regarding the VCAM. One future direction is to extend the VCAM to incorporate the influence of functional covariate histories on the response. Like function-on-function regression, the extension may involve theoretical and computational challenges associated with the inversion of a compact covariance operator. Another future direction is the development of simultaneous confidence bands.

Acknowledgments

The research of Xiaoke Zhang is supported by the USA National Science Foundation (NSF) grant DMS-1832046. The research of Qixian Zhong is supported by National Natural Science Foundation of China grant NSFC-11771241 and Chinese

Government Scholarship grant CSC-201806210163. The research of Jane-Ling Wang is supported by the USA NSF grant DMS-1512975 and DMS-1914917. The authors thank the Co-Editors, an Associate Editor, and two reviewers for helpful comments.

Appendix

A.1. Assumptions

For simplicity and without loss of generality we assume that $[a_k, b_k] = [0, 1]$ for all $k = 0, \dots, d$. Below is a list of assumptions to establish the asymptotic properties for $\hat{m}$, $\hat{\beta}_k$ and $\hat{\phi}_k$ obtained from (8).

Assumption 1. The joint density of $(T, \mathbf{Z}(T))$, denoted by $p(t, \mathbf{z})$, where $\mathbf{z} = (z_1, \dots, z_d)$, is bounded below and above on its domain $[0, 1]^{d+1}$, i.e.,

$$0 < m_p \leq \inf_{(t,\mathbf{z})\in[0,1]^{d+1}} p(t, \mathbf{z}) \leq \sup_{(t,\mathbf{z})\in[0,1]^{d+1}} p(t, \mathbf{z}) \leq M_p < \infty.$$

Assumption 2. All n_i are bounded, i.e., $n_i \leq M < \infty$ for all $i = 1, \dots, n$.

Assumption 3. The knots for ϕ_k, $k = 1, \dots, d$, are

$$0 = \tau_{k,1-p_{k,A}} = \cdots = \tau_{k,0} < \tau_{k,1} < \cdots < \tau_{k,K_{k,A}} < \tau_{k,K_{k,A}+1} = \cdots = \tau_{k,K_{k,A}+p_{k,A}} = 1,$$

and the knots for β_k, $k = 0, \dots, d$, are

$$0 = \zeta_{k,1-p_{k,C}} = \cdots = \zeta_{k,0} < \zeta_{k,1} < \cdots < \zeta_{k,K_{k,C}} < \zeta_{k,K_{k,C}+1} = \cdots = \zeta_{k,K_{k,C}+p_{k,C}} = 1.$$

They have bounded mesh ratios,

$$\limsup_{n\to\infty} \max_{1\leq k\leq d} \frac{\max_{1\leq l\leq K_{k,A}+1}\{\tau_{kl} - \tau_{k,l-1}\}}{\min_{1\leq l\leq K_{k,A}+1}\{\tau_{kl} - \tau_{k,l-1}\}} < \infty,$$

$$\limsup_{n\to\infty} \max_{0\leq k\leq d} \frac{\max_{1\leq l\leq K_{k,C}+1}\{\zeta_{kl} - \zeta_{k,l-1}\}}{\min_{1\leq l\leq K_{k,C}+1}\{\zeta_{kl} - \zeta_{k,l-1}\}} < \infty.$$

Assumption 4. All $p_{k,A}$, $k = 1, \dots, d$ and $p_{k,C}$, $k = 0, \dots, d$ are bounded:

$$\max\left\{\max_{1\leq k\leq d} p_{k,A}, \max_{0\leq k\leq d} p_{k,C}\right\} \leq p_{AC} < \infty.$$

Assumption 5. For $K_A = \max_{1\leq k\leq d} K_{k,A}$ and $K_C = \max_{0\leq k\leq d} K_{k,C}$, we assume

$$\limsup_{n\to\infty} \frac{K_A}{\min_{1\leq k\leq d} K_{k,A}} < \infty, \quad \text{and} \quad \limsup_{n\to\infty} \frac{K_C}{\min_{0\leq k\leq d} K_{k,C}} < \infty.$$

Assumption 6. $K_A, K_C \to \infty$ and $(\log K_A + \log K_C)K_A K_C/n \to 0$.

