TEST (2020) 29:282–306
https://doi.org/10.1007/s11749-019-00655-5

ORIGINAL PAPER

Oracally efficient estimation for dense functional data with holiday effects

Li Cai[1] · Lisha Li[2] · Simin Huang[2] · Liang Ma[2] · Lijian Yang[3]

Received: 13 April 2018 / Accepted: 14 April 2019 / Published online: 20 April 2019

Abstract

Existing functional data analysis literature has mostly overlooked data with spikes in mean, such as weekly sporting goods sales by a salesperson which spikes around holidays. For such functional data, two-step estimation procedures are formulated for the population mean function and holiday effect parameters, which correspond to the population sales curve and the spikes in sales during holiday times. The estimators are based on spline smoothing for individual trajectories using non-holiday observations, and are shown to be oracally efficient in the sense that both the mean function and holiday effects are estimated as efficiently as if all individual trajectories were known a priori. Consequently, an asymptotic simultaneous confidence band is established for the mean function and confidence intervals for holiday effects, respectively. Two sample extensions are also formulated and simulation experiments provide strong evidence that corroborates the asymptotic theory. Application to sporting goods sales data has led to a number of new discoveries.

Keywords B-spline · Dummy variables · Functional data · Holiday effects · Oracle efficiency · Simultaneous confidence band

Mathematics Subject Classification 62M10 · 62G08 · 62P20

Li Cai, Lisha Li: Co-first authors.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11749-019-00655-5) contains supplementary material, which is available to authorized users.

✉ Lijian Yang
yanglijian@tsinghua.edu.cn; yanglijian@mail.tsinghua.edu.cn

1 School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China

2 Department of Industrial Engineering, Tsinghua University, Beijing, China

3 Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing, China

1 Introduction

Functional data, also known as curve data, are commonly observed in various applications, such as food processing, meteorological and environmental studies, medical research, as well as other fields. There is a vast literature on the functional data, ranging from dense functional data: Cardot (2000), Rice and Wu (2001), Hall et al. (2006), Fan et al. (2007), Degras (2011), Cao et al. (2012, 2016); sparse functional data: James et al. (2000), James and Sugar (2003), Yao et al. (2005), Ma et al. (2012), Zheng et al. (2014) and Gu et al. (2014), to name a few. The book of Zhang (2013) provides a comprehensive review on functional data analysis.

While functional data in the aforementioned literature come in the form of random curves recorded at discrete points with measurement errors, other complications may occur as well, such as spikes in mean, which is also the main focus of this paper. Consider, for example, the weekly sales from August 2, 2015, to July 31, 2016, of 74 salespersons employed by a sportswear retailing company located in Shanghai, China, which are discussed in detail in Sect. 4.2. To make scientific decision on employee compensation, the company's management needs to have some reference on a typical salesperson's performance over time, which is called the learning curve. Since all salespersons work on one-day-on-duty-one-day-off schedule and often take sick leaves and other days off, daily sales data are irregular, incomplete and spurious; hence, we have used the completely regular and less noisy weekly sales data. Figure 1a shows the weekly sales of 10 randomly selected salespersons for the 52 consecutive weeks, and Fig. 1b average weekly sales of the 74 salespersons. The average weekly sales exhibits visible spikes for weeks {9, 10, 22, 26, 27, 28, 40}, marked by vertical lines, which correspond to National Day (October 1–7, 2015, weeks 9–10 counting from August 2, 2015), New Year's Day (January 1–7, 2016, week 22), Spring Festival (end of January to late February, 2016, weeks 26–28), and International Workers' Day (May 1, 2016, week 40), respectively.

As standard procedures for functional data are not designed for data with such anomaly, this paper proposes a spike-preserving curve fitting procedure for functional data by introducing dummy regressors to quantify the spikes. Take for example the

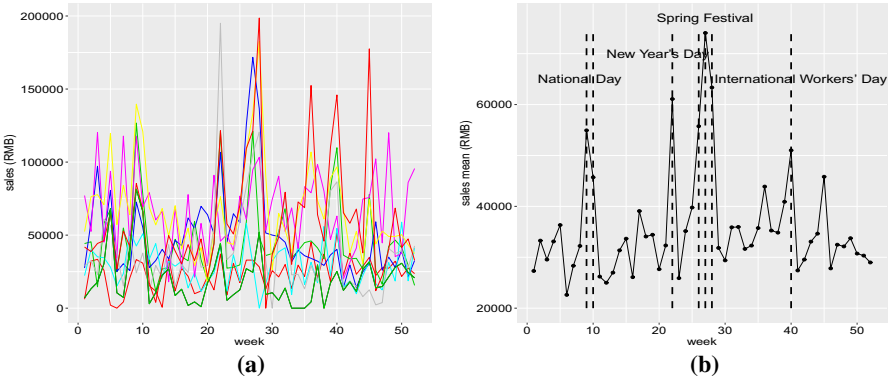

Fig. 1 **a** 10 sample weekly sales; **b** mean weekly sales

sales data of n salespersons over N time points (e.g., weeks), consisting of Y_{ij} the sales of the i-th salesperson at time point j, $1 \le i \le n$, $1 \le j \le N$, with spike β_ν at time point j_ν, $\nu = 1, \ldots, \mathcal{V}$. Following standard practice in functional data analysis, one converts time point j to a fractional point j/N of the scale free unit interval $[0, 1]$, and presumes a smooth trajectory $\xi_i(\cdot)$ (learning curve in management science terminology) of the i-th salesperson, a stochastic process of the variable time, for each $1 \le i \le n$. In our setup, these n trajectories are separated from the spikes, hence

$$\begin{aligned} Y_{ij_\nu} &= \xi_i(j_\nu/N) + \sigma(j_\nu/N)\,\varepsilon_{ij_\nu} + \beta_\nu, && \text{for the } j_\nu\text{-th holiday week,} \\ Y_{ij} &= \xi_i(j/N) + \sigma(j/N)\,\varepsilon_{ij}, && \text{for non-holiday week } j, \end{aligned}$$

where measurement errors ε_{ij} satisfy $\mathrm{E}\,\varepsilon_{ij} = 0, \mathrm{E}\,\varepsilon_{ij}^2 = 1$ and measurement variance $\sigma(\cdot)$.

Summarizing the above, a model describing functional data with holiday effects is as follows,

$$Y_{ij} = \xi_i(j/N) + \sigma(j/N)\uparrow_{ij} + \sum\nolimits_{\nu=1}^{\mathcal{V}} \beta_\nu D_\nu(j), \tag{1}$$

in which $D_\nu(j) = \delta_{jj_\nu}$, $1 \le \nu \le \mathcal{V}$, $1 \le j \le N$ are dummy variables, with Kronecker symbols $\delta_{jj'} = 1$ if $j = j'$, $\delta_{jj'} = 0$ otherwise. Although the above model (1) is motivated directly from the sportswear sales data, its use is also conceivable in other settings, for example, in the study of neural working curve of the brain, with spiked activity at certain time points.

This model should not to be confused with other models of discontinuity in stochastic processes, such as change-point in time series data, studied in Schröder and Fryzlewicz (2013), Fryzlewicz and Subba Rao (2014) and Cho and Fryzlewicz (2015). In regard to one Reviewer's comment, we note from Fig. 1a that the upward surge at the holiday weeks are uniform for each salesperson, rather than sporadic as functional data outliers discussed in Raña et al. (2015). Meanwhile, as upward surge in holiday sales does not die off quickly, it makes sense to aggregate the sales data in weeks rather than in days. In Fig. 1b, one can see that the National Day sales spike is over a 2 weeks period, while the Spring Festival (Chinese New Year) effect spreads over a 3 weeks period. This setup also addresses one Reviewer's concern that the holiday effect is modeled as spikes rather than smoothly changing phenomenon: had the data been collected in days rather than weeks, the holiday change in sale perhaps could have been smoothly changing instead of spiking. Also in response to Reviewer's comment, we note that our approach also differs from classification or clustering analysis as all salespersons' weekly sales spike up during holidays. Consequently, there is no natural separation of salespersons into two groups, one affected by holiday effects, another not.

The learning curve $\xi_i(\cdot)$ of the i-th salesperson is useful for tracking individual performance, a much more informative construction is the mean learning curve $m(\cdot) = \mathrm{E}\,\xi_i(\cdot)$ of all salespersons, a population mean of $\xi_i(\cdot)$. In our setup, the holiday effect parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{\mathcal{V}})^T$ is also of interest from management science point of view. Our goal is to provide efficient estimator for $m(\cdot)$ with asymptotic simultaneous confidence band (SCB) and estimator for $\boldsymbol{\beta}$ with asymptotic confidence intervals (CIs) for each holiday effect β_ν.

Studying the global shape of an unknown smooth function by using SCB has become a powerful tool of inference, see Zhou et al. (1998), Fan and Zhang (2000), Claeskens and Van Keilegom (2003), Wu and Zhao (2007), Huang et al. (2008), Zhao and Wu (2008), Song and Yang (2009), Wang and Yang (2009), Wang et al. (2014, 2016), Cai and Yang (2015), Gu and Yang (2015), Zheng et al. (2016), and Cai et al. (2019) for theory and applications of SCB. SCBs have been proposed in recent years for population mean curve of functional data in Ma et al. (2012), Cao et al. (2012, 2016) based on B-spline regression, and in Degras (2011) with local polynomial smoothing. The SCBs for $m(\cdot)$ and CIs for $\boldsymbol{\beta}$ are also extended to detect the difference between two populations. This is motivated by Fan and Lin (1998), Li and Yu (2008), Benko et al. (2009) and Cao et al. (2012) for testing equality between two populations of curves. In this paper, comparison has been made between various pairs of population learning curves and holiday effects such as male vs female, one brand of sportswear vs. another, one year vs. another, etc.

The paper is organized as follows. In Sect. 2, main results are stated on the oracally efficient estimation of $m(\cdot)$ and β_v and their SCBs and CIs. Section 3 describes the procedure to implement the proposed SCBs and CIs. Section 4 reports some simulation results and analyses of the sports footwear sales data. Technical proofs are collected in the online supplement.

2 Main results

2.1 Estimation procedure

We describe in this section how mean function $m(\cdot)$ and holiday effects $\boldsymbol{\beta}$ are estimated. If all the random trajectories $\xi_i(x), x \in [0,1], 1 \le i \le n$ were known by "oracle", one could compute the errors $e_{ij} = Y_{ij} - \xi_i(j/N)$ and obtain method of moment estimators for $m(x)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{\mathcal{V}})^T$ as

$$\bar{m}(x) = n^{-1} \sum_{i=1}^{n} \xi_i(x), x \in [0,1], \tag{2}$$

$$\tilde{\boldsymbol{\beta}} = \left(\tilde{\beta}_1, \ldots, \tilde{\beta}_{\mathcal{V}}\right)^T, \tilde{\beta}_v = e_{\cdot j_v} = n^{-1} \sum_{i=1}^{n} e_{ij_v}, 1 \le v \le \mathcal{V}. \tag{3}$$

The $\bar{m}(\cdot)$ and $\tilde{\boldsymbol{\beta}}$ are infeasible since they rely on unknown trajectories $\xi_i(\cdot), 1 \le i \le n$ in addition to the observed Y_{ij}'s. They serve, however, as useful benchmark and suggest that one mimics these would-be estimators by replacing trajectories $\xi_i(\cdot)$ with some reasonable estimates in order to obtain feasible estimators $\hat{m}(\cdot)$ and $\hat{\boldsymbol{\beta}}$ of $m(\cdot)$ and $\boldsymbol{\beta}$ as

$$\hat{m}(x) = n^{-1} \sum\nolimits_{i=1}^{n} \hat{\xi}_i(x), x \in [0,1], \tag{4}$$

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_{\mathcal{V}}\right)^T, \hat{\beta}_v = \hat{e}_{\cdot j_v} = n^{-1} \sum_{i=1}^{n} \hat{e}_{ij_v}, 1 \le v \le \mathcal{V}, \tag{5}$$

where residuals $\hat{e}_{ij_v} = Y_{ij_v} - \hat{\xi}_i(j_v/N)$ for some estimates $\hat{\xi}_i(\cdot)$ of $\xi_i(\cdot)$.

