Published in final edited form as:
J Stat Plan Inference. 2011 April 1; 141(4): 1554–1566. doi:10.1016/j.jspi.2010.11.003.

Recent History Functional Linear Models for Sparse Longitudinal Data

Kion Kim, **Damla Şentürk***, and **Runze Li**
Department of Statistics, The Pennsylvania State University

Abstract

We consider the recent history functional linear models, relating a longitudinal response to a longitudinal predictor where the predictor process only in a sliding window into the recent past has an effect on the response value at the current time. We propose an estimation procedure for recent history functional linear models that is geared towards sparse longitudinal data, where the observation times across subjects are irregular and total number of measurements per subject is small. The proposed estimation procedure builds upon recent developments in literature for estimation of functional linear models with sparse data and utilizes connections between the recent history functional linear models and varying coefficient models. We establish uniform consistency of the proposed estimators, propose prediction of the response trajectories and derive their asymptotic distribution leading to asymptotic point-wise confidence bands. We include a real data application and simulation studies to demonstrate the efficacy of the proposed methodology.

Keywords

Basis expansion; B-splines; Functional data analysis; Local least squares; Smoothing; Sparse design

1. Introduction

We address estimation in regression modeling of sparse longitudinal data. Sparse designs where the repeated measurements taken on each subject are irregular and the number of repetitions per subject is small, is encountered commonly in applications. An example is the longitudinal primary biliary liver cirrhosis data collected by the Mayo Clinic (see Appendix D of Fleming and Harrington, 1991). Due to missed visits the data is sparse and highly irregular where each patient visited the clinic at different times. We consider estimation in regression models where the sparse longitudinal predictor process only from the recent past has an effect on the sparse response trajectory.

Consider the commonly used functional linear model to relate a response trajectory to a predictor trajectory

$$Y(t)=\alpha(t)+\int_{\Omega_t} X(s)\beta(s,t)ds+\varepsilon(t),\quad t\in[0,T], \tag{1}$$

first introduced by Ramsay and Dalzell (1991). For excellent reviews of functional data analysis, see Ramsay and Silverman (2002; 2005), Rice (2004) and Müller (2005). In (1),

*Corresponding author: dsenturk@stat.psu.edu (Damla Şentürk).

$X(s)$ is the random predictor process, $Y(t)$ is the random response process, and $\varepsilon(t)$ is the mean zero error process with covariance function $\Gamma_\varepsilon(t, t')$, $t, t' \in [0, T]$. When the predictor and response processes are observed over the same time interval, i.e. $\Omega_t = [0, T]$, the entire predictor trajectory is assumed to affect the current value of the response process at time t in the functional linear model. This assumption may not be feasible in many applications and variations of the functional linear model have been considered corresponding to different choices of the support Ω_t. A more suitable choice of Ω_t for predictions was proposed by Malfait and Ramsay (2003) for their recently proposed historical functional linear model, where only the past of the predictor process is involved in predicting the current response. Another special case which had wide applications over the past fifteen years is the varying coefficient model (Cleveland et al., 1992; Hastie and Tibshirani, 1993) with point-wise support where only the current value of the predictor process is assumed to have an effect on the current value of the response process.

An intermediate model considered in this paper between the functional linear model with global support and the varying coefficient model with point-wise support is the recent history functional linear model of Kim et al. (2009)

$$E\{Y(t)|X(s), s \in \Omega_t\} = \alpha(t) + \sum_{k=1}^{K} b_k(t) \int_{\Omega_t} X(s)\varphi_k(s)ds, \tag{2}$$

where the two dimensional regression function $\beta(s, t)$ of the functional linear model is approximated by the product form $\sum_{k=1}^{K} b_k(t)\varphi_k(s)$ of one dimensional functions. In (2), the predictor process from the recent past in a sliding window is assumed to affect the current response value, i.e. $\Omega_t = [t-\delta_1, t-\delta_2]$ for $0 < \delta_2 < \delta_1 < T$. The two delay parameters help define the sliding window support where δ_2 denotes the delay for the predictor process to start affecting the response, and δ_1 is the delay beyond which the predictor process has no effect. Other regression models have also been proposed with a sliding window support such as the generalized varying coefficient model of Şentürk and Müller (2008) and the functional varying coefficient model of Şentürk and Müller (2010), where authors argue that the sliding support is useful in many applications where the response is affected by recent trends in the predictor process. Kim et al. (2009) consider the regression of river flow on rainfall as a motivating example where the river flow level would depend on the recent rainfall but not current rainfall or rainfall from distant past.

Kim et al. (2009) propose an estimation procedure for the recent history functional linear model for densely recorded functional data. In this paper, we propose a new estimation procedure for the recent history functional linear model specifically tailored for sparse longitudinal data. Sparsity is a real challenge in modeling longitudinal data, since nonparametric methods cannot feasibly explain a single trajectory for sparse designs. In addition, while standard semiparametric estimation approaches to longitudinal data have been studied extensively for irregular designs, they are not particularly designed to address sparsity issues and can yield inconsistent results for sparse noise contaminated longitudinal data.

Most recently there have been a number of proposals in the literature to broaden the reach of functional data analysis in general and functional linear models in specific, to include sparsely observed longitudinal data. Shi et al. (1996) proposed fixed and random effects which are linear combinations of B-splines, James et al. (2000) proposed an estimation method based on reduced rank mixed-effect model emphasizing the sparse data case among

others. Yao et al. (2005a) proposed an estimation procedure for the functional linear model with $\Omega_t = [0, T]$, which is applicable to sparse longitudinal data based on functional principal component analysis. The main idea is that information across all the subjects can still be pooled effectively even for sparse data in the estimation of mean functions and covariance surfaces of the random processes, which are needed for functional principle components analysis.

The challenge of estimation based on sparse longitudinal data is intensified when one considers sliding window supports for the regression function such as the support $[t - \delta_1, t - \delta_2]$ of the recent history functional linear model. That is even if the entire observed process is not very sparse, observations in the sliding window can easily get sparse fast as the window size decreases. We address this challenge by drawing connections between the proposed recent history functional linear model and the varying coefficient model and basing the estimation algorithm mainly on the estimation of auto-and cross-covariances following the proposals of Yao et al. (2005a). More specifically, note that for a set of K predetermined basis functions $\varphi_k(s)$, the recent history functional linear model in (2) reduces to a multiple varying coefficient model with K induced predictors $\int_{\Omega_t} X(s)\varphi_k(s)ds$. The unknown varying coefficient functions $b_k(t)$ are then targeted based on a set of covariance representations which can be estimated efficiently based on sparse longitudinal data, pooling information across subjects.

The paper is organized as follows. In Section 2, we introduce the recent history functional linear model. The proposed estimation method is outlined and uniform consistency of the proposed estimators are established in Section 3. The prediction of the response trajectories is also proposed in Section 3 utilizing Gaussian assumptions, where asymptotic distributions are derived leading to point-wise asymptotic confidence bands. In Section 4, we discuss numerical issues in implementation along with the choice of model parameters. We study the finite sample properties of the proposed estimators through simulations given in Section 5. We apply the proposed method to a primary biliary liver cirrhosis longitudinal data in Section 6, to study the dynamic relationship between serum albumin concentration and prothrombin time. Concluding remarks are given in Section 7 and technical details are assembled in an Appendix.

2. Data and model

Taking a functional approach we view the observed longitudinal data as noise contaminated realizations of a random process that produces smooth trajectories. We will reflect sparsity in the following representation through random number of repeated measurements per each subject at random time points. Let (X_i, Y_i), $i = 1, \ldots, n$, be the pairs of square integrable predictor and response trajectories, which are realizations of the smooth random processes (X, Y), defined on a finite and closed interval, $\mathcal{T} = [0, T]$. The smooth random processes have unknown smooth mean functions $\mu_X(t) = EX(t)$ and $\mu_Y(t) = EY(t)$, and auto-covariance functions $G_X(s, t) = \text{cov}\{X(s), X(t)\}$ and $G_Y(s, t) = \text{cov}\{Y(s), Y(t)\}$. Throughout this paper, s and t refer to time indices defined on $\mathcal{T}$. The observed trajectories of the i^{th} subject, $X_{ij} = X(T_{ij}) + \varepsilon_{ij}$, $Y_{ij} = Y(T_{ij}) + e_{ij}$, $j = 1, \ldots, N_i$ are noise contaminated realizations of the random processes X and Y measured at i.i.d. random time points, T_{ij}. Here, ε_{ij} and e_{ij} denote the additive i.i.d. zero mean finite variance measurement errors of the predictor and the response trajectories, respectively. The number of random time points per subject, N_i, are i.i.d. realizations of the random variable N where $P(N > 1) > 0$.