Assumption 7. There exists a constant $0 < L < \infty$ such that
(a) The error e satisfies $E\exp(be) \leq \exp\left(Lb^2\right)$ for any scalar b;
(b) For any number of arbitrary time points $t_1, \dots, t_p \in [0, 1]$, the random vector $\mathbf{W}_p = (W(t_1), \dots, W(t_p))^\top$ satisfies $E\exp\left(\mathbf{b}^\top\mathbf{W}_p\right) \leq \exp\left(pL\mathbf{b}^\top\mathbf{b}\right)$ for any vector $\mathbf{b} = (b_1, \dots, b_p)^\top$.

Assumptions 1 and 2 are typical in the literature on smoothing (e.g., Stone, 1985; Wang and Yang, 2007) and on longitudinal data analysis respectively. Assumptions 3 and 4 are standard for B-spline methods. Assumption 5, also used by Huang et al. (2004) and Zhang and Wang (2015), implies that the numbers of knots for all ϕ_k, $k = 1, \dots, d$ are of the same order, and likewise for all β_k, $k = 0, \dots, d$, which will be used to prove Property 2 in the supplementary material. Assumption 6 will be useful to prove Lemma 2 in the supplement. Assumption 7 indicates that e is a sub-Gaussian random variable and $\mathbf{W}_p$ is a sub-Gaussian random vector. This assumption is common when using the technique of empirical processes to obtain a convergence rate (e.g., van der Vaart and Wellner, 1996; van de Geer, 2000). Obviously Assumption 7(a) is automatically satisfied if e is Gaussian. A sufficient condition for Assumption 7(b) is that $\{W(t) : t \in [0, 1]\}$ is a Gaussian process with $\sup_{t\in[0,1]} \mathrm{Var}\{W(t)\} < \infty$ (see the proof in the Appendix A.2).

A.2. Sufficient condition for Assumption 7(b)

Proposition 1. *Assumption 7(b) holds if* $\{W(t) : t \in [0,1]\}$ *is a Gaussian process with* $\sup_{t\in[0,1]} Var\{W(t)\} < \infty$.

Proof. Denote the eigenvalues of the covariance matrix $\mathrm{Cov}(\mathbf{W}_p)$ in the descending order by $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. For arbitrary $\mathbf{b} = (b_1, \ldots, b_p)^\top$, we have

$$\begin{aligned} E\left\{\exp\left(\mathbf{b}^\top \mathbf{W}_p\right)\right\} &= \exp\left\{\frac{1}{2}\mathbf{b}^\top \mathrm{Cov}(\mathbf{W}_p)\mathbf{b}\right\} \leq \exp\left\{\frac{1}{2}\lambda_1 \mathbf{b}^\top \mathbf{b}\right\} \\ &\leq \exp\left\{\frac{1}{2}\left(\sum_{l=1}^p \lambda_l\right)\mathbf{b}^\top \mathbf{b}\right\} = \exp\left[\frac{1}{2}\mathrm{tr}\left\{\mathrm{Cov}(\mathbf{W}_p)\right\}\mathbf{b}^\top \mathbf{b}\right] \\ &\leq \exp\left(\frac{p}{2}\left[\sup_{t\in[0,1]} \mathrm{Var}\{W(t)\}\right]\mathbf{b}^\top \mathbf{b}\right) \leq \exp\left(pL\mathbf{b}^\top \mathbf{b}\right), \end{aligned}$$

if we choose $L \geq \sup_{t\in[0,1]} \mathrm{Var}\{W(t)\}/2$. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2020.106912. An online supplementary material document is given that includes the proof of the identifiability condition (2), the proofs for the theoretical results in Section 3, and additional simulation results.