One convenient way of estimating of $\hat{\xi}_i(\cdot)$ is by spline, so we first introduce the space of spline functions defined on $[0, 1]$. Divide $[0, 1]$ into $(N_m + 1)$ subintervals $J_j = \left[\chi_j, \chi_{j+1}\right), j = 0, \ldots, N_m - 1, J_{N_m} = \left[\chi_{N_m}, 1\right]$ where the endpoints $\left\{\chi_j\right\}_{j=1}^{N_m}$ are equally spaced, and called interior knots. Let $\mathcal{H}_{N_m}^{(p-2)}$ =space of functions that are polynomials of degree $(p-1)$ on each J_j and have continuous $(p-2)$-th derivatives on $[0, 1]$, the p-th order spline space. Following de Boor (1978), denote by $b_{J,p}(x)$ the B-spline basis of order $p, J = 1-p, \ldots, N_m$, and denote by $\mathbf{b}(x) = \left(b_{1-p,p}(x), \ldots, b_{N_m,p}(x)\right)$ the row vector of all such basis, then $\mathcal{H}_{N_m}^{(p-2)} = \left\{\sum_{J=1-p}^{N_m} \lambda_J b_{J,p}(x), \lambda_J \in \mathbb{R}, x \in [0, 1]\right\}$. In what follows, we denote by $\mathcal{J}_\mathcal{V} = \{j_1, \ldots, j_\mathcal{V}\}$ the set of all time points where holiday effects occur, and $\bar{\mathcal{J}}_\mathcal{V} = \{1, \ldots, N\} \setminus \{j_1, \ldots, j_\mathcal{V}\}$ the complement set, i.e., those time points without holiday effects.

The spline estimator of $\xi_i(\cdot)$ is then defined as

$$\hat{\xi}_i(\cdot) = \sum_{J=1-p}^{N_m} \hat{\lambda}_{i,J} b_{J,p}(\cdot), \left(\hat{\lambda}_{i,J}\right)_{J=1-p}^{N_m} = \underset{\{\lambda_{i,J}\} \in \mathbb{R}^{N_m+p}}{\arg\min} \sum_{j \in \bar{\mathcal{J}}_\mathcal{V}} \left(Y_{ij} - \sum_{J=1-p}^{N_m} \lambda_{i,J} b_{J,p}(j/N)\right)^2$$

and standard algebra leads to

$$\hat{\xi}_i(x) = \mathbf{b}(x) \left(\mathbf{B}_{\_\mathcal{V}}^T \mathbf{B}_{\_\mathcal{V}}\right)^{-1} \mathbf{B}_{\_\mathcal{V}}^T \mathbf{Y}_{i\_\mathcal{V}}, 1 \le i \le n, \tag{6}$$

in which $\mathbf{Y}_{i\_\mathcal{V}} = \left(Y_{ij}, j \in \bar{\mathcal{J}}_\mathcal{V}\right)^T, 1 \le i \le n$ and

$$\mathbf{B}_{\_\mathcal{V}} = \{\mathbf{b}(j/N)\}_{j \in \bar{\mathcal{J}}_\mathcal{V}} = \begin{pmatrix} b_{1-p,p}(1/N) & \cdots & b_{N_m,p}(1/N) \\ \vdots & \vdots & \vdots \\ b_{1-p,p}(j_1/N - 1/N) & \cdots & b_{N_m,p}(j_1/N - 1/N) \\ b_{1-p,p}(j_1/N + 1/N) & \cdots & b_{N_m,p}(j_1/N + 1/N) \\ \vdots & \vdots & \vdots \\ b_{1-p,p}(j_\mathcal{V}/N - 1/N) & \cdots & b_{N_m,p}(j_\mathcal{V}/N - 1/N) \\ b_{1-p,p}(j_\mathcal{V}/N + 1/N) & \cdots & b_{N_m,p}(j_\mathcal{V}/N + 1/N) \\ \vdots & \vdots & \vdots \\ b_{1-p,p}(N/N) & \cdots & b_{N_m,p}(N/N) \end{pmatrix}$$

is a $(N - \mathcal{V}) \times (N_m + p)$ design matrix.

While we agree with one Reviewer that other estimators of $\xi_i(\cdot)$ such as Fourier series estimator may be as viable as spline for purpose of intermediate use in (4) and (5), theoretical development in what follows focuses on the B-spline method. This is due primarily to the computational advantage of B-spline as its empirical inner

product matrix is diagonally banded, in contrast to the more complicated empirical inner product matrix of Fourier basis. This simplicity of empirical inner product matrix also affords great theoretical convenience so technical results in the Supplement such as Lemmas S.3 to S.6 are proved much more easily.

2.2 Asymptotic properties

In the following, for any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_t)$, let $\|\boldsymbol{\zeta}\|_\infty = \max(|\zeta_1|, \ldots, |\zeta_t|)$, $\|\boldsymbol{\zeta}\|_r = \left(\zeta_1^r + \cdots + \zeta_t^r\right)^{1/r}$, $1 \leq r < +\infty$, and denote by $\psi^{(s)}(x)$ the s-th order derivative of a function $\psi(x)$. For $\theta \in (0, 1]$ and integer $p \geq 0$, let $C^{p,\theta}[0, 1]$ be the space of functions with θ-H ölder continuous p-th order derivatives on $[0, 1]$,

$$C^{p,\theta}[0,1] = \left\{ \phi(x) : \|\phi\|_{p,\theta} = \sup_{x \neq x', x, x' \in [0,1]} \frac{\left|\phi^{(p)}(x) - \phi^{(p)}\left(x'\right)\right|}{|x - x'|^\theta} < +\infty \right\},$$

and denote by $C^{(p)}[0, 1]$ the space of p -times continuously differentiable functions. For sequences of positive real numbers c_n and d_n, $c_n \ll d_n$ means $c_n/d_n \to 0$, as $n \to \infty$.

For the definition of $m(\cdot) = \mathrm{E}\,\xi_i(\cdot)$ to be valid, one assumes that all the n stochastic processes $\xi_i(\cdot)$ are independently and identically distributed, taking values in $L^2[0, 1]$. Denote by $C\left(x, x'\right)$ the covariance function of $\xi_i(x)$, i.e., $C\left(x, x'\right) = \mathrm{cov}\left\{\xi_i(x), \xi_i\left(x'\right)\right\}$. Assuming continuity of $C\left(x, x'\right)$ on $[0, 1]^2$, then Lemma 1.3 in Bosq (2012), also known as Mercer Lemma, implies that

$$C\left(x, x'\right) = \sum\nolimits_{k=1}^{\infty} \lambda_k \psi_k(x) \psi_k\left(x'\right),$$

where $\{\lambda_k\}_{k=1}^\infty$ and $\{\psi_k(x)\}_{k=1}^\infty$ are the eigenvalues and the corresponding eigenfunctions of $C\left(x, x'\right)$ as an integral operator on $L^2[0, 1]$. By Theorem 1.5 in Bosq (2012), also known as Karhunen-Loève expansion

$$\xi_i(x) = m(x) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(x),$$

where ξ_{ik} are uncorrelated random coefficients with mean 0, variance 1 and rescaled eigenfunctions $\phi_k(x) = \sqrt{\lambda_k}\psi_k(x)$.

We need the following technical assumptions:

(A1) *For integer $p > 1$, the function $m(\cdot) \in C^{p-1,1}[0, 1]$, while for some $\mu \in (0, 1]$, the standard deviation function $\sigma(\cdot) \in C^{0,\mu}[0, 1]$.*

(A2) *As $n \to \infty$, for some $\theta > 1/(2p)$, $N = \mathcal{O}\left(n^\theta\right)$ and $N \gg n^{1/(2p)} \log n$, while for the number of interior knots N_m, $n^{1/(2p)} \ll N_m \ll N \log^{-1} n$.*

(A3) *For some $C > 0$, $C(x, x) > C, x \in [0, 1]$, while $\phi_k \in C^{0,\mu}[0, 1], k = 1, 2, \ldots, \sum_{k=1}^\infty \|\phi_k\|_\infty < \infty$ and as $n \to \infty$, $N_m^{-\mu} \sum_{k=1}^{\kappa_n} \|\phi_k\|_{0,\mu} = o(1)$ for a sequence $\{\kappa_n\}_{n=1}^\infty$ of increasing integers, with $\lim_{n\to\infty} \kappa_n = \infty$.*

(A4) *There are i.i.d.* $N(0,1)$ *variables* $\left\{Z_{ik,\xi}\right\}_{i=1,k=1}^{n,\infty}$, $\left\{Z_{ik,\varepsilon}\right\}_{i=1,k=1}^{n,\infty}$ *and constants* $C_1, C_2 \in (0,\infty)$, $\gamma_1, \gamma_2 \in (1,\infty)$, $\rho \in (0,1/2)$ *such that*

$$\max_{1\le k\le\infty} P\left\{\max_{1\le t\le n}\left|\sum_{i=1}^{t}\xi_{ik}-\sum_{i=1}^{t}Z_{ik,\xi}\right|>C_1n^{\rho}\right\}<C_2n^{-\gamma_1},$$
$$P\left\{\max_{1\le j\le N}\max_{1\le t\le n}\left|\sum_{i=1}^{t}\varepsilon_{ij}-\sum_{i=1}^{t}Z_{ik,\varepsilon}\right|>C_1n^{\rho}\right\}<C_2n^{-\gamma_2}.$$

Assumption (A1) on smoothness of $m(\cdot)$ and $\sigma(\cdot)$ is a typical condition for spline smoothing adapted from Huang and Yang (2004), Wang and Yang (2009) and Ferraty and Vieu (2006), while Assumption (A2) with the choice of knots number N_m and the number of observations for each subject ensures $n^{-1/2}$ oracle efficiency. Assumption (A3) concerns the bounded smoothness of principal components, which is the same as Assumption (A4) in Cao et al. (2012). While Lemma 1.3 of Bosq (2000) ensures the continuity of eigenfunctions $\phi_k(x)$ when the covariance function $C(x,x')$ is continuous, the Hölder continuity and other conditions in Assumption (A3) do not follow automatically. Assumption (A4) provides the Gaussian approximation of the error process and according to Lemma S.2 in the Supplement, it is ensured by the following elementary moment condition:

(A4') There exist $\eta_1 > 2, \eta_2 > 2+2\theta$ such that $\mathrm{E}\left|\xi_{ik}\right|^{\eta_1+2} + \mathrm{E}\left|\varepsilon_{ij}\right|^{\eta_2+2} < +\infty$, for $1 \le i < \infty, 1 \le k \le \infty, 1 \le j < \infty$. The number of nonzero eigenvalues is finite or it is infinite while the variables $\{\xi_{ik}\}_{i=1,k=1}^{n,\infty}$ are iid

Theorem 1 *Under Assumptions (A1)–(A4), the two-step estimators* $\hat{m}(x)$ *and* $\hat{\boldsymbol{\beta}}$ *are asymptotically equivalent to* $\bar{m}(x)$ *and* $\tilde{\boldsymbol{\beta}}$*, respectively, with the order* $n^{-1/2}$,

$$\sup_{x\in[0,1]}\left|\hat{m}(x)-\bar{m}(x)\right| = o_p\left(n^{-1/2}\right), \tag{7}$$
$$\left\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\right\|_{\infty} = o_p\left(n^{-1/2}\right). \tag{8}$$

Theorem 1 shows that the two-step estimators $\hat{m}(x)$ of $m(x)$ and $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ approximate the infeasible estimator $\bar{m}(x)$ and $\tilde{\boldsymbol{\beta}}$, respectively, at a rate smaller than $n^{-1/2}$, which is negligible compared to the convergence rate of parametric estimators. $\hat{m}(x)$ and $\hat{\boldsymbol{\beta}}$, therefore, are oracally efficient up with the order $n^{-1/2}$.