Under mild conditions, auto-covariance functions defined above have orthogonal expansions in terms of eigenfunctions $\psi_m(\cdot)$ and $\phi_p(\cdot)$ with nonincreasing eigenvalues γ_m and η_p,

$$G_X(s,t)=\sum_{m=1}^{\infty}\gamma_m\psi_m(s)\psi_m(t),\quad G_Y(s,t)=\sum_{p=1}^{\infty}\eta_p\phi_p(s)\psi_p(t),\quad \text{for } s,t\in[0,T]. \tag{3}$$

Then, based on Karhunen-Loéve expansion (Ash and Gardner, 1975), the observations X_{ij} and Y_{ij} can be represented as

$$X_{ij}=\mu_X(T_{ij})+\sum_{m=1}^{\infty}\zeta_{im}\psi_m(T_{ij})+\varepsilon_{ij},\qquad Y_{ij}=\mu_Y(T_{ij})+\sum_{p=1}^{\infty}\xi_{ip}\phi_p(T_{ij})+e_{ij}, \tag{4}$$

where ζ_{im} and ξ_{ik} denote the mean zero functional principal component scores with the second moments equal to the corresponding eigenvalues γ_m and η_p for $\sum_{m=1}^{\infty}\gamma_m<\infty$ and $\sum_{p=1}^{\infty}\eta_p<\infty$.

Let Δ_t denote the interval, $[t-\delta_1, t-\delta_2]$, where $0<\delta_2<\delta_1<T$, $t\in[\delta_1, T]$, and let Δ denote the interval $[0, \delta_1-\delta_2]$. Note that the first lag δ_1 is the time point beyond which the predictor function does not have an effect on the response function and the second lag δ_2 allows a delay for the predictor function to start having an effect on the response function.

The recent history functional linear model where $\beta(s,t)=\sum_{k=1}^{K}b_k(t)\varphi_k(s-t+\delta_1)$, can be given as

$$\begin{aligned}E\{Y(t)|X(s),\ s\in\Delta_t\}&=\alpha(t)+\sum_{k=1}^{K}b_k(t)\int_{\Delta_t}X(s)\varphi_k(s-t+\delta_1)ds\\&=\alpha(t)+\sum_{k=1}^{K}b_k(t)\tilde{X}_k(t),\end{aligned} \tag{5}$$

for K predetermined basis functions $\varphi_k(t)$ defined on Δ. In (5) $\tilde{X}_k(t)=\int_{\Delta_t}X(s)\varphi_k(s-t+\delta_1)ds$, $k=1,\ldots,K$, are the induced covariates and $b_k(t)$, $k=1,\ldots,K$, are the unknown time varying coefficient functions of interest. Defining $\tilde{X}(t)=[\tilde{X}_1(t),\ldots,\tilde{X}_K(t)]^{\mathrm{T}}$ and $b(t)=[b_1(t),\ldots,b_K(t)]^{\mathrm{T}}$, the model in (5) can be rewritten in vector form as $E\{Y(t)|X(s),\ s\in\Delta_t\}=\alpha(t)+b^{\mathrm{T}}(t)\tilde{X}(t)$.

In (5), K controls the resolution of the fit and should be chosen based on the data. Depending on the specific features of the regression function, various basis functions such as Fourier, truncated power, eigen and B-spline basis can be used in (5). Because of their fast computation and good properties, we will use B-spline basis in the following calculations. For more discussions on the B-spline basis, see Fan and Gijbels (1996) and Ramsay and Silverman (2005).

3. Estimation and asymptotic properties

3.1. Proposed estimation algorithm

In the proposed estimation procedure, we utilize connections of the proposed model to varying coefficient models. There are three main estimation methods for varying coefficient models proposed in the literature: local polynomial smoothing (Wu et al., 1998; Hoover et al., 1998; Fan and Zhang, 2000; 2008; Kauermann and Tut, 1999), polynomial spline

(Huang et al., 2002; 2004; Huang and Shen, 2004) and smoothing spline (Hastie and Tibshirani, 1993; Hoover et al., 1998; Chiang et al., 2001).

Note that the previously proposed methods cannot be directly employed here, since the induced covariates, $\tilde{X}_k(t)$'s in (5) cannot be estimated well for sparse designs due to the difficulty in numerically approximating the integral in their definition. In fact, the induced covariates may not be well approximated not only in sparse designs but also in longitudinal data in general, since the integration involved is over a narrow window into the past, where there may not be enough points. We propose to base the estimation on the covariance structure to address this difficulty, which will be shown to adjust for measurement error in predictors as well.

Let $\tilde{X}_{ik}(t)$ denote observation of the k^{th} covariate in (5) for the i^{th} subject taken at time point t, i.e., $X_{ik}(t) = \int_{\Delta_t} X_i(s)\varphi_k(s)ds$, and $\tilde{X}_i(t) = [\tilde{X}_{i1}(t), \ldots, \tilde{X}_{iK}(t)]^{\mathrm{T}}$. The proposed estimation rests on the following equality that follows from (5)

$$v(t)b(t)=\theta(t), \tag{6}$$

where $\theta(t)$ is the $K \times 1$ vector with the k^{th} element equal to $\mathrm{cov}\{\tilde{X}_k(t), Y(t)\}$, $v(t)$ is the $K \times K$ matrix with the $(k, \ell)^{th}$ element equal to $\mathrm{cov}\{\tilde{X}_k(t), X_\ell(t)\}$ and $b(t)$ is the $K \times 1$ vector of K varying coefficient functions. The elements $v_{k\ell}(t)$ and $\theta_k(t)$ can be given in terms of auto- and cross-covariance functions of the predictor and response processes as

$$v_{k\ell}(t)=\mathrm{cov}\{\tilde{X}_k(t), \tilde{X}_\ell(t)\}=\int_{\Delta_t}\int_{\Delta_t} G_X(s_1, s_2)\varphi_k(s_1 - t+\delta_1)\varphi_\ell(s_2 - t+\delta_1)ds_1 ds_2 \tag{7}$$

and

$$\theta_k(t)=\mathrm{cov}\{\tilde{X}_k(t), Y(t)\}=\int_{\Delta_t} G_{XY}(s, t)\varphi_k(s - t+\delta_1)ds, \tag{8}$$

where $G_{XY}(s, t) = \mathrm{cov}\{X(s), Y(t)\}$. The estimation of the components given in (7) and (8) begins with estimation of the mean functions, $\mu_X(t)$ and $\mu_Y(t)$, via locally smoothing the aggregated data (T_{ij}, X_{ij}) and (T_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N_i$. For the estimation of the auto- and cross-covariance functions, we apply two-dimensional local linear smoothing to the raw auto-and cross-covariances defined as $G_{X,i}(T_{ij}, T_{ij'}) = \{X_{ij} - \hat{\mu}_X(T_{ij})\}\{X_{ij'} - \hat{\mu}_X(T_{ij'})\}$, $G_{XY,i}(T_{ij}, T_{ij'}) = \{X_{ij} - \hat{\mu}_X(T_{ij})\}\{Y_{ij'} - \hat{\mu}_Y(T_{ij'})\}$, $i = 1, \ldots, n, j, j' = 1, \ldots, N_i$, respectively. To obtain smooth estimates, the raw covariances are fed into a two dimensional local smoothing algorithm where special care needs to be taken in estimating the auto-covariance surface. Since the only terms in the raw auto-covaraince matrix that are perturbed by the additive measurement error on the predictors are along the diagonal, we remove the diagonal before the application of two dimensional smoothing, following (Yao et al., 2005a). Explicit forms of mean and covariance function estimators are given in Appendix C. The smoothing parameters used in the one and two dimensional smoothing procedures for the estimation of the mean functions and covariance surfaces respectively, can be chosen with respect to one-curve-leave-out cross-validation (Rice and Silverman, 1991). For computational efficiency we utilize generalized cross-validation (Liu and Müller, 2008) in simulation studies and data analysis. Plugging the covariance function estimators into equations (7) and (8), and

performing numerical integrations, we can obtain the estimators of $\theta(t)$ and $v(t)$, $\hat{\mu}(t)$ and $\hat{v}(t)$, respectively. Then the estimator for varying coefficient vector, $b(t)$, is defined as

$$\widehat{b}(t)=\widehat{v}^{-1}(t)\widehat{\theta}(t),$$

leading to the final estimator $\widehat{\beta}(s,t)=\sum_{k=1}^{K}\widehat{b}_k(t)\varphi_k(s)$ for the regression surface $\beta(s,t)=\sum_{k=1}^{K}b_k(t)\varphi_k(s)$.