References

Barber, R.F., Reimherr, M., Schill, T., 2017. The function-on-scalar LASSO with applications to longitudinal GWAS. Electron. J. Stat. 11 (1), 1351–1389.

Berhane, K., Tibshirani, R.J., 1998. Generalized additive models for longitudinal data. Canad. J. Statist. 26, 517–535.

Blanchflower, D.G., Oswald, A.J., 1994. The Wage Curve. MIT press.

Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc. 80 (391), 580–598.

Brumback, B.A., Rice, J.A., 1998. Smoothing spline models for the analysis of nested and crossed samples of curves. J. Amer. Statist. Assoc. 93 (443), 961–976.

Carroll, R.J., Maity, A., Mammen, E., Yu, K., 2009. Nonparametric additive regression for repeatedly measured data. Biometrika 96, 383–398.

Chen, K., Lei, J., 2015. Localized functional principal component analysis. J. Amer. Statist. Assoc. 110 (511), 1266–1275.

de Boor, C., 2001. A Practical Guide to Splines, Vol. 27. Springer Verlag.

Fan, J., Zhang, J.-T., 2000. Two-step estimation of functional linear models with applications to longitudinal data. J. R. Stat. Soc. Ser. B Stat. Methodol. 62 (2), 303–322.

Fan, J., Zhang, W., 2008. Statistical methods with varying coefficient models. Stat. Interface 1 (1), 179.

Goldsmith, J., Schwartz, J.E., 2017. Variable selection in the functional linear concurrent model. Stat. Med. 36 (14), 2237–2250.

Goldsmith, J., Zipunnikov, V., Schrack, J., 2015. Generalized multilevel function-on-scalar regression and principal component analysis. Biometrics 71 (2), 344–353.

Greven, S., Scheipl, F., 2017. A general framework for functional regression modelling. Stat. Model. 17 (1–2), 1–35.

Guo, W., 2002. Functional mixed effects models. Biometrics 58 (1), 121–128.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models, Vol. 43. Chapman and Hall, pp. 205–208.

Hoover, D., Rice, J., Wu, C., Yang, L., 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 85 (4), 809–822.

Hu, L., Huang, T., You, J., 2018. Estimation and identification of a varying-coefficient additive model for locally stationary processes. J. Amer. Statist. Assoc. 114 (527), 1191–1204.

Hu, L., Huang, T., You, J., 2019. Robust inference in varying-coefficient additive models for longitudinal/functional data. Statist. Sinica (in press).

Huang, J., Wu, C., Zhou, L., 2002. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 89 (1), 111–128.

Huang, J.Z., Wu, C.O., Zhou, L., 2004. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statist. Sinica 14 (3), 763–788.

Jiang, C.-R., Wang, J.-L., 2010. Covariate adjusted functional principal components analysis for longitudinal data. Ann. Statist. 38, 1194–1226.

Li, Y., Hsing, T., 2010. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. Ann. Statist. 38 (6), 3321–3351.

Lin, X., Zhang, D., 1999. Inference in generalized additive mixed models by using smoothing splines. J. R. Stat. Soc. Ser. B Stat. Methodol. 61, 381–400.

Luo, R., Qi, X., 2017. Function-on-function linear regression by signal compression. J. Amer. Statist. Assoc. 112 (518), 690–705.

Luo, R., Qi, X., Wang, Y., 2016. Functional wavelet regression for linear function-on-function models. Electron. J. Stat. 10 (2), 3179–3216.

Morris, J.S., 2015. Functional regression. Annu. Rev. Stat. Appl. 2 (1), 321–359.

Morris, J.S., Carroll, R.J., 2006. Wavelet-based functional mixed models. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (2), 179–199.