Let $\mathcal{G}(x) = C(x,x)^{-1/2}\sum_{k=1}^{\infty} Z_k\phi_k(x)$, in which the Z_k's, $k = 1, 2, \ldots$ are iid standard normal variables. Then, $\mathcal{G}(x)$ is a Gaussian process such that $\mathrm{E}\,\mathcal{G}(x) = 0$, $\mathrm{E}\,\mathcal{G}^2(x) = 1$ with covariance function

$$\mathrm{E}\,\mathcal{G}(x)\,\mathcal{G}\left(x'\right) = C\left(x,x'\right)\left\{C(x,x)\,C\left(x',x'\right)\right\}^{-1/2}, x, x' \in [0,1].$$

In addition, the expected value of the absolute maximum $\sup_{x\in[0,1]}|\mathcal{G}(x)|$ is finite according to Assumption (A3), so one can denote by $Q_{1-\alpha}$ the $100(1-\alpha)$-th percentile of its distribution, i.e., $P\left\{\sup_{x\in[0,1]}|\mathcal{G}(x)| \le Q_{1-\alpha}\right\} = 1-\alpha, \forall\alpha \in (0,1)$. The following result on $\bar{m}(x)$ is cited from Theorem 2.1 of Cao et al. (2012).

Proposition 1 *Under Assumptions (A1)–(A4), for any* $\alpha \in (0,1)$, *as* $n \to \infty$, *"the infeasible estimator"* $\bar{m}(x)$ *converges at the rate of* $n^{-1/2}$ *uniformly with*

$$P\left\{\sup_{x\in[0,1]} n^{1/2}\left|\bar{m}(x)-m(x)\right| C^{-1/2}(x,x) \le Q_{1-\alpha}\right\} \to 1-\alpha.$$

The following corollary is a direct result of Theorem 1 and Proposition 1 above, which summarizes the explicit expression of a SCB for $m(x)$.

Corollary 1 *Under Assumptions (A1)–(A4), for any* $\alpha \in (0,1)$, *as* $n \to \infty$, *an asymptotic* $100(1-\alpha)\%$ *SCB for* $m(x)$ *is*

$$\hat{m}(x) \pm n^{-1/2} C^{1/2}(x,x)\, Q_{1-\alpha}, x \in [0,1].$$

We next describe the asymptotic distribution of $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ and hence obtain the CIs and confidence region for $\boldsymbol{\beta}$.

Theorem 2 *Under Assumptions (A1)–(A4), as* $n \to \infty$, *one has that*

$$n^{1/2}\Sigma^{-1/2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \to_D \mathcal{N}(\mathbf{0}, \mathbf{I}_{\mathcal{V}}),$$

in which $\Sigma = \operatorname{diag}\left(\sigma^2(j_1/N), \ldots, \sigma^2(j_{\mathcal{V}}/N)\right)$ *and* $\mathbf{I}_{\mathcal{V}}$ *is the* $\mathcal{V}$ *dimensional identity matrix.*

Theorem 2 implies that $100(1-\alpha)\%$ asymptotically correct CIs for β_ν, $1 \le \nu \le \mathcal{V}$ are

$$\hat{\beta}_\nu \pm n^{-1/2}\sigma(j_\nu/N)\, z_{1-\alpha/2},$$

and $100(1-\alpha)\%$ asymptotically correct confidence region for $\boldsymbol{\beta}$ is

$$\|\Sigma^{-1/2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\|_2^2 \le n^{-1}\chi^2_{1-\alpha}(\mathcal{V}),$$

where $z_{1-\alpha/2}$ and $\chi^2_{1-\alpha}(\mathcal{V})$ are the $100(1-\alpha/2)$th percentile of the standard normal and the $100(1-\alpha)$th percentile of the $\mathcal{V}$ degrees chi-squared distribution, respectively. Thus, one can test the difference of holiday effects in the model (1), e.g., $H_0: \beta_\nu = 0, 1 \le \nu \le \mathcal{V}$ or $H_0: \beta_1 = \cdots = \beta_{\mathcal{V}} = 0$ by applying the CIs and the confidence region given above.

Theorem 1 is established by decomposing the discrepancy $\hat{m}(x) - m(x)$ into simpler terms. For any function $\phi \in C[0,1]$, denote the vector $\boldsymbol{\phi}_{\mathcal{V}} = \left(\phi(j/N), j \in \bar{\mathcal{I}}_{\mathcal{V}}\right)^T$ and the function

$$\tilde{\phi}(x) = \mathbf{b}(x)\left(\mathbf{B}_{\mathcal{V}}^T\mathbf{B}_{\mathcal{V}}\right)^{-1}\mathbf{B}_{\mathcal{V}}^T\boldsymbol{\phi}_{\mathcal{V}}. \tag{9}$$

Projecting via (4) and (6) the relationship in model (1) onto the linear space of $\mathbb{R}^{N_m+p}$ spanned by $\left\{b_{J,p}\left(j/N\right)\right\}_{j\in\bar{\mathcal{J}}_{\mathcal{V}},1-p\leq J\leq N_m}$, we obtain the following crucial decomposition in the space $\mathcal{H}_{N_m}^{(p-2)}$ of spline functions:

$$\hat{m}\left(x\right)=n^{-1}\sum_{i=1}^{n}\left\{\tilde{\xi}_i\left(x\right)+\tilde{\varepsilon}_i\left(x\right)\right\}, \tag{10}$$

$$\tilde{\xi}_i\left(x\right)=\tilde{m}\left(x\right)+\sum_{k=1}^{\infty}\xi_{ik}\tilde{\phi}_k\left(x\right),\ \tilde{\varepsilon}_i\left(x\right)=\mathbf{b}\left(x\right)\left(\mathbf{B}_{\_\mathcal{V}}^{T}\mathbf{B}_{\_\mathcal{V}}\right)^{-1}\mathbf{B}_{\_\mathcal{V}}^{T}\mathbf{E}_{i\_\mathcal{V}}, \tag{11}$$

where $\tilde{m}\left(x\right)$ and $\tilde{\phi}_k\left(x\right)$ are defined according to (9), and $\mathbf{E}_{i\_\mathcal{V}}=\left(\sigma\left(j/N\right)\varepsilon_{ij}\right.$, $\left.j\in\bar{\mathcal{J}}_{\mathcal{V}}\right)^{T}$.

The oracle efficiency on $\hat{m}\left(x\right)$ and $\hat{\boldsymbol{\beta}}$ in Theorem 1 immediately follows from the next two propositions concerning $n^{-1}\sum_{i=1}^{n}\tilde{\xi}_i\left(x\right)$ and $n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_i\left(x\right)$.

Proposition 2 *Under Assumptions (A1)–(A4), as* $n\to\infty$,

$$\sup_{x\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}\left\{\tilde{\xi}_i\left(x\right)-\bar{m}\left(x\right)\right\}C\left(x,x\right)^{-1/2}\right|=o_p\left(n^{-1/2}\right).$$

Proposition 3 *Under Assumptions (A1)–(A4), as* $n\to\infty$,

$$\sup_{x\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_i\left(x\right)C\left(x,x\right)^{-1/2}\right|=o_p\left(n^{-1/2}\right).$$

The online Supplement contains proofs for the above two propositions.

2.3 Extension to two-sample problems

When two sets of data $\left\{Y_{ij,1}\right\}_{i,j=1}^{n_1,N_1}$, $\left\{Y_{ij,2}\right\}_{i,j=1}^{n_2,N_2}$ are observed with holiday effects $\boldsymbol{\beta}_1=\{\beta_{11},\ldots,\beta_{1\mathcal{V}}\}$ and $\boldsymbol{\beta}_2=\{\beta_{12},\ldots,\beta_{2\mathcal{V}}\}$ satisfying model (1) with the covariance function $C_1\left(x,x'\right)$ and $C_2\left(x,x'\right)$, mean function $m_1\left(x\right)$ and $m_2\left(x\right)$, and the variance function $\sigma_1^2\left(x\right)$ and $\sigma_2^2\left(x\right)$, respectively, analogous SCB can be constructed for the difference of their mean curves. Denote the ratio of two sample sizes as $\hat{r}=n_1/n_2$, one assumes that $\lim_{n_1,n_2\to\infty}\hat{r}=r>0$. Let $\hat{m}_1\left(x\right)$, $\hat{m}_2\left(x\right)$ and $\hat{\boldsymbol{\beta}}_1=\left\{\hat{\beta}_{11},\ldots,\hat{\beta}_{1\mathcal{V}}\right\}$, $\hat{\boldsymbol{\beta}}_2=\left\{\hat{\beta}_{21},\ldots,\hat{\beta}_{2\mathcal{V}}\right\}$ be the two-step estimates described in Sect. 2.1 of mean functions $m_1\left(x\right),m_2\left(x\right)$ and parameters $\boldsymbol{\beta}_1=\{\beta_{11},\ldots,\beta_{1\mathcal{V}}\},\boldsymbol{\beta}_2=\{\beta_{12},\ldots,\beta_{2\mathcal{V}}\}$ for each group, respectively. Set $C_{12}\left(x,x'\right)=C_1\left(x,x'\right)+rC_2\left(x,x'\right)$ and $\sigma_{12}^2\left(x\right)=\sigma_1^2\left(x\right)+r\sigma_2^2\left(x\right)$, and let $\mathcal{G}_{12}\left(x\right),x\in[0,1]$ denote a standardized Gaussian process such that $\mathrm{E}\,\mathcal{G}_{12}\left(x\right)=0$, $\mathrm{E}\,\mathcal{G}_{12}^2\left(x\right)=1$ and covariance function

$$\mathrm{E}\,\mathcal{G}_{12}\left(x\right)\mathcal{G}_{12}\left(x'\right)=C_{12}\left(x,x'\right)\left\{C_{12}\left(x,x\right)C_{12}\left(x',x'\right)\right\}^{-1/2}.$$

Denote by $Q_{12,1-\alpha}$, the $(1-\alpha)$-th quantile of the absolute maxima deviation of $\mathcal{G}_{12}(x), x \in [0,1]$ as the above. Proof of the following analog of two-sample t-test is similar to that of Theorem 1 and Proposition 1.

Theorem 3 *Under Assumptions (A1)–(A4), modified for each group accordingly, as* $n_1 \to \infty, \hat{r} \to r > 0$, *a* $100(1-\alpha)\%$ *asymptotically correct SCB for* $m_1(x) - m_2(x), x \in [0,1]$ *is*

$$\{\hat{m}_1(x) - \hat{m}_2(x)\} \pm n_1^{-1/2} C_{12}(x,x)^{1/2} Q_{12,1-\alpha}, \tag{12}$$

and $100(1-\alpha)\%$ *asymptotically correct CIs for* $\beta_{1\nu} - \beta_{2\nu}, 1 \le \nu \le \mathcal{V}$ *are*

$$(\hat{\beta}_{1\nu} - \hat{\beta}_{2\nu}) \pm n_1^{-1/2} \sigma_{12}(j_\nu/N) z_{1-\alpha/2}. \tag{13}$$

The SCB in (12) can be used to test any hypotheses on $m_1(x) - m_2(x)$, while the CIs in (13) any hypotheses on the difference of holiday effects between the groups, e.g., the null hypothesis $H_0 : \beta_{1\nu} - \beta_{2\nu} = 0, 1 \le \nu \le \mathcal{V}$.