There are two major advantages of the proposed estimation procedure. The first advantage is that, since it depends on estimation of the mean functions and covariance surfaces which are estimated from the entire data, it enables us to surmount the sparsity of the design by pooling information. The second is that it naturally adjusts for measurement error in the predictor process by considering the covariance structure and removing the diagonal terms before smoothing. In addition note that we can get estimates of the covariance surfaces on a fine set of grid points through smoothing, which allows precise numerical integration approximations in the estimation of $\theta(t)$ and $v(t)$.

For a given number of components, K, uniform consistency of the proposed estimator is established by the below Theorem.

Theorem 1—Under Assumptions (A.1)–(A.5) given in Appendix A,

$$\sup_{s,t\in\Delta_t\times[\delta_1,T]}|\widehat{\beta}(s,t)-\beta(s,t)|=O_p\left\{\frac{1}{\sqrt{n}}\left(\frac{1}{h_X^2}+\frac{1}{h_1h_2}\right)\right\}.$$

Here, h_X is the bandwidth used in obtaining the smooth auto-covariance surface of the predictor process and, h_1 and h_2 are used in obtaining the cross-covariance surface between the response and predictor processes. The set of bandwidths depend on n and are all required to converge to 0 as $n\to\infty$, for further details see Appendix A.

3.2. Prediction of response trajectories

We will establish the prediction of a new response trajectory, Y^*, based on the sparse predictor trajectory X^*. In what follows, let us define $P_m(t)=\int_{\Delta_t}\beta(s,t)\psi_m(s)ds=\sum_{k=1}^{K}b_k(t)\int_{\Delta_t}\varphi_k(s)\psi_m(s)ds$, where $\psi_m(s)$ are the eigen-functions of the auto-covariance of the predictor process defined in (3). From the functional representation of the predictor trajectory given in (4), the predicted response trajectory can be given as

$$E\{Y^*(t)|X^*(s),\ s\in\Delta_t\}=\mu_Y(t)=\sum_{m=1}^{\infty}\zeta_m^*P_m(t), \tag{9}$$

where $\zeta_m^*=\int_{\mathcal{T}}\{X^*(s)-\mu_X(s)\}\psi_m(s)ds$ is the m^{th} functional principal component score of the predictor function X^*. For estimation of (9), $\mu_Y(t)$ can be estimated from aggregated data as described above and estimators of the eigenfunctions, $\psi_m(t)$, can be obtained from the eigen-

decomposition of the estimated auto-covariance surface. Details on these estimation procedures are given in Appendix C. For estimation of ζ_m^* in sparse designs, we invoke Gaussian assumptions following Yao et al. (2005b).

Let us define the collection of N^* error contaminated observations from the predictor trajectory as $X'^* = (X_1^*, \ldots, X_{N^*}^*)^{\mathrm{T}}$, where X'^*_ℓ is the ℓ^{th} measurement observed at time point T_ℓ^*. We assume that $(\zeta_m^*, \varepsilon_\ell^*)$, for $\ell = 1, \ldots, N^*$, are jointly Gaussian. Further define $X^* = \{X^*(T_1^*), \ldots, X^*(T_{N^*}^*)\}^{\mathrm{T}}$, $\mu_X^* = \{\mu_X^*(T_1^*), \ldots, \mu_X^*(T_{N^*}^*)\}^{\mathrm{T}}$, $\psi_m^* = \{\psi_m^*(T_1^*), \ldots, \psi_m^*(T_{N^*}^*)\}^{\mathrm{T}}$ and $T^* = (T_1^*, \ldots, T_{N^*}^*)^{\mathrm{T}}$. Under the Gaussian assumption, the best linear predictor for ζ_m^* given X'^*, N^* and T^* is obtained by

$$\tilde{\zeta}_m^* = \gamma_m \psi_m^{*T} \sum\nolimits_{X'^*}^{-1} (X'^* - \mu_X^*), \tag{10}$$

where $\sum_{X'^*} = \mathrm{cov}(X'^* | N^*, T^*) = \mathrm{cov}(X^* | N^*, T^*) + \sigma_\varepsilon^2 I_{N^*}$ with I_{N^*} denoting the $N^* \times N^*$ identity matrix and $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon)$ denoting the variance of the measurement error on the predictor process. Defining $\hat{\mu}_X^*$ and $\hat{\psi}_m^*$ to be the estimators of $\hat{\mu}_X^*$ and $\hat{\psi}_m^*$ respectively, the estimator of $\tilde{\zeta}_m^*$ can be given as

$$\hat{\zeta}_m^* = \hat{\gamma}_m \hat{\psi}_m^{*T} \hat{\sum}\nolimits_{X'^*}^{-1} (X'^* - \hat{\mu}_X^*),$$

where the $(i, j)^{th}$ element of $\hat{\Sigma}_{X'^*}$ is defined as $\widehat{\mathrm{cov}}\{X^*(T_i^*), X^*(T_j^*)\} + \hat{\sigma}_\varepsilon^2 \delta_{ij}$ where $\hat{\sigma}_\varepsilon^2$ is the estimator of the measurement error variance and $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ otherwise. Explicit form of $\hat{\sigma}_\varepsilon^2$ is deferred to Appendix C. Hence the predicted response trajectory is obtained by

$$\hat{Y}_M^*(t) = \hat{\mu}_Y(t) + \sum_{m=1}^{M} \hat{\zeta}_m^* \hat{P}_m(t), \tag{11}$$

where $\hat{P}_m(t) = \int_{\Delta_t} \hat{\beta}(s,t) \hat{\psi}_m(s) ds = \sum_{k=1}^{K} \hat{b}_k(t) \int_{\Delta_t} \varphi_k(s) \hat{\psi}_m(s) ds$. The number M of eigenfunctions used in the decomposition of the predictor auto-covariance surface, given in (11) can be selected by leave-one-curve-out cross validation, generalized cross validation (GCV), or the Akiake information criterion (AIC). For more details on the selection of M, see Yao et al. (2005a). A consistency result of the predicted trajectory for the target trajectory $\tilde{Y}^*(t) = \mu_Y(t) + \sum_{m=1}^{\infty} \tilde{\zeta}_m^* P_m(t)$ is established in Theorem 2.

Theorem 2—Under (A1)–(A5), (B1)–(B3) of Appendix A, given T^* and N^*, for all $t \in [\delta_1, T]$, the predicted trajectories satisfy

$$\lim_{n\to\infty} \widehat{Y}_M^*(t) = \tilde{Y}^*(t) \quad \text{in probability}.$$

Note that, the number M of eigenfunctions used in the eigen-decomposistion of the predictor process is a function of n and tends to infinity as $n \to \infty$.

3.3. Asymptotic confidence bands for the predicted response trajectories

In this section, we construct point-wise asymptotic confidence intervals for the predicted response trajectory, $\widehat{Y}_M^*(t)$. For $M \geq 1$, let us define $\zeta^{*M} = (\zeta_1^*, \ldots, \zeta_M^*)^{\mathrm{T}}$, $\tilde{\zeta}^{*M} = (\tilde{\zeta}_1^*, \ldots, \tilde{\zeta}_M^*)^{\mathrm{T}}$, where $\tilde{\zeta}_m^*$ is as defined in (10). Note that $\mathrm{cov}(\zeta_m^*, X'^*) = \gamma_m \psi_m^*$. Defining the $M \times N^*$ matrix $H = \mathrm{cov}(\zeta^{*M}, X'^* | T^*, N^*) = (\gamma_1 \psi_1^*, \ldots, \gamma_M \psi_M^*)^{\mathrm{T}}$, covariance matrix of $\tilde{\zeta}^{*M}$ can be given in terms of H as $\mathrm{cov}(\tilde{\zeta}^{*M} | T^*, N^*) = H \sum_{X'^*}^{-1} H^{\mathrm{T}}$. Observing that $\mathrm{cov}(\tilde{\zeta}^{*M}, \zeta^{*M} | T^*, N^*) = H \sum_{X'^*}^{-1} \mathrm{cov}(X'^*, \zeta^{*M}) = H \sum_{X'^*}^{-1} H^{\mathrm{T}}$, we have $\mathrm{cov}(\tilde{\zeta}^{*M} - \zeta^{*M} | T^*, N^*) = \mathrm{cov}(\tilde{\zeta}^{*M} | T^*, N^*) + \mathrm{cov}(\zeta^{*M} | T^*, N^*) - 2\mathrm{cov}(\tilde{\zeta}^{*M}, \zeta^{*M} | T^*, N^*) = \mathrm{cov}(\zeta^{*M} | T^*, N^*) - \mathrm{cov}(\tilde{\zeta}^{*M} | T^*, N^*) \equiv \Omega_M$, where $\Omega_M = D - H \Sigma_{X'^*} H^{\mathrm{T}}$ with $D = \mathrm{diag}\{\gamma_1, \ldots, \gamma_M\}$. Hence, under the Gaussian assumption, conditioning on T^* and N^*, $\tilde{\zeta}^{*M} - \zeta^{*M}$ is distributed as $N(0, \Omega_M)$.