Murnane, R.J., Willett, J.B., Boudett, K.P., 1999. Do male dropouts benefit from obtaining a GED, postsecondary education, and training? Eval. Rev. 23 (5), 475–503.

Paganoni, A.M., Sangalli, L.M., 2017. Functional regression models: Some directions of future research. Stat. Model. 17 (1–2), 94–99.

Park, B.U., Mammen, E., Lee, Y.K., Lee, E.R., 2015. Varying coefficient regression models: a review and new developments. Internat. Statist. Rev. 83 (1), 36–64.

Qi, X., Luo, R., 2018. Function-on-function regression with thousands of predictive curves. J. Multivariate Anal. 163, 51–66.

Qi, X., Luo, R., 2019. Nonlinear function-on-function additive model with multiple predictor curves. Statist. Sinica 29 (2), 719–739.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis. Springer, New York.

Reimherr, M., Sriperumbudur, B., Kang, H.B., 2019. Optimal function-on-scalar regression over complex domains. arXiv preprint arXiv:1902.07284.

Reiss, P.T., Goldsmith, J., Shang, H.L., Ogden, R.T., 2017. Methods for scalar-on-function regression. Internat. Statist. Rev. 85 (2), 228–249.

Reiss, P.T., Huang, L., Mennes, M., 2010. Fast function-on-scalar regression with penalized basis expansions. Int. J. Biostat. 6 (1).

Rice, J., Silverman, B., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. J. R. Stat. Soc. Ser. B Stat. Methodol. 53 (1), 233–243.

Scheipl, F., Gertheiss, J., Greven, S., 2016. Generalized functional additive mixed models. Electron. J. Stat. 10 (1), 1455–1492.

Scheipl, F., Staicu, A.-M., Greven, S., 2015. Functional additive mixed models. J. Comput. Graph. Statist. 24 (2), 477–501.

Schumaker, L., 2007. Spline Functions: Basic Theory. Cambridge University Press.

Şentürk, D., Nguyen, D.V., 2011. Varying coefficient models for sparse noise-contaminated longitudinal data. Statist. Sinica 21 (4), 1831.

Singer, J.D., Willett, J.B., 2003. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press.

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F., 2017. Flexible Regression and Smoothing: Using GAMLSS in R. Chapman and Hall/CRC.

Stone, C.J., 1985. Additive regression and other nonparametric models. Ann. Statist. 13, 689–705.

van de Geer, S.A., 2000. Empirical Processes in M-Estimation. Cambridge University Press.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York.

Wang, L., Yang, L., 2007. Spline-backfitted kernel smoothing of nonlinear additive autoregression model. Ann. Statist. 35 (6), 2474–2503.

Wong, R.K.W., Zhang, X., 2019. Nonparametric operator-regularized covariance function estimation for functional data. Comput. Statist. Data Anal. 131, 131–144.

Wood, S.N., 2017. Generalized Additive Models: An Introduction With R, second ed. Chapman and Hall/CRC.

Xue, L., Qu, A., Zhou, J., 2010. Consistent model selection for marginal generalized additive model for correlated data. J. Amer. Statist. Assoc. 105 (492), 1518–1530.

Yao, F., Müller, H.-G., Wang, J.-L., 2005. Functional data analysis for sparse longitudinal data. J. Amer. Statist. Assoc. 100 (470), 577–590.

You, J., Zhou, H., 2007. Two-stage efficient estimation of longitudinal nonparametric additive models. Statist. Probab. Lett. 77, 1666–1675.

Zhang, X., Park, B.U., Wang, J.-L., 2013. Time-varying additive models for longitudinal data. J. Amer. Statist. Assoc. 108 (503), 983–998.

Zhang, X., Wang, J.-L., 2015. Varying-coefficient additive models for functional data. Biometrika 102 (1), 15–32.

Zhang, X., Wang, J.-L., 2016. From sparse to dense functional data and beyond. Ann. Statist. 44 (5), 2281–2321.