3 Implementation

This section describes steps to implement the oracally efficient estimators, the SCB and the CIs in Corollary 1, Theorems 2 and 3, respectively. The issues to be addressed are: choosing the number of knots N_m for spline smoothing, estimating the covariance function $C(x, x')$, the error variance $\sigma^2(x)$ and the percentile $Q_{1-\alpha}$.

3.1 Knot selection and estimating covariance function $C(x, x')$

Although Ma (2014) provided optimal order for the number N_m of interior knots for individual curve estimates $\hat{\xi}_i(\cdot)$, the more appropriate form of N_m for the purpose of this paper is $\lfloor an^{1/(2p)} \log n \rfloor$, satisfying Assumption (A2), where $[c]$ denotes the integer part of c. Simulation examples lead us to settle for the choice of constant $a = 0.2, 0.3, 0.5, 1.2$, with default $a = 0.5$. Similar empirical formulas were also used in Cao et al. (2012, 2016) to select the number of knots when constructing SCBs for mean functions and covariance function, respectively.

The covariance function $C(x, x')$ is estimated by tensor product B-spline:

$$\hat{C}_p(x, x') = \text{argmin}_{g(\cdot,\cdot) \in \mathcal{H}_{N_C}^{(p-2),2}} \sum_{j \ne j', j, j' \in \bar{\mathcal{J}}_\mathcal{V}} \left\{ V_{\cdot jj'} - g(j/N, j'/N) \right\}^2,$$

with $V_{\cdot jj'} = n^{-1} \sum_{i=1}^n \left\{ Y_{ij} - \hat{m}(j/N) \right\} \left\{ Y_{ij'} - \hat{m}(j'/N) \right\}$, the tensor product spline space $\mathcal{H}_{N_C}^{(p-2),2} = \left\{ \sum_{J,J'=1-p}^{N_C} \omega_{JJ'} b_{J,p}(t) b_{J',p}(s), \omega_{JJ'} \in \mathbb{R}, t, s \in [0,1] \right\}$ and the optimal interior knots number $N_C = \lfloor n^{1/(2p)} \log\log n \rfloor$.

3.2 Estimating the variance of the measurement error

Set $\sigma_Y^2(x) = C(x,x) + \sigma^2(x)$, then $\mathrm{Var}(Y_{ij}) = \sigma_Y^2(j/N)$, hence $\sigma_Y^2(x)$ is estimated by smoothing the moment estimator $V_{\cdot jj} = n^{-1}\sum_{i=1}^{n}\left\{Y_{ij} - \hat{m}(j/N)\right\}^2$ of $\sigma_Y^2(j/N)$

$$\hat{\sigma}_Y^2(x) = \mathrm{argmin}_{g\in\mathcal{H}_{N_\sigma}^{(p-2)}}\sum_{j\in\bar{\mathcal{J}}_\mathcal{V}}\left\{V_{\cdot jj} - g(j/N)\right\}^2.$$

where the interior knots number N_σ satisfies the same condition as for N_m in (A3), and by default, follows the same empirical formula of N_m in Sect. 3.1. One then estimates $\sigma^2(x)$ by $\hat{\sigma}^2(x) = \hat{\sigma}_Y^2(x) - \hat{C}_p(x,x)$. According to the argument for Theorem 1 of Cao et al. (2012), if $\sigma_Y^2(x)$ is a smooth function that satisfies the Hölder condition as for $m(x)$, one then obtains that $\sup_x\left|\hat{\sigma}_Y^2(x) - \sigma_Y^2(x)\right| = \mathcal{O}_p\left(n^{-1/2}\right)$ and hence $\sup_x\left|\hat{\sigma}^2(x) - \sigma^2(x)\right| = \mathcal{O}_p\left(n^{-1/2}\right)$.

The asymptotic CIs for β_ν, $1 \le \nu \le \mathcal{V}$ are

$$\hat{\beta}_\nu \pm n^{-1/2}\hat{\sigma}(j_\nu/N)\, z_{1-\alpha/2}.$$

For the two-sample problems, one estimates $\sigma_1^2(x)$ and $\sigma_2^2(x)$, respectively, analogous to $\hat{\sigma}^2(x)$, and then plug them in $\sigma_{12}^2(x)$ to obtain a estimator $\hat{\sigma}_{12}^2(x)$. Hence, one can compute the asymptotic CIs for $\beta_{1\nu} - \beta_{2\nu}$ as

$$\left(\hat{\beta}_{1\nu} - \hat{\beta}_{2\nu}\right) \pm n_1^{-1/2}\hat{\sigma}_{12}(j_\nu/N)\, z_{1-\alpha/2},$$

which can be used to test and compare the difference of holiday effects.

3.3 Estimating the percentile $Q_{1-\alpha}$

To evaluate $Q_{1-\alpha}$, one needs to simulate the Gaussian process $\mathcal{G}(x)$ in Sect. 2.2. Let $\hat{\mathcal{G}}(x) = C(x,x)^{-1/2}\sum_{k=1}^{\infty} Z_k\phi_k(x)$ where Z_k are i.i.d. standard normal variables. Hence, $\hat{\mathcal{G}}(x)$ is a Gaussian process with mean 0, variance 1 and covariance function $C\left(x,x'\right)\left\{C(x,x) \times C\left(x',x'\right)\right\}^{-1/2}$ which is the same random process with $\mathcal{G}(x)$.

To simulate $\hat{\mathcal{G}}(x)$, one first needs the eigenfunction decomposition of $\hat{C}_p\left(x,x'\right)$ described in Sect. 3.1 to obtain the estimated eigenvalues $\hat{\lambda}_k$, orthonormal eigenfunctions $\hat{\psi}_k(x)$, and rescaled eigenfunctions $\hat{\phi}_k(x) = \hat{\lambda}_k^{1/2}\hat{\psi}_k(x)$, respectively. We next truncate the infinite expansion of $\hat{C}_p\left(x,x'\right) = \sum_{k=1}^{\infty}\hat{\phi}_k(x) \times \hat{\phi}_k\left(x'\right)$ at a chosen order κ. The number of principal component κ is chosen by the "fraction of variation explained"method in Cao et al. (2016): one selects the number of eigenvalues that can explain 95% of the variation in the data, that is, $\kappa = \mathrm{argmin}_{1\le l\le L}\left\{\sum_{k=1}^{l}\hat{\lambda}_k/\sum_{k=1}^{L}\hat{\lambda}_k > 0.95\right\}$, where $\left\{\hat{\lambda}_k\right\}_{k=1}^{L}$ are the first L estimated positive eigenvalues.

Finally, one simulates $\hat{\mathcal{G}}_b(x) = \hat{C}_p(x,x)^{-1/2}\sum_{k=1}^{\kappa} Z_{k,b}\hat{\phi}_k(x)$ where $Z_{k,b}$ are i.i.d. standard normal variables and $b = 1,\ldots,b_M$, in which b_M is a preset large

integer, with default value $b_M = 5000$. The maximal absolute deviation for each copy of $\hat{\mathcal{G}}_b(x)$, is taken and $Q_{1-\alpha}$ is estimated as the empirical percentile $\hat{Q}_{1-\alpha}$ of these maximum values.

Hence, a SCB for mean function is

$$\hat{m}(x) \pm n^{-1/2}\hat{C}_p(x,x)^{1/2}\,\hat{Q}_{1-\alpha}, x \in [0,1]. \tag{14}$$

One estimates $Q_{12,1-\alpha}$ analogous to $\hat{Q}_{1-\alpha}$ and computes

$$\{\hat{m}_1(x) - \hat{m}_2(x)\} \pm n^{-1/2}\hat{C}_{12}(x,x)^{1/2}\,\hat{Q}_{12,1-\alpha}, x \in [0,1],$$

as a SCB for $m_1(x) - m_2(x)$ which can be used to test and compare the mean curve differences of two groups.

4 Simulation studies and analysis of sports footwear sales data

This section illustrates the use of the proposed method via simulation and by analyzing sports footwear sales data of a company in Shanghai, China.

4.1 Simulation studies

We generate data from the model

$$Y_{ij} = m(j/N) + \sum_{k=1}^{2}\xi_{ik}\phi_k(j/N) + \sigma\varepsilon_{ij} + \sum_{\nu=1}^{\mathcal{V}}\beta_\nu D_\nu(j), \tag{15}$$

where $\xi_{ik}, \varepsilon_{ij} \sim \mathcal{N}(0,1)$ are independent variables with $\phi_1(x) = -2\cos\{\pi(x-1/2)\}$, $\phi_2(x) = \sin\{\pi(x-1/2)\}$. Following four cases are studied: $m_1(x) = 3 - \exp(-x/2)$, a commonly used exponential sales curve discussed in Anzanello and Fogliatto (2011), and $m_2(x) = \sin\{2\pi(1/2 - x)\}$, a sine signal function; the noise levels are set to be $\sigma_1 = 0.5$ and $\sigma_2 = 1$.

Case 1: m_1, σ_1 Case 2: m_1, σ_2
Case 3: m_2, σ_1 Case 4: m_2, σ_2

To have a similar design of the real data example in Sect. 4.2, the number of subjects n is set to be 50, 100, 200, 300, with corresponding number of observations per curve $N = \lceil 3.3n^{0.3}\log n \rceil = 41, 60, 85, 104$, respectively. The holiday effect points $j_\nu, \nu = 1, 2, \ldots, 6$ to be $\{9, 22, 26, 27, 28, 40\}$ with $\beta_\nu = \nu \times m(j_\nu/N)$. Confidence levels $1 - \alpha = 0.95, 0.99$ are used, while each combination of Case and sample size is replicated 500 times.

Table 1 shows the empirical frequencies that at the points $\{1/N, \ldots, N/N\}$ the true curve $m(x)$ is covered by the asymptotic SCB and the bootstrap SCB, together with computing time. The bootstrap SCB is constructed by randomly selecting n subjects

Table 1 Empirical coverage frequencies of the asymptotic SCB and the bootstrap SCB based on 500 replications generated from model (15) and computing time in minutes (inside parentheses)

n	$1-\alpha$	Case 1		Case 2	
		Asymptotic	Bootstrap	Asymptotic	Bootstrap
50	0.950	0.888 (1.46)	0.938(6.26)	0.802 (1.76)	0.932 (6.10)
	0.990	0.978 (1.47)	0.982 (6.94)	0.924 (1.65)	0.978 (6.06)
100	0.950	0.914 (1.84)	0.934 (6.71)	0.852 (2.05)	0.932 (6.51)
	0.990	0.978 (1.81)	0.986 (6.59)	0.950 (1.99)	0.980 (6.51)
200	0.950	0.922 (2.76)	0.940 (8.58)	0.868 (2.80)	0.938 (7.81)
	0.990	0.970 (2.75)	0.990 (7.90)	0.946 (2.80)	0.990 (7.83)
300	0.950	0.936 (4.25)	0.948 (9.49)	0.892 (3.62)	0.954 (9.12)
	0.990	0.984 (3.91)	0.994 (9.46)	0.968 (3.55)	0.998 (9.36)
n	$1-\alpha$	Case 3		Case 4	
		Asymptotic	Bootstrap	Asymptotic	bootstrap
50	0.950	0.888 (2.34)	0.940 (5.60)	0.804 (1.44)	0.932 (6.20)
	0.990	0.976 (2.70)	0.982 (5.61)	0.922 (1.44)	0.978 (5.88)
100	0.950	0.912 (3.71)	0.936 (6.27)	0.852 (2.08)	0.926 (6.44)
	0.990	0.978 (3.68)	0.982 (6.20)	0.946 (2.07)	0.980(6.97)
200	0.950	0.922 (3.52)	0.940 (7.35)	0.886 (2.79)	0.938 (7.62)
	0.990	0.970 (2.52)	0.990 (7.20)	0.944 (2.94)	0.990 (7.67)
300	0.950	0.936 (3.54)	0.948 (8.56)	0.886 (3.63)	0.952 (8.60)
	0.990	0.984 (3.49)	0.992 (8.54)	0.970 (3.62)	0.998 (8.58)

with replacement from the original n subjects and computing the cubic spline ($p = 4$) estimate of $m(x)$, a procedure which is repeated 1000 times.