Let $\widehat{\zeta}^{*M} = (\widehat{\zeta}_1^*, \ldots, \widehat{\zeta}_M^*)^{\mathrm{T}}$ and $\widehat{\Omega}_M = \widehat{D} - \widehat{H} \widehat{\sum}_{X'^*}^{-1} \widehat{H}^{\mathrm{T}}$, where $\hat{D} = \mathrm{diag}\{\hat{\gamma}_1, \ldots, \hat{\gamma}_M\}$ and $\widehat{H} = (\widehat{\gamma}_1 \widehat{\psi}_1^*, \ldots, \widehat{\gamma}_M \widehat{\psi}_M^*)^{\mathrm{T}}$. Defining $\hat{P}(t) = \{\hat{P}_1(t), \ldots, \hat{P}_M(t)\}^{\mathrm{T}}$, Theorem 3 gives the asymptotic distribution of the predicted response trajectory $\widehat{Y}_M^*(t) = \widehat{\mu}_Y(t) + \widehat{P}^{\mathrm{T}}(t) \widehat{\zeta}^{*M}$.

Theorem 3—Under (A1)–(A5), (B1)–(B3), and (C1) of Appendix A, given N^* and T^*, for all $t \in [\delta_1, T]$ and $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left\{ \frac{\widehat{Y}_M^*(t) - E\{Y^*(t) | X^*(s) \in \Delta_t\}}{\sqrt{\widehat{\omega}_M(t,t)}} \leq x \right\} = \Phi(x), \tag{12}$$

where $\omega_M(t, t) = P^{\mathrm{T}}(t) \Omega_M P(t)$, $\hat{\omega}_M(t, t) = \hat{P}^{\mathrm{T}}(t) \hat{\Omega}_M \hat{P}(t)$ and Φ denotes the Gaussian cdf.

From Theorem 3, it follows that ignoring the bias from the truncation in $\widehat{Y}_M^*$ at M eigen-components, the $(1 - \alpha)100(\%)$ asymptotic pointwise confidence interval for $E\{Y^*(t) | X^*(s), s \in \Delta_t\}$ is

$$\widehat{Y}_M^*(t) \pm \Phi\left(1 - \frac{\alpha}{2}\right) \sqrt{\widehat{P}_M^{\mathrm{T}}(t) \widehat{\Omega}_M \widehat{P}_M(t)}.$$

4. Numerical issues in implementation and parameter selection

An important issue in the implementation of the proposed estimation algorithm is the inversion of the matrix $\hat{v}(t)$ in equation (6). To obtain stable estimators for β, penalized

solutions have been studied in literature (Cardot et al., 2003) minimizing the penalized least squares

$$\{\widehat{v}(t)b(t) - \widehat{\theta}(t)\}^{\mathrm{T}}\{\widehat{v}(t)b(t) - \widehat{\theta}(t)\} + \lambda b^{\mathrm{T}}(t)Qb(t),$$

where Q is a $K \times K$ matrix that determines the type of penalty used. Here, λ is a tuning parameter that controls the amount of regularization and should be chosen from data balancing the stability and validity of the resulting estimator. The penalized estimator is then given by $\hat{b}_\lambda(t) = \{\hat{v}(t) + \lambda Q\}^{-1}\hat{\mu}(t)$. A common choice of Q is the $K \times K$ identity matrix, which leads to the ridge solution, $\hat{b}_\lambda(t) = \{\hat{v}(t) + \lambda I\}^{-1}\hat{\mu}(t)$. Another common choice of penalty used is on the degree of smoothness of the penalized solution where Q is chosen such that its $(i, j)^{th}$ element is equal to $\int_\Delta \varphi_i^{(m)}(s)\varphi_j^{(m)}(s)ds$ with $\varphi_k^{(m)}(t)$ denoting the m^{th} derivative of the k^{th} predetermined basis function. In the following applications, we use the ridge solution.

Hence, the proposed estimation procedure for the recent history functional linear model, includes three sets of parameters with different roles: K, λ, and (δ_1, δ_2). The number of predetermined basis functions used, K, controls the resolution of the fit; the tuning parameter, λ, controls the stability of the estimates; and the window, (δ_1, δ_2), determines the model used via controlling the predictor window affecting the response. An important observation in the current estimation set-up is that the proposed estimation procedure is not sensitive to the choice of K provided that there are enough number of basis functions used in the estimation, since the penalized solution employed via λ prevents over-fitting. Instead the choice of the tuning parameter λ is a more important choice controlling the stability of the estimates. This fact has been pointed out before in similar functional estimation problems, see Cardot et al. (2003) for further details. We run multiple simulation studies comparing the estimated regression surfaces obtained with different choices of K and λ to confirm this observation, where the results are summarized in Section 5.2. Based on the above argument, following Cardot et al. (2003), we fix K at a value that guarantees good precision and concentrate on the choice of λ and (δ_1, δ_2). For example, K is fixed at 10 in both the simulation studies and the data example that follow, to include 10 B-spline basis functions of order 4 with 6 interior equi-distance knots in the window Δ, which prove to provide good precision.

For the selection of λ and (δ_1, δ_2), consider the following two criteria. Define the normalized prediction error (NPE) as

$$\mathrm{NPE}\{\lambda;(\delta_1, \delta_2)\} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i^\delta}\frac{|\widehat{Y}_{ij} - Y_{ij}|}{|Y_{ij}|},$$

where $\hat{Y}_{ij}$ is the predicted value for the j^{th} measurement on the i^{th} response trajectory obtained using λ and (δ_1, δ_2), N_i^δ is the number of observations from the i^{th} subject obtained in the interval, $[\delta_1, \mathrm{T}\,]$, and $N = \sum_{i=1}^{n} N_i^\delta$. Note that NPE measures the relative absolute prediction error which will also be used in evaluating the finite sample performance of the proposed estimates in the simulations given in Section 5. We also define leave-one-curve-out cross validation squared prediction error as

$$CV\{(\delta_1, \delta_2);\lambda\} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i^{\delta}}\{\widehat{Y}_{ij}^{(-i)} - Y_{ij}\}^2},$$

where $\widehat{Y}_{ij}^{(-i)}$ is the fitted response function value of the i^{th} subject at time point T_{ij} obtained from the data excluding the i^{th} subject. Note that in both criteria, NPE and the cross validation score, the fitted response values are obtained via the prediction methods proposed in Section 3.2.

We propose to estimate λ and (δ_1, δ_2) in a hierarchical manner, where first λ is chosen for a set of (δ_1, δ_2) values using NPE and finally the set of (δ_1, δ_2) values are compared to yield the final choice of (δ_1, δ_2) with minimum cross validation score. This hierarchical approach has the advantage that the computationally faster NPE criterion is used within the inside loop of choosing λ and that the relatively more refined cross validation score is utilized for the selection of (δ_1, δ_2), which is similar to model selection in an outer loop. More specifically, the algorithm can be outlined as follows. Let Λ and D be the predetermined sets of λ and (δ_1, δ_2) values considered. For a fixed $(\delta_1, \delta_2)_0 \in D$, NPE values are calculated for all $\lambda \in \Lambda$. The λ that gives the smallest NPE value is selected as the optimal λ for the given $(\delta_1, \delta_2)_0$ and is thus used for calculating the cross validation score for $(\delta_1, \delta_2)_0$. Repeating the above steps for all $(\delta_1, \delta_2) \in D$, the optimal (δ_1, δ_2) is selected to be the one with the minimum cross validation score.

5. Simulation studies

In this section, we investigate the finite sample properties of the proposed estimators through two simulation studies. In the first simulation, we study efficiency of the NPE criterion for selecting the tuning parameter λ for a fixed (δ_1, δ_2) choice, along with the finite sample performance of the proposed estimator based on the selected λ from the algorithm. In the second simulation, we evaluate the performance of the cross validation method for choosing (δ_1, δ_2) with varying sample size. The following simulation results are reported based on 500 Monte Carlo runs. Bandwidths used in the smoothing of the mean and covariance functions are chosen by generalized cross-validation.