Table 1 shows that (i) for all Cases, the coverage frequencies of the SCBs improve and approach the nominal levels 0.95 and 0.99 as sample size n increases, a positive confirmation of (7) in Theorem 1 and of Corollary 1; (ii) when sample size is small, the bootstrap SCB performs better than the asymptotic one; however, when sample size becomes larger, performance of the two is similar. It is consistent with the findings of Claeskens and Van Keilegom (2003) which studied bootstrap SCB for the nonparametric regression function; see Table 1 in Claeskens and Van Keilegom (2003) for the comparisons; (iii) the asymptotic SCB is much faster to compute than the bootstrap SCB. The results for linear spline estimator ($p = 2$) are similar but omitted to save space.

Figure 2 depicts the boxplots of $\sqrt{n}\left(\hat{\beta}_1 - \tilde{\beta}_1\right)$ for all four Cases, positively corroborating with (8) in Theorem 1: the boxes become narrower and closer to 0 as the sample size increases from 50 to 300. Performing standard z- and χ^2-tests derived from Theorem 2, null hypotheses of no holiday effects $H_0 : \beta_\nu = 0$ with $j_\nu \in \{9, 22, 26, 27, 28, 40\}$ and $H_0 : \beta_1 = \cdots = \beta_6 = 0$ are all strongly rejected with p-values less than 1.28×10^{-3}, for all cases with sample size $n = 100$.

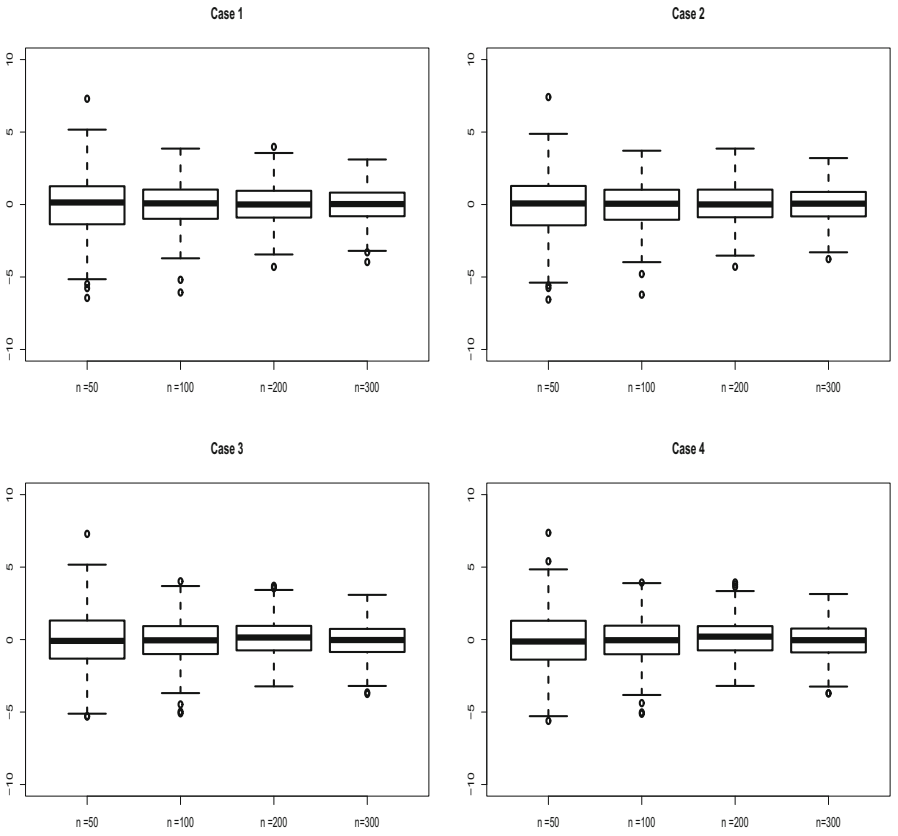

Fig. 2 Boxplots of $\sqrt{n}\left(\hat{\beta}_1 - \tilde{\beta}_1\right)$ based on cubic spline for the data generated from model (15) over 500 replications in four cases

For visual impression of the actual function estimates, SCB (dashed) and cubic spline estimator $\hat{m}(x)$ (dashed-dotted) for the true function $m(x)$ (solid) are plotted in Figs. 3 and 4 for Cases 1 and 3, respectively. As expected, as n increases, the SCB and spline estimator zero in to the true function.

Following one reviewer's suggestion, the following model is examined:

$$Y_{ij} = m(j/N) + \sum_{k=1}^{10} \xi_{ik}\phi_k(j/N) + \sigma\varepsilon_{ij} + \sum_{\nu=1}^{\mathcal{V}} \beta_\nu D_\nu(j) \tag{16}$$

with $\phi_{2s-1}(x) = \cos\left\{6^{-1}(2s-1)\pi x\right\}$, $\phi_{2s}(x) = \sin\left\{3^{-1}s\pi x\right\}$, $s = 1, 2, \ldots, 5$, and $m(x)$, $\xi_{ik}, \varepsilon_{ij}, \sigma, \beta_\nu$ being the same as above in (15) so there are 10 nonzero eigenvalues. Table 2 shows the empirical coverage frequencies for both the asymptotic SCB and the bootstrap SCB. One can see that the two SCBs perform similarly even when the sample size is small.

Figure 5 shows the boxplots of $\sqrt{n}\left(\hat{\beta}_1 - \tilde{\beta}_1\right)$ for the four cases based on the data generated from model (16), while Figs. 6 and 7 depict the true function $m(x)$ (solid),

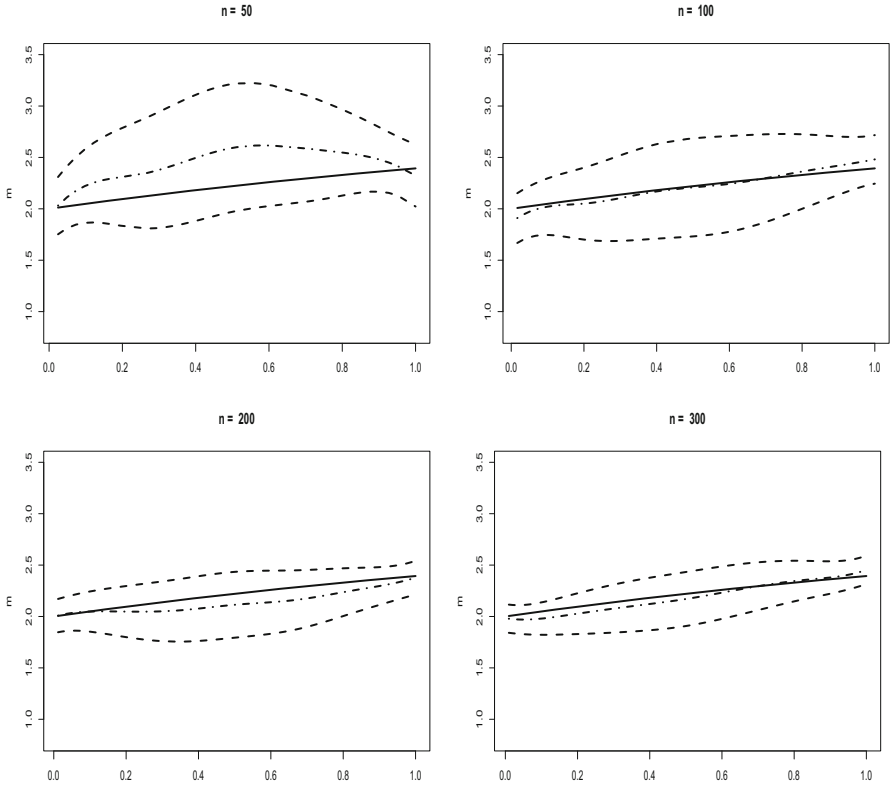

Fig. 3 Plots of 95% SCB (dashed) for $m(x)$ (solid) and cubic spline estimator $\hat{m}(x)$ (dashed-dotted) for Case 1, which are computed by (14) based on data generated from model (15)

the cubic spline estimator $\hat{m}(x)$ (dashed-dotted) and the SCB (dashed) for Cases 1 and 3, respectively. One can see that these results are similar to the scenario of $k = 2$ above and positively confirm to our asymptotic theories. These also imply that the number of positive eigenvalues does not significantly impact the performance of the proposed estimates and SCBs.

4.2 Sports footwear sales data

In this subsection, various SCBs and CIs are constructed and hypotheses tested about sports footwear sales data provided by a sportswear retailing company located in Shanghai, China, collected from August 3, 2014 to July 31, 2016, consisting of weekly sales of salespersons at retail stores.

Inference is made on population mean curve $m(x)$ of weekly sales data from August 2, 2015 to July 31, 2016 and August 3, 2014 to August 1, 2015, respectively, by SCB to determine whether these two learning curves are flat or with trend. Also examined is the difference of these two learning curves to detect any year-to-year change, the difference of female and male learning curves from August 3, 2014, to August 1, 2015,

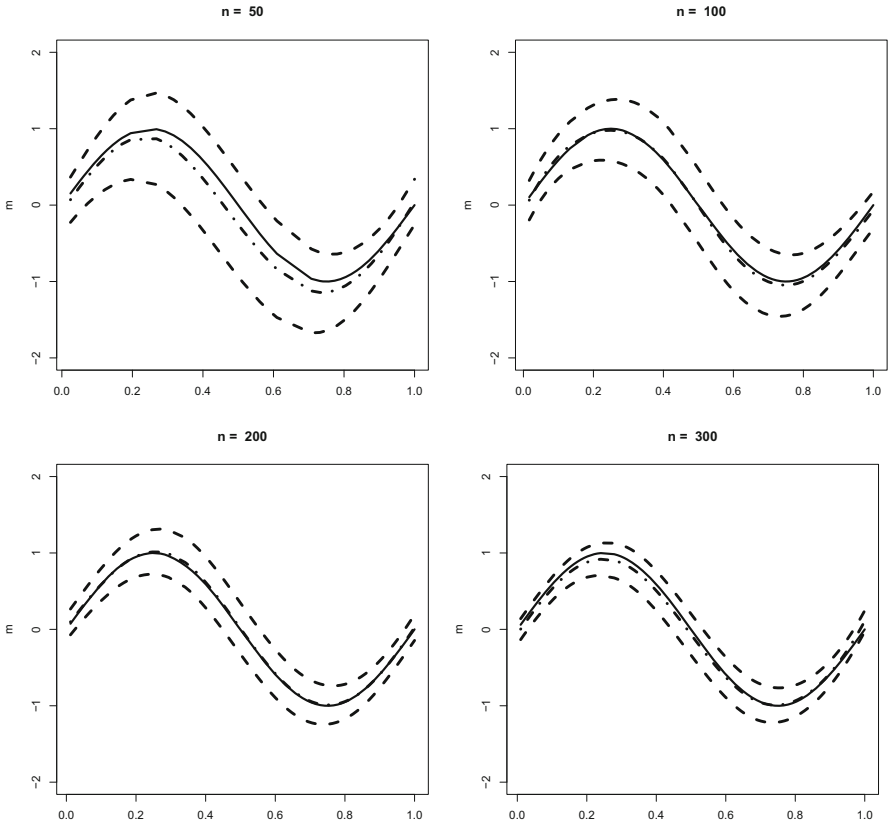

Fig. 4 Plots of 95% SCB (dashed) for $m(x)$ (solid) and cubic spline estimator $\hat{m}(x)$ (dashed-dotted) for Case 3, which are computed by (14) based on data generated from model (15)

and the difference of learning curves of two brands of sports footwear for the same year.

For the sales data from August 2, 2015, to July 31, 2016, there are $n = 74$ salespersons, each having $N = 52$ weekly sales records. Figure 1a and b show 10 randomly sampled weekly sales records and the average of all weekly sales. Holiday spikes are clearly visible at weeks $\{9, 10, 22, 26, 27, 28, 40\}$ with the holidays being National Day (October 1–7, 2015, weeks 9–10 counting from August 2, 2015), New Year's Day (January 1–7, 2016, week 22), Spring Festival (end of January to late February, 2016, week 26–28) and International Workers' Day (May 1, 2016, week 40). Four randomly selected cubic spline trajectory estimates $\hat{\xi}_i(x)$ (solid, as described in Sect. 2.1), with corresponding sales records (circles), are displayed in Fig. 8. These individual learning curves do not exhibit much similarity, hence useful information is only extractable from the population learning curve.

Figure 9a shows the cubic spline estimate $\hat{m}(x)$ (thick solid) of population curve, the 95% SCB (dashed) and a horizontal line (solid) to test hypothesis $H_0 : m(x) = 32,326.7$. The constant 32, 326.7 is simply taken as the sales' average over all non-

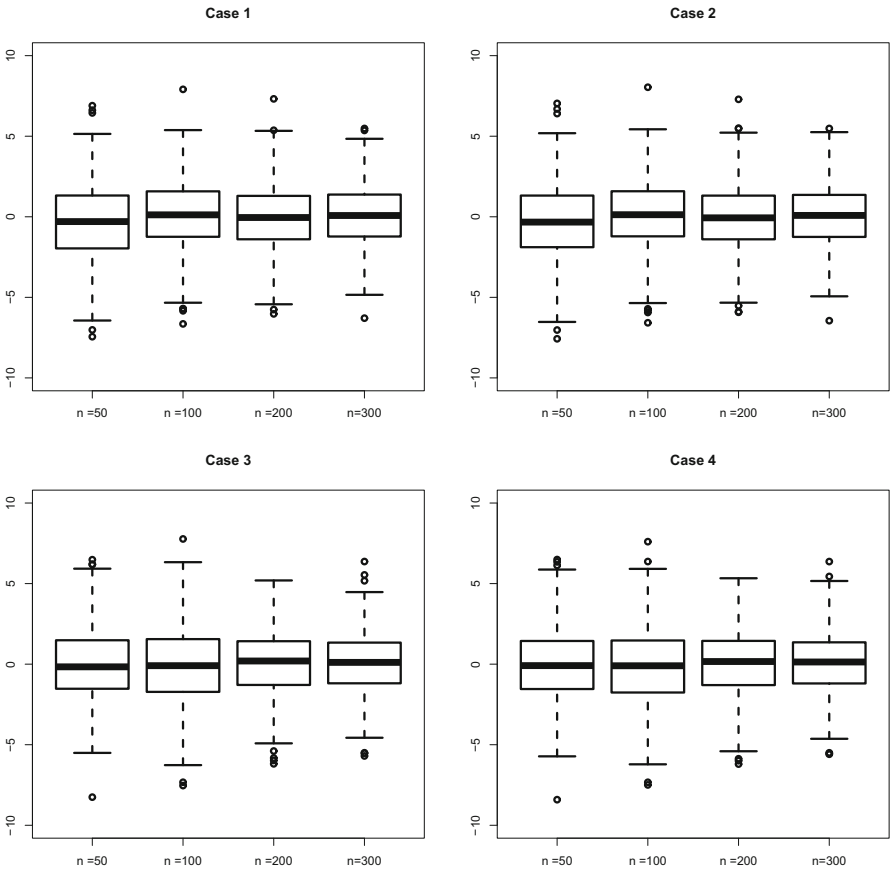

Fig. 5 Boxplots of $\sqrt{n}\left(\hat{\beta}_1 - \tilde{\beta}_1\right)$ based on cubic spline for the data generated from model (16) over 500 replications in four cases

holiday weeks, i.e., $n^{-1}\,(N-\mathcal{V})^{-1}\sum_{i=1}^{n}\sum_{j\in\bar{\mathcal{J}}_{\mathcal{V}}} Y_{ij}$. Since the minimum confidence level of SCB containing the entire null constant 32, 326.7 curve is 72.54%, one retains the null hypothesis with p-value 0.2746, see Fig. 9b. One therefore concludes that mean weekly sales during August 2, 2015 to July 31, 2016 of salespersons can be taken as 32, 326.7. For the sales data from August 3, 2014 to August 1, 2015 with holiday weeks {9, 10, 22, 27, 28, 29, 39}, Fig. 10 shows similar patterns, with the null hypothesis of constancy retained at p-value 0.0798.

The same technique extended to two sample allows one to test hypothesis if there is any change in population learning curve for the 2 years. Figure 11 shows that the zero null hypothesis for the difference of learning curves for August 3, 2014, to August 1, 2015 ($n_1 = 105$) and August 2, 2015 to July 31, 2016 ($n_2 = 74$) is retained at p-value 0.132. That is, there is no significant change on the sales curve from 1 year to the next.

Similarly, Fig. 12 shows that the null hypothesis of female sales curve from August 3, 2014, to August 1, 2015, being the male curve plus −12,350.6 is retained with p-

Table 2 Empirical coverage frequencies of the asymptotic SCB and the bootstrap SCB based on 500 replications generated from model (16) and computing time in minutes (inside parentheses)

n	$1-\alpha$	Case 1		Case 2	
		Asymptotic	Bootstrap	Asymptotic	Bootstrap
50	0.950	0.924 (1.53)	0.924 (7.15)	0.900 (1.54)	0.922 (5.93)
	0.990	0.970 (1.52)	0.978(6.18)	0.966 (1.55)	0.978 (7.32)
100	0.950	0.926 (1.89)	0.940 (6.28)	0.892 (1.86)	0.938 (7.76)
	0.990	0.980 (1.90)	0.990 (6.73)	0.974 (1.94)	0.986 (8.07)
200	0.950	0.928 (2.63)	0.958 (7.40)	0.916 (2.65)	0.952 (9.63)
	0.990	0.986 (2.61)	0.992 (8.14)	0.986 (2.62)	0.994 (9.60)
300	0.950	0.940 (8.72)	0.944 (8.72)	0.926 (3.89)	0.940 (8.74)
	0.990	0.982 (3.82)	0.988 (9.78)	0.980 (3.91)	0.984 (9.37)
n	$1-\alpha$	Case 3		Case 4	
		asymptotic	bootstrap	asymptotic	bootstrap
50	0.950	0.924 (1.46)	0.920 (5.62)	0.896 (1.44)	0.916 (5.94)
	0.990	0.968 (1.49)	0.978 (5.90)	0.968 (1.49)	0.978 (5.67)
100	0.950	0.928 (1.81)	0.942 (6.23)	0.892 (1.79)	0.938 (6.38)
	0.990	0.980 (1.97)	0.990 (6.44)	0.974 (1.86)	0.986 (6.56)
200	0.950	0.928 (2.50)	0.958 (7.23)	0.916 (2.51)	0.952 (8.57)
	0.990	0.986 (2.66)	0.992 (7.61)	0.986 (2.57)	0.994 (8.08)
300	0.950	0.938 (3.68)	0.944 (8.61)	0.926 (3.63)	0.940 (9.51)
	0.990	0.984 (3.77)	0.988 (8.59)	0.978 (3.81)	0.984 (8.99)

value 0.7404, while the constant $-12{,}350.6$ is estimated by the difference in average non-holiday sales, i.e., $(N-\mathcal{V})^{-1}\left\{n_1^{-1}\sum_{i=1}^{n_1}\sum_{j\in\bar{\mathcal{J}}_{\mathcal{V}}}Y_{ij,1} - n_2^{-1}\sum_{i=1}^{n_2}\sum_{j\in\bar{\mathcal{I}}_v}Y_{ij,2}\right\}$. Thus male salespersons on average sales about 12,350 RMB per week more than the female counterpart. In addition, the null hypothesis H_0 : female sales curve from August 3, 2014 to August 1, 2015 being equal to that of male is also tested, with p-value of 0.0104. Therefore, the null hypothesis is rejected, indicating that the female sales curve are significantly different with the male's.

The null hypothesis of sales curve from August 3, 2014, to August 1, 2015, of one brand of footwear being equal to that of another brand is rejected with p-value of 0.0574. Upon further examination, the null hypothesis of sales curve from August 3, 2014 to August 1, 2015 of one brand of footwear being equal to that of another brand plus -8293.6 is retained with p-value 0.4682, and the constant -8293.6 is estimated by the difference in average non-holiday sales as described above. Thus, the average sales of one brand sport footwear was significantly less than those of the other brand sport footwear about 8293 RMB per week (Fig. 13).

The CIs for $\beta_{1v}-\beta_{2v}$ in Theorem 3 are used to conduct a two-sided test on whether the holiday effects differ between two samples, i.e., $H_0:\beta_{1v}-\beta_{2v}=0$ v.s. $H_1:\beta_{1v}-\beta_{2v}\neq 0$.

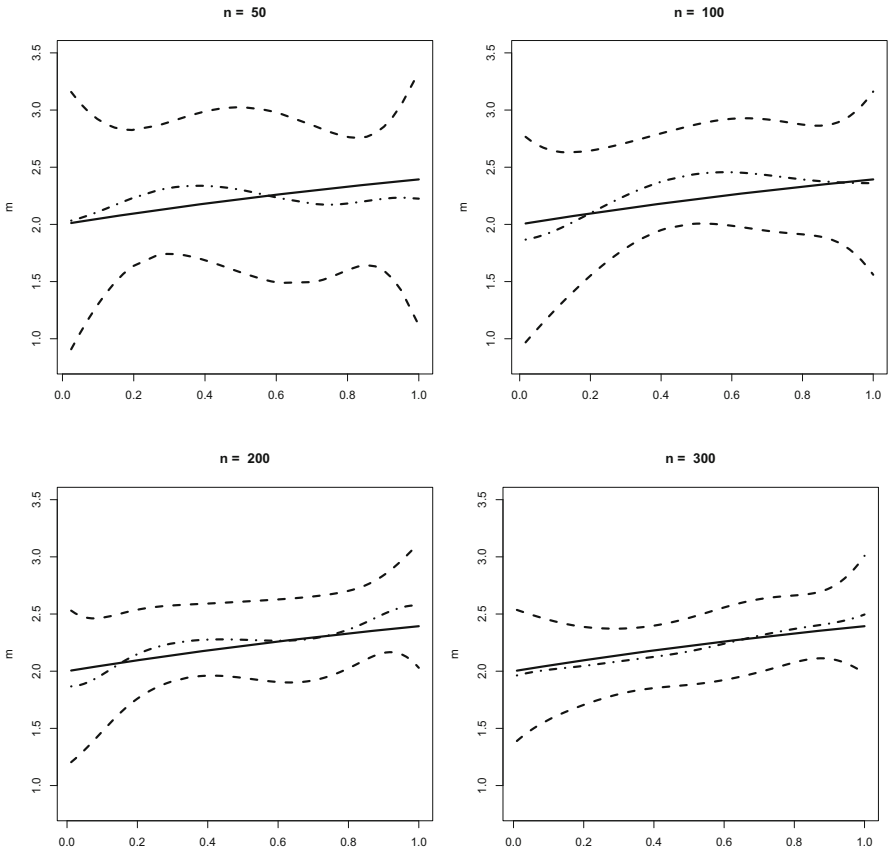

Fig. 6 Plots of 95% SCB (dashed) for $m(x)$ (solid) and cubic spline estimator $\hat{m}(x)$ (dashed-dotted) for Case 1, which are computed by (14) based on data generated from model (16)

For the two data from August 3, 2014 to August 1, 2015 and from August 2, 2015 to July 31, 2016, the p-values are 0.0013, 0.000012, 0.82, 0.0000021, 0.000000001, 0.46 and 0.735 for zero difference of holiday effects on National Day, New Year's Day, Spring Festival and International Workers' Day, respectively. Therefore, except very significant difference in holiday effects for the National Day and Spring Festival, all other holiday effects exhibit no such difference, with p-values greater than 0.46. To visualize, we plot the CIs for $\beta_{1\nu} - \beta_{2\nu}$ with $\nu = 1, 2, \ldots, 7$; see Fig. 14a for the CIs between the 2 years.