5.1. Data generation

For n subjects, the number of measurements made on the i^{th} predictor and response processes, N_i, are randomly selected from 3, 4, and 5. The design points, $T_{i1}, \ldots, T_{iN_i}$ are randomly selected from the uniform distribution between 0 and 50. At a given time point t, the predictor function is evaluated around the mean function $t + \sin(t)$ using two principal components, $\psi_1(t) = -\cos(\pi t/50)/\sqrt{5}$ and $\psi_2(t) = \sin(\pi t/50)/\sqrt{5}$ for $t \in [0, 50]$. The corresponding two eigen component scores, ζ_1 and ζ_2, are independently sampled from the Gaussian distributions with mean 0 and the variances, $\rho_1 = 4$ and $\rho_2 = 1$, respectively. Hence, the predictor process at time t is generated via $X(t) = t + \sin(t) + \sum_{m=1}^{2}\zeta_m\psi_m(t)$. In addition, i.i.d. measurement error, ε, simulated from the Gaussian distribution with mean 0 and variance $\sigma_\varepsilon^2 = 0.025$, is added to the predictor observations in accordance with (4). The regression function is generated based on the same basis functions used for the predictor process via, $\beta(s, t) = \sum_{i=1}^{2}\sum_{j=1}^{2} c_{ij}\psi_i(s)\psi_j(t)$, $s \in \Delta_t$, where $c_{11} = 2$, $c_{12} = 2$, $c_{21} = 1$, and $c_{22} = 2$. The true window combination, (δ_1, δ_2), is set at (20, 0). The error function, $\varepsilon(t)$, is also generated using the same basis functions as the predictor process with eigen-component

scores independently sampled from the Gaussian distributions with mean 0 and variance 0.025, 0.004, respectively. The response function, $Y(t)$ is generated by the equation, $\int_{\Delta_t} \beta(s, t)X(s)ds+\varepsilon(t)$ using the numerical integration procedure. To obtain a noisy version of the response function, we add i.i.d measurement error, e, generated from the Gaussian distribution with mean 0 and variance 0.025.

5.2. Simulation results

To evaluate the performance of the λ selection criterion, NPE, let us define relative square deviation (RSD) at time t as

$$\mathrm{RSD}(t)=\frac{\int_{\Delta_t}\{\widehat{\beta}(s,t)-\beta(s,t)\}^2 ds}{\int_{\Delta_t}\beta(s,t)^2 ds}.$$

The RSD measures the relative size of the squared difference between the estimated and true regression functions at time t. The RSD integrated over the entire support, $t \in [\delta_1, T]$, will be called integrated RSD, denoted IRSD.

Before reporting on results from the two simulation set-ups, we present the true and estimated regression functions with different λ and K choices to demonstrate the relative importance of λ over K in the estimation procedure discussed in Section 4. The true and estimated regression functions are plotted in Figure 1 from three Monte-Carlo simulation runs at $n = 200$ with (K, λ) equal to (5, 7.4) (Figure 1b, for the NPE minimizer ($\lambda = 7.4$)), (5, 1) (Figure 1c) and (10, 7.4) (Figure 1d), respectively. While the difference between the estimators plotted in Figures 1b and 1d is small corresponding to doubling of the K choice, the estimator given in Figure 1c at the wrong λ value of 1 cannot recover the true regression function with IRSD= 298. This confirms that the choice of λ plays a more important role in the proposed estimation algorithm than the choice of K.

The estimated regression function $\hat{\beta}(s, t)$ with the tuning parameter λ equal to the minimizer of IRSD can be thought of as the "optimal estimate", since information on the true regression function is utilized in the comparison. Hence, this estimate can only be obtained in a simulation setting where the true regression function is known. We compare this choice of λ with the minimizer of NPE, which is the only estimate that can be obtained from the data in reality. The performance of the two estimators are compared in Figure 2 in terms of IRSD, where boxplots of estimated IRSD values are given for the optimal and proposed estimates at $n = 200$ and 500 from 500 Monte-Carlo simulation runs. In the displayed boxplots, 22 and 12 outliers are removed for $n = 200$ and $n = 500$, respectively.

Figure 2 suggests that the estimator with the λ choice selected via NPE and the one with the optimal choice of λ, both improve with increasing sample size. For the proposed estimators with λ chosen by NPE, the median estimated IRSD is 0.2937 at $n = 200$ and is 0.20670 at $n = 500$, which implies that the estimated regression surface is close to the true one. We also give in Figure 2 the boxplots of NPE values of the proposed estimator with λ chosen by NPE. The median NPE value at $n = 200$ is 0.3196 and the median NPE drops to 0.2928 for $n = 500$.

The performance of leave-one-curve-out cross validation score for the selection of δ is studied through the second simulation. For the computational efficiency, we use 10 fold cross validation. The true value of (δ_1, δ_2), is set to be (20, 0), where 6 candidate (δ_1, δ_2) pairs are considered at (30, 0), (30, 5), (20, 0), (20, 5), and (10, 5). The correct (δ_1, δ_2)

choice ratio out of 500 Monte Carlo runs are 0.8929 and 0.9214 for $n = 200$ and 500, respectively. There seems to be improvement with increasing sample size.

6. Data analysis

To demonstrate the proposed method, we include an application to the longitudinal primary biliary liver cirrhosis data collected between January 1974 and May 1984 by the Mayo Clinic. In the study design, patients were scheduled to visit the clinic at six months, one year and annually thereafter post diagnosis, where certain blood characteristics were recorded. However, due to missed visits, the data is sparse and highly irregular where each patient visited the clinic at different times. We explore the dynamic relationship between serum albumin level in mg/dl (predictor) and prothrombin time in seconds (response). Both variables are used as an indicator of the liver function, where a decrease in serum albumin levels and elevated prothrombin times are typically associated with malfunctioning of the liver (Murtaugh et al., 1994). We include 201 female patients in the analysis where predictor and response measurements before 2500 days are considered. The number of observations per subject ranges from 1 to 9, with a median of 5 measurements. Individual trajectories of the serum albumin level and prothrombin time overlaying their respective estimated mean functions are given in Figure 3. The generalized cross-validation choice of smoothing bandwidths used in estimation of the prothrombin time and serum albumin level mean functions are 1100 and 1250, respectively. Generalized cross-validation choice for smoothing bandwidths of auto- and cross-covariance surfaces are (400, 400) and (550, 550), respectively. The estimated mean functions indicate opposite patterns as expected, where there is a decreasing trend for the predictor process and an increasing trend for the response.

We next fit the proposed recent history functional linear model to the data with $K = 10$ B-spline functions. Among the five candidate (δ_1, δ_2) choices considered, [1000, 0], [500, 0], [700, 200], [1000, 200], and [1000, 500], the minimizer of cross-validation error was [1000, 0], and the NPE choice of λ was 6502. To save on computational time, we report results from 10 fold cross validation.

The estimated regression surface is displayed in Figure 4 viewed from two different angles. Most of the estimated regression surface is negative stressing the general opposing trends also observed in literature between serum albumin levels and prothrombin time. For a given time point, the albumin concentration has the strongest effect in magnitude on prothrombin time with a delay of about 500 days where the affect decays as the lag increases. Note also that the observed negative effect of the past albumin concentration levels on the current prothrombin time seems to get more pronounced towards the later stages of the study and hence the disease.

Predicted response trajectories for four randomly selected subjects obtained from the proposals of Section 3.2 and 3.3 and the corresponding 95% asymptotic confidence bands are given in Figure 5. We also include for comparison predicted trajectories obtained from the functional linear model proposed by Yao et al. (2005a). The predicted trajectories seem to be quite close to those from a functional linear model fit which uses the entire predictor trajectory including observations from future and distant past measurement times in the predictions as well. Hence, we conclude that the recent history functional linear model which models the effects of the predictor process from the recent past of 1000 days till the present time, provide a reasonable model for the data in terms of prediction. Given its ease in interpretation due to its restricted regression support when compared with the full functional linear model, the proposed model merges as a viable alternative for the analysis of the current data set.

7. Discussion

We proposed an estimation algorithm for the recent history functional linear models which are useful in applications. The sliding window support of the recent history functional linear model strikes a useful balance between the global support of the functional linear models and the point wise support of the varying coefficient models. The assumption that only the predictor process from the recent past has an effect on the response rather than the future predictor values or only the current predictor value, is useful in many applications where changes in the response process can be explained using recent trends of the predictor process. In addition the product form assumed for the regression surface uses only one dimensional smooth functions considerably easing and speeding estimation.

Our proposal is geared towards sparse longitudinal data where the estimation procedure proposed also accommodates measurement error in variables. Sparsity and measurement error are both commonly encountered in longitudinal designs. We provide asymptotic properties of our estimators that enable the estimation of the predicted response trajectories and that lead to asymptotic confidence bands. Choice of model parameters is also addressed, where favorable properties of the proposed estimators are demonstrated in simulations and data applications.