Likewise, for the female and male sales data from August 3, 2014 to August 1, 2015, the zero difference null hypotheses is rejected for the first National Day effect and the last spring festival effect with p-values being 0.0000029 and 0.012, while other zero difference null hypotheses are retained with p-values being 0.97, 0.96, 0.71, 0.18 and 0.89, respectively. See Fig. 14b for the CIs. For the two brand data from August 3, 2014 to August 1, 2015, the results are also mixed with p-values being 0.17, 0.062, 0.00086, 0.08, 0.149, 0.147 and 0.05 (see Fig. 14c for the CIs). Hence,

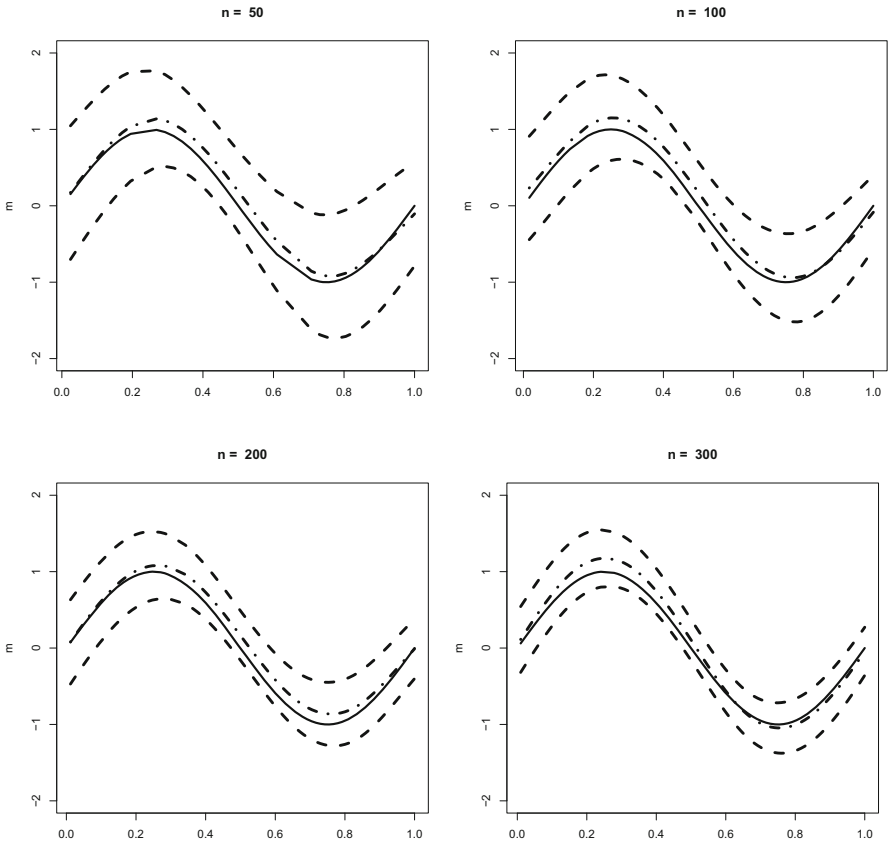

Fig. 7 Plots of 95% SCB (dashed) for $m(x)$ (solid) and cubic spline estimator $\hat{m}(x)$ (dashed-dotted) for Case 3, which are computed by (14) based on data generated from model (16)

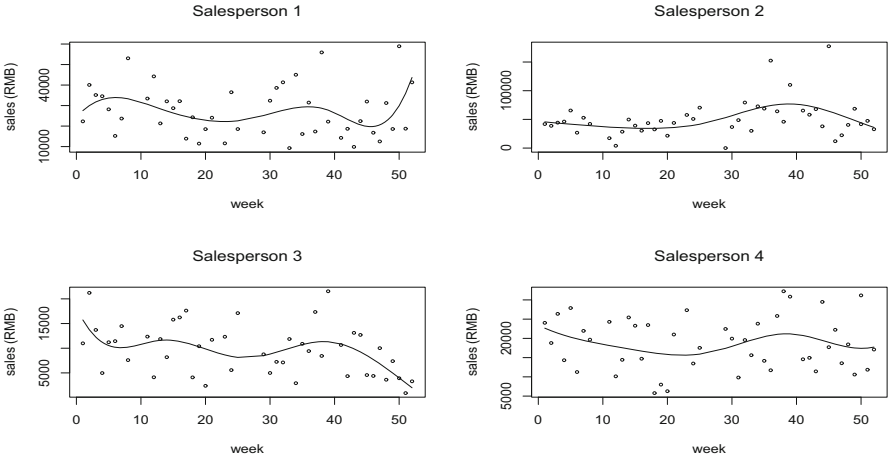

Fig. 8 Plots of cubic spline trajectories of 4 randomly selected salespersons, for sales data from August 2, 2015 to July 31, 2016

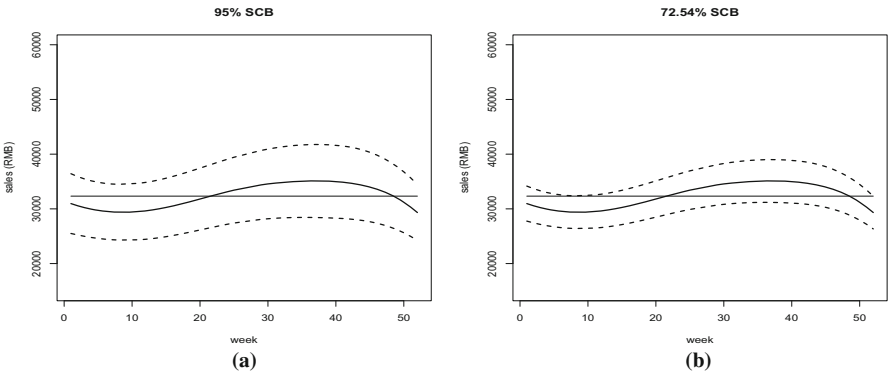

Fig. 9 Plots of the null hypothesis curve $m(x) = 32,326.7$ (solid), cubic spline estimator $\hat{m}(x)$ (thick solid), SCB (dashes) for $m(x)$ with **a** $\alpha = 0.05$ and **b** $\alpha = 0.2746$ for the sales data from August 2, 2015 to July 31, 2016

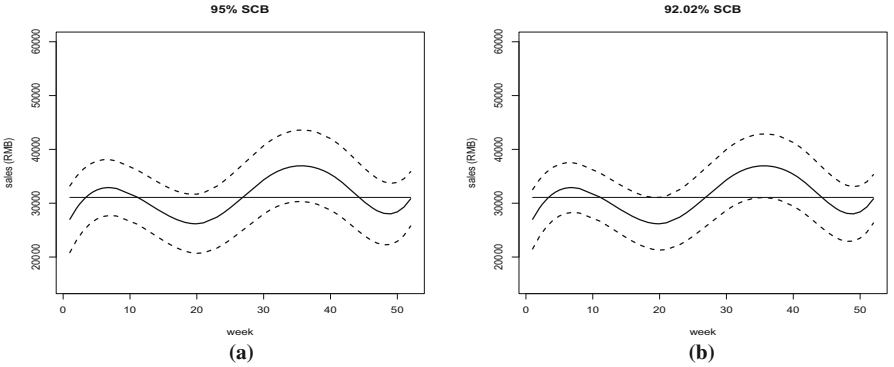

Fig. 10 Plots of the null hypothesis curve $m(x) = 31,077.5$ (solid), cubic spline estimator $\hat{m}(x)$ (thick solid), SCB (dashes) for $m(x)$ with **a** $\alpha = 0.05$ and **b** $\alpha = 0.0798$ for the sales data from August 3, 2014 to August 1, 2015

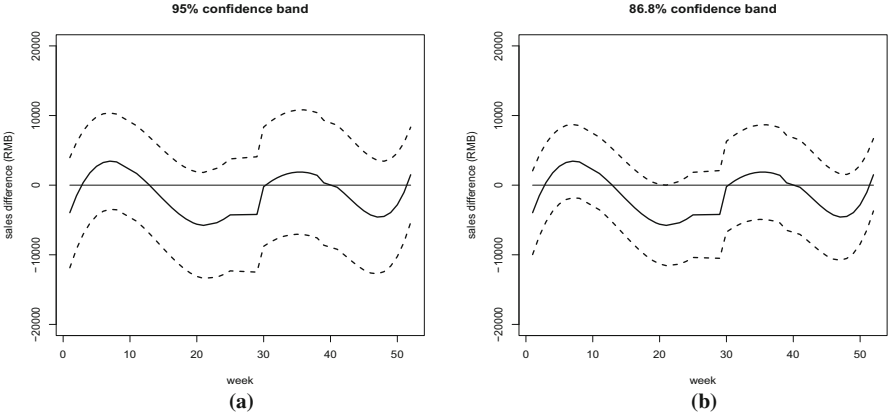

Fig. 11 For data of August 3, 2014 to August 1, 2015 and August 2, 2015 to July 31, 2016, plots of the null hypothesis curve $m_1(x) - m_2(x) = 0$ (solid), cubic spline estimator $\hat{m}_1(x) - \hat{m}_2(x)$ (thick solid), SCB (dashes) for $m_1(x) - m_2(x)$ with **a** $\alpha = 0.05$ and **b** $\alpha = 0.132$

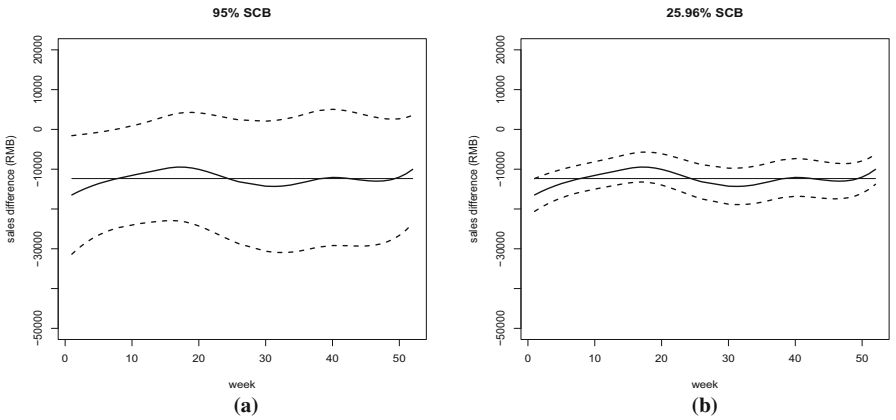

Fig. 12 For data of female and male salespersons from August 3, 2014 to August 1, 2015, plots of the null hypothesis curve $m_1(x) - m_2(x) = -12,350.6$ (solid), cubic spline estimator $\hat{m}_1(x) - \hat{m}_2(x)$ (thick solid), SCB (dashes) for $m_1(x) - m_2(x)$ with (a) $\alpha = 0.05$ and (b) $\alpha = 0.7404$