Acknowledgments

We are extremely grateful to an anonymous referee and the Editor for helpful remarks that improved the paper. Kim's research is supported by NIDA, NIH grants R21 DA024260 as a research assistant. Li's research is supported by NIDA, NIH grants P50 DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

References

Ash, RB.; Gardner, MF. Topics in Stochastic Processes. Academic Press; New York: 1975.

Cardot H, Ferraty F, Sarda P. Spline estimators for the functional linear model. Statistica Sinica. 2003; 13:571–592.

Chiang CT, Rice JA, Wu CO. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. Journal of the American Statistical Association. 2001; 96:605–619.

Cleveland WS, Grosse E, Shyu WM. Local regression models. Statistical Models in S. 1991:309–376.

Fan, J.; Gijbels, I. Local Polynomial Modeling and Its Applications. Chapman and Hall; London: 1996.

Fan J, Zhang W. Two-step estimation of functional linear models with applications to longitudinal data. Journal of the Royal Statistical Society Series B (Methodological). 2000; 62:303–322.

Fan J, Zhang W. Statistical methods with varying coefficient models. Statistics and Its Interface. 2008; 1:179–195. [PubMed: 18978950]

Fleming, T.; Harrington, D. Counting Processes and Survival Analysis. Wiley; New York: 1991.

Hastie T, Tibshirani R. Varying coefficient models. Journal of the Royal Statistical Society Series B (Methodological). 1993; 55:757–796.

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

Huang J, Shen H. Functional coefficient regression models for nonlinear time series: a polynomial spline approach. Scandinavian Journal of Statistics. 2004; 31:515–534.

Huang JZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika. 2002; 89:111–128.

Huang JZ, Wu CO, Zhou L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statistica Sinica. 2004; 14:763–788.

James G, Hastie TJ, Sugar CA. Principal component models for sparse functional data. Biometrika. 2000; 87:587–602.

Kauermann G, Tutz G. On model diagnostics using varying coefficient models. Biometrika. 1999; 86:119–128.

Kim, K.; Şentürk, D.; Li, R. Technical report. The Pennsylvania State University; 2009. The recent history functional linear models.

Liu, B.; Müller, HG. Functional data analysis for sparse auction data. Wiley and Sons Inc; 2008. p. 269-290.

Malfait N, Ramsay JO. The historical functional linear model. Canadian Journal of Statistics. 2003; 31:115–128.

Müller HG. Functional modelling and classification of longitudinal data. Scandinavian Journal of Statistics. 2005; 32:223–240.

Murtaugh P, Dickson E, Van Dam G, Malinchoc M, Grambsch P, Langworthy A, Gips C. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. Hepatology. 1994; 20:126–134. [PubMed: 8020881]

Ramsay, J.; Silverman, B. Functional Data Analysis. Springer-Verlag; New York: 2002.

Ramsay, J.; Silverman, B. Functional Data Analysis. Springer-Verlag; New York: 2005.

Rice J. Functional and longitudinal data analysis: Perspectives on smoothing. Statistica Sinica. 2004; 14:631–647.

Rice J, Silverman B. Estimating the mean and covariance structure nonparametrically when data are curves. Journal of the Royal Statistical Society Series B (Methodological). 1991; 53:233–243.

Şentürk D, Müller HG. Generalized varying coefficient models for longitudinal data. Biometrika. 2008; 95:653–666.

Şentürk D, Müller HG. Functional varying coefficient models for longitudinal data. Journal of the American Statistical Association. 2010 in-press.

Shi M, Weiss RE, Taylor JMG. An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. Applied Statistics, Journal of the Royal Statistical Society Series C. 1996; 45:151–163.

Wu CO, Chiang CT, Hoover D. Asymptotic confidence regions for kernel smoothing of a varying coefficient model with longitudinal data. Journal of the American Statistical Association. 1998; 93:1388–1389.

Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association. 2005a; 100:577–591.

Yao F, Müller HG, Wang JL. Functional linear regression analysis for longitudinal data. Annals of Statistics. 2005b; 33:2873–2903.

Appendix A. Assumptions

We present the assumptions in three groups, assumptions (A) are needed for all three Theorems, assumptions (B) are needed for the consistency and asymptotic normality of the predicted response trajectories given in Theorems 2 and 3 respectively, and assumption (C1) is only used in Theorem 3.

The data (T_{ij}, X_{ij}) and (T_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, are assumed to be the i.i.d. samples from the joint densities, $g_1(t, x)$ and $g_2(t, y)$. Assume also that the observation times T_{ij} are i.i.d. with marginal densities $f_T(t)$. Let T_1 and T_2 be two different time points, and X_1 (respectively Y_1) and X_2 (respectively Y_2) be the repeated measurements of X (respectively Y) made on the same subject at times T_1 and T_2. The predictor (and response) measurements made on the same subject at different times are allowed to be dependent. Assume $(T_{ij}, T_{il}, X_{ij}, X_{il})$, $1 \le j \ne l \le N_i$, is identically distributed as (T_1, T_2, X_1, X_2) with joint density function $g_{XX}(t_1, t_2, x_1, x_2)$ and analogously for $(T_{ij}, T_{il}, Y_{ij}, Y_{il})$ with identical joint density function $g_{YY}(t_1, t_2, y_1, y_2)$. The following regularity conditions are assumed on $f_T(t)$, $g_1(t, x)$, $g_2(t, y)$, $g_{XX}(t_1, t_2, x_1, x_2)$ and $g_{YY}(t_1, t_2, y_1, y_2)$. Let p_1, p_2 be integers with $0 \le p_1 + p_2 \le 4$.

(A1) The derivative $(d^p/dt^p)\, f_T(t)$ exists and is continuous on $[0, T]$ with $f_T(t) > 0$ on $[0, T]$, $(d^p/dt^p)g_1(t, x)$ and $(d^p/dt^p)g_2(t, y)$ exist and are continuous on $[0, T] \times \mathbb{R}$, and $\{d^p/(dt^{p1}dt^{p2})\}g_{XX}(t_1, t_2, x_1, x_2)$ and $\{d^p/(dt^{p1}dt^{p2})\}g_{YY}(t_1, t_2, x_1, x_2)$ exist and are continuous on $[0, T]^2 \times \mathbb{R}^2$ for $p_1 + p_2 = p$, $0 \le p_1, p_2 \le p$.

(A2) The number of measurements N_i made for the i^{th} subject is a random variable such that $N_i \overset{i.i.d}{\sim} N$, where N is a positive discrete random variable with $P(N > 1) > 0$. The observation times and measurements are assumed to be independent of the number of observations for any subset $J_i \in \{1,\ldots, N_i\}$ and for all $i = 1, \ldots, n$, i.e. $\{T_{ij}, X_{ij}, Y_{ij}: j \in J_i\}$ is independent of N_i.

Let $K_1(\cdot)$ and $K_2(\cdot, \cdot)$ be the nonnegative univariate and bivariate kernel functions for smoothing the mean functions μ_X and μ_Y, and auto-covariance surface, G_X, and cross-covariance surface, G_{XY}. Assume that K_1 and K_2 are densities with zero means and finite variances on a compact support.

(A3) The Fourier transformations of $K_1(u)$ and $K_2(u, v)$, defined by $\kappa_1(t) = \int e^{-iut}K_1(u)du$ and $\kappa_2(t, s) = \int e^{-(iut+ivs)}K_2(u, v)dudv$ are required to absolutely integrable, i.e. $\int|\kappa_1(t)|dt < \infty$ and $\int\int|\kappa_2(t, s)|dtds < \infty$.

Let h_X and h_Y be the bandwidths used for estimating μ_X and μ_Y, respectively. Also let h_G be the bandwidth used for estimating G_X, and let (h_1, h_2) be bandwidths used in the estimation of G_{XY}.

(A4) As $n \to \infty$, the following are assumed about the bandwidths.

(A4.1) $h_X \to 0$, $h_Y \to 0$, $nh_X^4 \to \infty, nh_Y^4 \to \infty$, and $nh_X^6<\infty, nh_Y^6<\infty$.

(A4.2) $h_G \to 0$, $nh_G^6 \to \infty$, and $nh_G^8<\infty$.

(A4.3) Without loss of generality, $h_1/h_2 \to 1$, and $nh_1^6 \to \infty, nh_1^8<\infty$.

(A5) Assume that the fourth moments of Y and X are finite.

(B1) The number of eigenfunctions used in (11), $M = M(n)$, is an integer valued sequence that depends on sample size n and satisfies the rate conditions given in assumption (B5) of Yao et al. (2005a).

(B2) The number and locations of measurements for a given subject does not change as the sample size $n \to \infty$.

(B3) For all $1 \le i \le n$, $m \ge 1$ and $1 \le \ell \le N_i$, the functional principal component scores ζ_{im} and the measurement errors $\varepsilon_{i\ell}$ in (4) are jointly Gaussian.

(C1) There exists a continuous positive definite function $\omega(s, t)$ such that $\omega_M(s, t) \to \omega(s, t)$, as $M \to \infty$.

Appendix B. Proofs

Proof of Theorem 1

Uniform consistency of $\hat{G}_X(s, t)$ is given in Theorem 1 of Yao et al. (2005b) and that of $\hat{G}_{XY}(s, t)$ is given in Lemma A.1 of Yao et al. (2005a). Consistency of $\hat{v}_{kl}(t)$ and $\hat{\mu}_k(t)$ for v_{kl} and $\theta_k(t)$ follow from uniform consistency of $\hat{G}_X(s, t)$ and $\hat{G}_{XY}(s, t)$. This implies consistency of $\hat{b}(t)$ for $b(t)$, and hence that of $\hat{\beta}(s, t)$.

Proof of Theorem 2

For fixed M, define $\tilde{Y}^*_M(t)=\mu_Y(t)+\sum_{m=1}^{M}\tilde{\zeta}^*_m P_m(t)$, where $\hat{\zeta}_m$ is as defined in (10) and $P_m(t) = \int_{\Delta_t}\beta(s,t)\psi_m(s)ds$. Then, it follows that

$$|\widehat{Y}^*_M(t) - \tilde{Y}^*(t)| \le |\widehat{Y}^*_M(t) - \tilde{Y}^*_M(t)|+|\tilde{Y}^*_M(t) - \tilde{Y}^*(t)|=Q_1+Q_2.$$

The convergence of Q_2 to 0 as $n \to \infty$ follows from Lemma A.3 in Yao et al. (2005a). Note that, for Q_1,

$$Q_1=|\widehat{Y}^*_M(t) - \tilde{Y}^*_m(t)| \le |\widehat{\mu}_Y(t) - \mu_Y(t)|+\sum_{m=1}^{M}|\widehat{\zeta}^*_m\widehat{P}_m - \widehat{\zeta}^*_m P_m(t)|.$$

Uniform consistency of $\hat{\mu}_Y(t)$ for $\mu_Y(t)$ follows from Theorem 1 in Yao et al. (2005b), and consistency of $\widehat{\zeta}^*_m$ for $\tilde{\zeta}^*_m$ follows from Theorem 3 in Yao et al. (2005b). From uniform consistency of $\hat{\beta}(s,t)$ established in Theorem 1 of Section 3.1 and that of $\hat{\psi}_k(t)$ shown in Yao et al. (2005a), uniform consistency of $\hat{P}_m(t)$ follows. Combining these results, we have

$$\sup_{t\in[\delta_1,T]}|\widehat{Y}^*_M(t) - \tilde{Y}^*_M(t)| \xrightarrow{p} 0 \text{ as } n \to \infty,$$

and by Slutsky's Theorem, Theorem 2 follows.

Proof of Theorem 3

Under the Gaussian assumption, for any fixed $M \ge 1$, we have $\tilde{\zeta}^*_m - \zeta^*_m \sim \mathcal{N}(0,\Omega_M)$. In proof of Theorem 2, it is shown that $\lim_{n\to\infty}\sup_{t\in\mathcal{T}}|\widehat{Y}^*_M(t) - \tilde{Y}^*_M(t)| \xrightarrow{p} 0$. Observing that $\widehat{Y}^*_M(t) - Y^*_M(t)=\widehat{Y}^*_M(t) - \tilde{Y}^*_M(t)+\tilde{Y}^*_M(t) - Y^*_M(t)$, we have $\{\widehat{Y}^*_M(t) - Y^*_M(t)\} \xrightarrow{\mathcal{D}} Z_m \overset{\mathcal{D}}{=} \mathcal{N}(0,\omega_M(t,t))$. Under assumption (C1), letting $M \to \infty$ leads to $Z_M \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,\omega(t,t))$. From the Karhunen-Loéve Theorem, $|Y^*_M(t) - Y^*(t)| \xrightarrow{p} 0$. Therefore, $\lim_{M\to\infty}\lim_{n\to\infty}|\widehat{Y}^*_M(t) - Y^*(t)| \overset{\mathcal{D}}{=} Z$. From the convergence of $\hat{\psi}(t)$, $\widehat{\zeta}^*_m$, $\hat{\gamma}_m$ and $\hat{P}_m$ for $\psi(t)$, ζ^*_m, γ_m and $\tilde{P}_m$, we can deduce $\widehat{\omega}_M(t,t) \xrightarrow{p} \omega_M(t,t)$ as $n \to \infty$. Under the assumption (C1), it follows that $\lim_{M\to\infty}\lim_{n\to\infty}\hat{\omega}_M(t,t) = \omega(t,t)$ in probability. Applying Slutsky's theorem, (12) follows.

Appendix C. Estimation procedures

In this section, we provide explicit forms for the local polynomial smoothing procedures used in estimating the mean functions and covariance surfaces. Eigendecompositions for the estimated covariance surfaces and explicit form of the measurement error variance estimator are also provided.

The estimator of mean function for the predictor process, $\hat{\mu}_X(t)$, can be obtained by local linear regression via minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i} K_1\left(\frac{T_{ij}-t}{h_X}\right)\{X_{ij}-\eta_0-\eta_1(t-T_{ij})\}^2,$$

with respect to η_0, η_1, which leads to $\hat{\mu}_X(t) = \hat{\eta}_0$. Estimation of $\mu_Y(t)$ follows similarly.

For estimation of the cross-covariance surface G_{XY}, the two dimensional local linear smoother is fitted to the raw covariances by minimizing

$$\sum_{i=1}^{n}\sum_{1\le j,\ell\le N_i} K_2\left(\frac{T_{ij}-s}{h_1},\frac{T_{i\ell}-t}{h_2}\right)[G_{XY,i}(T_{ij},T_{i\ell})-f\{\eta,(s,t),(T_{ij},T_{i\ell})\}], \tag{C.1}$$

with respect to $\eta = (\eta_0, \eta_1, \eta_2)$, yielding $\hat{G}_{XY}(s, t) = \hat{\eta}_0$. Estimation of the auto-covariance surface of the predictor process can be obtained similarly where the diagonal elements of the raw auto covariance matrix are not included in the smoothing as described in Section 3.1. Hence the second sum in (C.1) is taken over $1 \le j \ne \ell \le N_i$, excluding the diagonal terms.

In order to obtain the estimator for the measurement error variance σ_ε^2, we first estimate the diagonal elements of the auto-covariance surface $G_X(t, t)$ excluding the diagonal raw covariances contaminated with error, by applying a local linear smoother along the diagonal and local quadratic smoother along the direction perpendicular to the diagonal. The resulting estimators of the diagonal elements are denoted by $\tilde{G}_X(t)$. This estimator is then compared to a linear smoother fit only to the diagonal raw covariance terms $\{T_{ij}, G_{X,i}(T_{ij}, T_{ij})\}$, estimating $G_X(t, t)+\sigma_\varepsilon^2(t)$. This estimator is denoted by $\hat{V}(t)$. We estimate the error variance by these two estimators, yielding

$$\widehat{\sigma}_\varepsilon^2=\frac{2}{T}\int_{\mathcal{T}_1}\{\widehat{V}(t)-\tilde{G}(t)\}dt,$$

where $\mathcal{T}_1 = [T/4, 3T/4]$. Here, the integration is taken over the middle half of $\mathcal{T}$ in order to remove the boundary effect of the local polynomial smoother.

The eigenfunctions and eigenvalues of the estimated auto-covariance surface, $\hat{G}_X(s, t)$, for the predictor process are the solutions, $\hat{\psi}_k$ and $\hat{\gamma}_k$, of the eigenequation given by

$$\int_{[0,T]}\widehat{G}_X(s,t)\widehat{\psi}_k(s)ds=\widehat{\gamma}_k\widehat{\psi}_k(t),$$

where $\int_{[0,T]}\widehat{\psi}_k^2(t)dt=1$ and $\int_{[0,T]}\hat{\psi}_m(t)\hat{\psi}_k(t)dt = 0$ for $m \ne k$. For numerical solutions, discretization of the smoothed covariance function can be used following Rice and Silverman (1991).