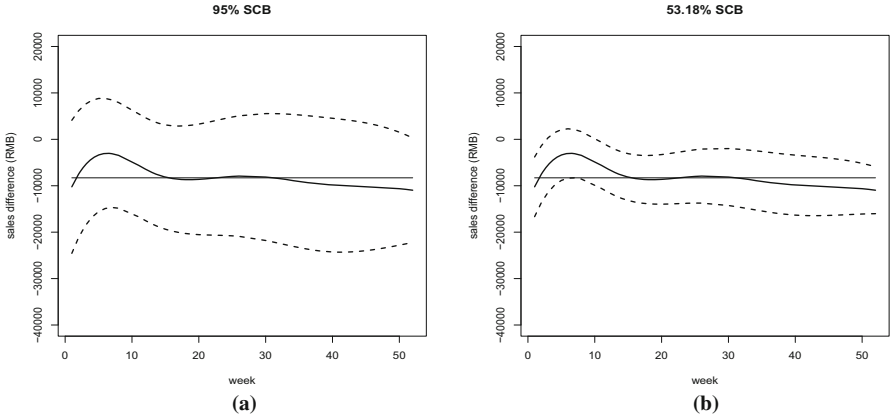

Fig. 13 For sales data of two brands of sports shoes from August 3, 2014 to August 1, 2015, plots of the null hypothesis curve $m_1(x) - m_2(x) = -8293.6$ (solid), cubic spline estimator $\hat{m}_1(x) - \hat{m}_2(x)$ (thick solid), SCB (dashes) for $m_1(x) - m_2(x)$ with (a) $\alpha = 0.05$ and (b) $\alpha = 0.4682$

except significant difference in holiday effect for the New Year's Day, all other holiday effects exhibit no such difference.

5 Conclusions

Motivated by the need to remove holiday spikes from sports footwear sales data, a two-step procedure is proposed to estimate population mean learning curve and holiday effects of dense functional data under holiday effects contamination. Individual trajectories are collectively estimated via spline smoothing over non-holiday time points and used in place of true trajectories. Our theoretical arguments establish that

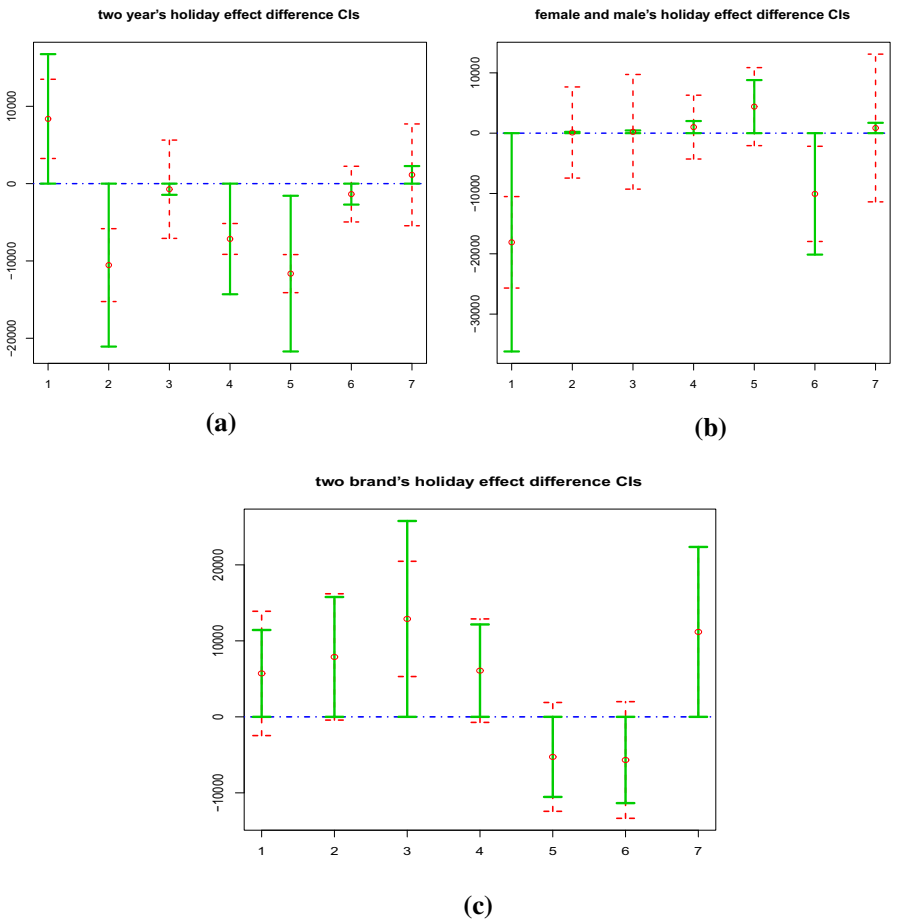

Fig. 14 Plots of the holiday effect difference estimators $\hat{\beta}_{1,\nu} - \hat{\beta}_{2,\nu}$ (points) between two samples with $\nu = 1, 2, \ldots, 7$, the corresponding 95% CIs (dashed), and the $100(1 - p\text{-value})\%$ CIs (solid) with (a): the CIs between the years from August 3, 2014 to August 1, 2015 and from August 2, 2015 to July 31, 2016; (b): the CIs between the female and the male from August 3, 2014 to August 1, 2015; (c): the CIs between the two brands from August 3, 2014 to August 1, 2015

the two-step estimators are oracally efficient in the sense that they are asymptotically equivalent in the order of $n^{-1/2}$ to those with all individual trajectories being known oracally. This oracle efficiency in turn has yielded asymptotic SCBs for the population mean learning curve and CIs for holiday effect parameters. All results are extended to two-sample problems, allowing detection of any significant difference in the mean learning curve and the holiday effects between two samples. Monte Carlo experiments corroborate theoretical findings and application to the weekly sales data has produced a number of interesting discoveries.

Further research may lead to similar procedures for sparse or unbalanced function data with holiday effects, and improved covariance and variance estimation in the presence of holiday effects.

Acknowledgements This research was supported in part by National Natural Science Foundation of China Awards 11371272 and 11771240, and the Tsinghua University Center for Data-Centric Management in the Department of Industrial Engineering. Part of the research was carried out when the first author was a visitor at the Department of Statistics, Texas A & M University. The first author thanks the China Scholarship Council (CSC) for providing financial support to visit Texas A & M University. The helpful comments from Editor-in-Chief Lola Ugarte, an Associate Editor and two Reviewers are gratefully acknowledged.

References

Anzanello M, Fogliatto F (2011) Learning curve models and applications: literature review and research directions. Int J Ind Ergon 41:573–583

Benko M, Härdle W, Kneip A (2009) Common functional principal components. Ann Statist 37:1–34

Bosq D (2000) Linear processes in function spaces: theory and applications. Springer, New York

Cai L, Yang L (2015) A smooth simultaneous confidence band for conditional variance function. TEST 24:632–655

Cai L, Liu R, Wang S, Yang L (2019) Simultaneous confidence bands for mean and variance functions based on deterministic design. Stat Sin 29:505–525

Cao G, Wang L, Li Y, Yang L (2016) Oracle efficient confidence envelopes for covariance functions in dense functional data. Stat Sin 26:359–383

Cao G, Yang L, Todem D (2012) Simultaneous inference for the mean function based on dense functional data. J Nonparametr Statist 24:359–377

Cardot H (2000) Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. J Nonparametr Stat 12:503–538

Cho H, Fryzlewicz P (2015) Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. J R Stat Soc B 77:475–507

Claeskens G, Van Keilegom I (2003) Bootstrap confidence bands for regression curves and their derivatives. Ann Stat 31:1852–1884

de Boor C (1978) A practical guide to splines. Springer, New York

Degras D (2011) Simultaneous confidence bands for nonparametric regression with functional data. Stat Sin 21:1735–1765

Fan J, Huang T, Li R (2007) Analysis of longitudinal data with semiparametric estimation of covariance function. J Am Stat Assoc 102:632–642

Fan J, Lin S (1998) Tests of significance when data are curves. J Am Stat Assoc 93:1007–1021

Fan J, Zhang W (2000) Simultaneous confidence bands and hypothesis testing in varying coefficient models. Scand J Stat 27:715–731

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, New York

Fryzlewicz P, Subba Rao S (2014) Multiple-change-point detection for auto-regressive conditional heteroscedastic processes. J R Stat Soc B 76:903–924

Gu L, Wang L, Härdle W, Yang L (2014) A simultaneous confidence corridor for varying coefficient regression with sparse functional data. TEST 23:806–843

Gu L, Yang L (2015) Oracally efficient estimation for single-index link function with simultaneous confidence band. Electron J Stat 9:1540–1561

Hall P, Müller H, Wang J (2006) Properties of principal component methods for functional and longitudinal data analysis. Ann Stat 34:1493–1517

Huang J, Yang L (2004) Identification of nonlinear additive autoregressive models. J R Stat Soc B 66:463–477

Huang X, Wang L, Yang L, Kravchenko A (2008) Management practice effects on relationships of grain yields with topography and precipitation. Agron J 100:1463–1471

James G, Hastie T, Sugar C (2000) Principal component models for sparse functional data. Biometrika 87:587–602

James G, Sugar C (2003) Clustering for sparsely sampled functional data. J Am Stat Assoc 98:397–408

Komlós J, Major P, Tusnády G (1976) An approximation of partial sums of independent RV's, and the sample DF II. Z. Wahrscheinlichkeitstheorie Verw. Gebiete 34:33–58

Li B, Yu Q (2008) Classification of functional data: a segmentation approach. Comput Stat Data Anal 52:4790–4800

Ma S, Yang L, Carroll RJ (2012) A simultaneous confidence band for sparse longitudinal regression. Stat Sin 22:95–122
Ma S (2014) A plug-in the number of knots selector for polynomial spline regression. J Nonparametr Stat 26:489–507
Raña P, Aneiros G, Vilar JM (2015) Detection of outliers in functional time series. Environmetrics 26:178–191
Rice J, Wu C (2001) Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57:253–259
Schröder AL, Fryzlewicz P (2013) Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. Stat Interface 6:449–461
Song Q, Yang L (2009) Spline confidence bands for variance function. J Nonparametric Stat 21:589–609
Wang J, Liu R, Cheng F, Yang L (2014) Oracally efficient estimation of autoregressive error distribution with simultaneous confidence band. Ann Stat 42:654–668
Wang J, Wang S, Yang L (2016) Simultaneous confidence bands for the distribution function of a finite population and of its superpopulation. TEST 25:692–709
Wang J, Yang L (2009) Polynomial spline confidence bands for regression curves. Stat Sin 19:325–342
Wu W, Zhao Z (2007) Inference of trends in time series. J R Stat Soc B 69:391–410
Yao F, Müller H, Wang J (2005) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100:577–590
Zhang J (2013) Analysis of variance for functional data. Chapman & Hall/CRC, Boca Raton
Zhao Z, Wu W (2008) Confidence bands in nonparametric time series regression. Ann Stat 36:1854–1878
Zheng S, Liu R, Yang L, Hädle W (2016) Statistical inference for generalized additive models: simultaneous confidence corridors and variable selection. TEST 25:607–626
Zheng S, Yang L, Härdle W (2014) A smooth simultaneous confidence corridor for the mean of sparse functional data. J Am Stat Assoc 109:661–673
Zhou S, Shen X, Wolfe D (1998) Local asymptotics of regression splines and confidence regions. Ann Stat 26:1760–1782