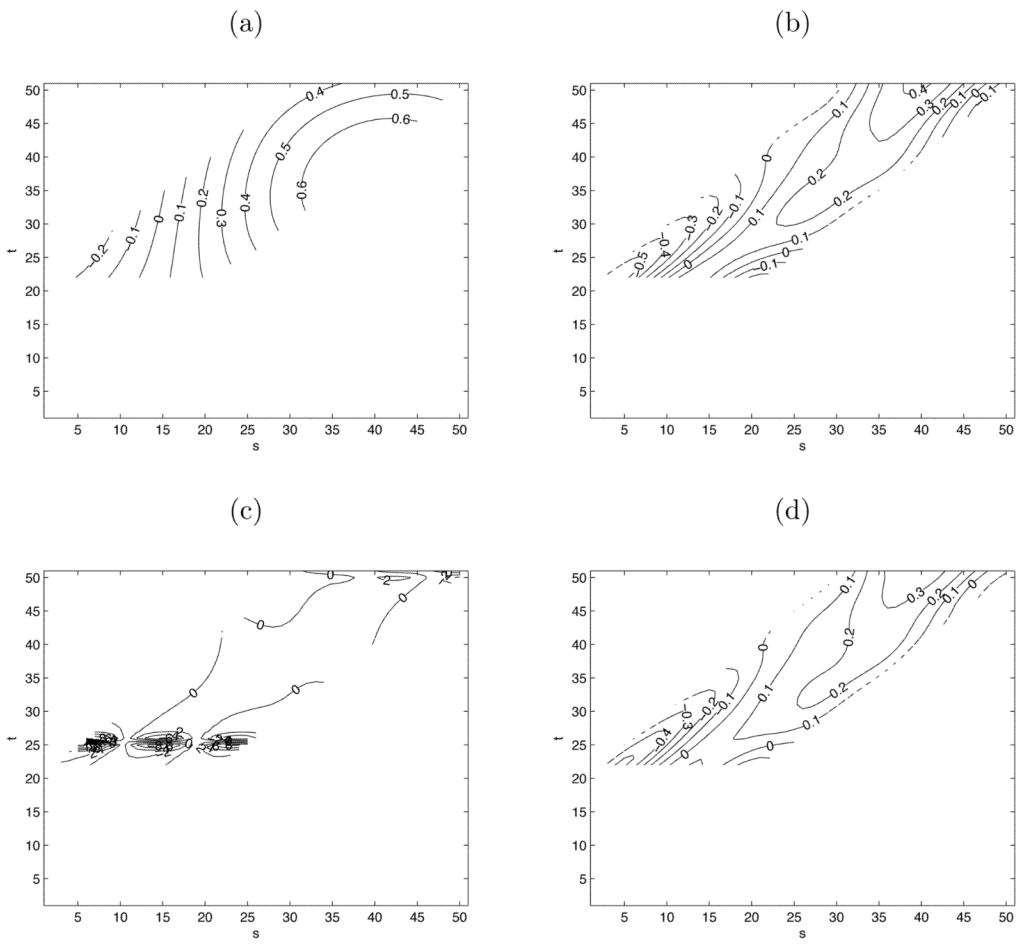

Figure 1.

(a) The true regression function defined on $\Delta = [0, 20]$. The estimated regression function with (K, λ) equal to (5, 7.4) (plot b, estimated IRSD= 0.3632), (5, 1) (plot c, estimated IRSD= 298) and (10, 7.4) (plot d, estimated IRSD= 0.3274).

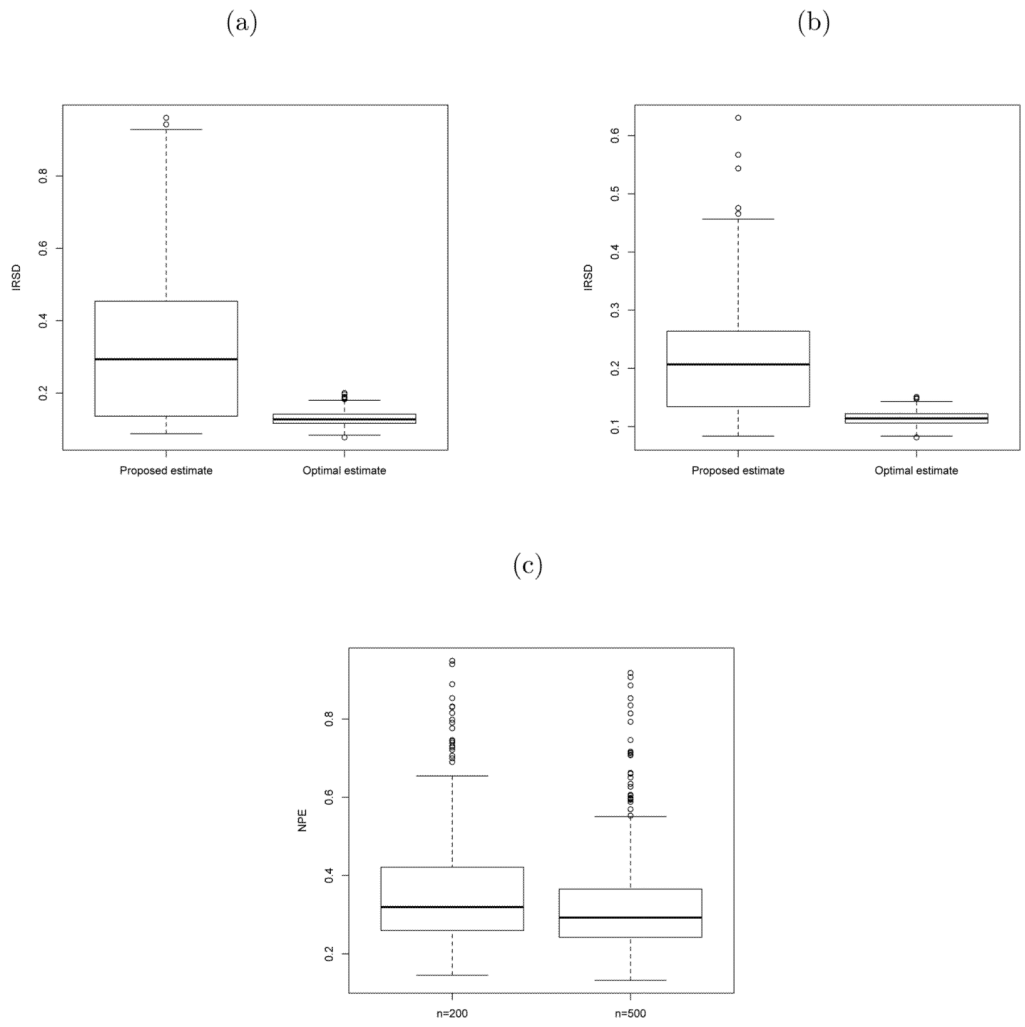

Figure 2.
Boxplots of the estimated IRSD values of the proposed and optimal estimators for $n = 200$ (a) and $n = 500$ (b). (c) Boxplot of NPE values of the proposed estimators for $n = 200$ and $n = 500$.

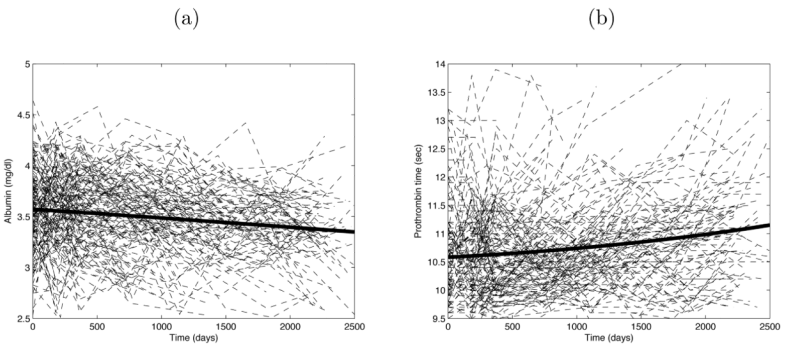

Figure 3.
(a) Individual predictor trajectories (dashed) overlaying the estimated cross-sectional mean of the predictor process (solid). (b) Individual response trajectories (dashed) overlaying the cross-sectional mean of the response process (solid).

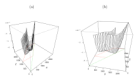

Figure 4.
The estimated regression surface defined on $[t - 1000, t] \times [1000, 2500]$ in the longitudinal primary biliary liver cirrhosis data obtained by $\lambda = 6502$ and $K = 10$.

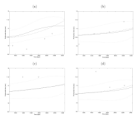

Figure 5.
Predicted trajectories for 4 randomly selected subjects. In each plot, circles denote the original response observations, the solid line is the response trajectory predicted via the proposed method, the dashed line is the predicted trajectory using the functional linear fit of Yao et al. (2005a), and the dotted lines represent the 95% asymptotic confidence bands.