Research Article

Statistics in Medicine

Received 11 March 2012, Accepted 20 December 2012 Published online 18 January 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5740

# Modeling time-varying effects with generalized and unsynchronized longitudinal data

**Damla Şentürk,[a*†] Lorien S. Dalrymple,[b] Sandra M. Mohammed,[c] George A. Kaysen[b,d] and Danh V. Nguyen[c]**

**We propose novel estimation approaches for generalized varying coefficient models that are tailored for unsynchronized, irregular and infrequent longitudinal designs/data. Unsynchronized longitudinal data refer to the time-dependent response and covariate measurements for each individual measured at distinct time points. Data from the Comprehensive Dialysis Study motivate the proposed methods. We model the potential age-varying association between infection-related hospitalization status and the inflammatory marker, C-reactive protein, within the first 2 years from initiation of dialysis. We cannot directly apply traditional longitudinal modeling to unsynchronized data, and no method exists to estimate time-varying or age-varying effects for generalized outcomes (e.g., binary or count data) to date. In addition, through the analysis of the Comprehensive Dialysis Study data and simulation studies, we show that preprocessing steps, such as binning, needed to synchronize data to apply traditional modeling can lead to significant loss of information in this context. In contrast, the proposed approaches discard no observation; they exploit the fact that although there is little information in a single subject trajectory because of irregularity and infrequency, the moments of the underlying processes can be accurately and efficiently recovered by pooling information from all subjects using functional data analysis. We derive subject-specific mean response trajectory predictions and study finite sample properties of the estimators.**


**Keywords:** binning; functional data analysis; generalized linear models; sparse design; United States Renal Data System; varying coefficient models

## 1. Introduction

The public health burden directly related to infection in the dialysis population is substantial. Current projections estimate that the USA will have as many as 710,000 prevalent end-stage renal disease patients by the year 2015 [1]. Recently, Dalrymple *et al.* [2] showed a significant burden of infection in patients on dialysis, finding that among patients aged 65 to 100 years, the rate of infection-related hospitalization was 52 and 53 per 100 person-years for patients on peritoneal and hemodialysis, respectively. The Comprehensive Dialysis Study (CDS) is a prospective cohort study of end-stage renal disease patients who newly initiated dialysis between September 2005 and June 2007 [3] where longitudinal serum samples were collected on a subset of participants within the first 2 years from the start of dialysis. In this work, we aim to examine the potential age-varying association between infection-related hospitalization status and serum C-reactive protein (CRP), a positive acute-phase protein marker of inflammation that is increased during inflammation.

Varying coefficient models [4, 5] are designed to capture complex age-varying (or, more generally, time-varying) effects/associations in regression relationships. They have been widely used in

[a]*Department of Biostatistics, University of California, Los Angeles, CA, U.S.A.*
[b]*Division of Nephrology, Department of Medicine, University of California, Davis, CA, U.S.A.*
[c]*Division of Biostatistics, University of California, Davis, CA, U.S.A.*
[d]*Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, U.S.A.*
**Correspondence to: Damla Şentürk, Department of Biostatistics, University of California, Los Angeles, CA 90095, U.S.A.*
[†]*E-mail: dsenturk@ucla.edu*

longitudinal data analysis in the past decade because of their ability to capture time-varying associations and their ease in interpretation and natural graphical display of time-varying dynamics [6–12]. Cai *et al.* [13] developed extensions to generalized varying coefficient models for modeling longitudinal generalized responses, such as binary variables (e.g., infection-related hospitalization status) or counts of events. The generalized varying coefficient model for longitudinal data is

$$E\{Y(t)|X(t)\} = g\{\beta_0(t) + \beta_1(t)X(t)\}, \tag{1}$$

where $g(\cdot)$ is a known inverse link (transformation) function that relates the mean outcome to the longitudinal covariate $X$. Qu and Li [14] studied estimation in generalized varying coefficient models for longitudinal data using penalized spline expansions coupled with quadratic inference function approaches, whereas Zhang [15] used a generalized linear mixed-model approach where a double-penalized quasi-likelihood approach is used for estimation.

There are several major challenges in the estimation of the generalized varying coefficient models from the CDS data. First, the longitudinal measurements on the binary response, the infection-related hospitalization status and the continuous covariate, serum CRP concentration, are obtained at distinct time points (unsynchronized) within each subject. Hospitalization times are stochastic and do not coincide with serum CRP measurement times. In addition, longitudinal hospitalization data are naturally highly irregular and infrequent over time. Thus, longitudinal data characterized by a combination of unsynchronized, irregular and infrequent measurements pose substantial challenges to modeling time-varying effects. We note that in other longitudinal studies, these issues may arise because of missed and/or rescheduled visits, despite diligent plans to collect data contemporaneously and on regular follow-up schedules.

Existing methods for the regression modeling of longitudinal data cannot handle unsynchronized data. Xiong and Dubin [16] have proposed binning, as a data preprocessing step (if feasible), to synchronize the response and covariate measurements to make existing methods applicable. However, binning can lead to significant loss of data in irregular and infrequent longitudinal designs as will be shown in the simulation studies of Section 5. For the CDS data, binning leads to a loss of 69% of the repeated measurements (and 51% loss in subject sample size). Hence, in this work, we develop novel estimation procedures for generalized varying coefficient models based on unsynchronized, irregular and infrequent longitudinal data, obviating data loss due to preprocessing steps.

The proposed estimation procedures build on recent developments in the longitudinal data literature that introduce functional data analysis techniques [17–21] to address irregular and infrequent designs [22–24]. Yao *et al.* [25,26] used estimates of the covariance structure and mean function of the longitudinal trajectories for functional linear regression; Hall *et al.* [27] modeled a single generalized longitudinal response trajectory (without any covariates) using latent Gaussian processes based on functional data analysis. Recently, Senturk and Mueller [28], Senturk and Nguyen [29] and Kim *et al.* [30] developed a functional data analysis framework for estimation in the standard varying coefficient models and reported improved estimation results over standard estimation procedures such as local least squares in regression modeling of sparse continuous error-prone longitudinal data. However, there has been no work on modeling generalized longitudinal outcome, including binary and count data, in the framework of generalized varying coefficient modeling geared towards unsynchronized, irregular and infrequent designs.

The remainder of the paper is organized as follows. We detail the proposed estimation approaches in Section 2, where binning coupled with local maximum likelihood estimation is also outlined as a comparison/baseline approach in Section 2.4. We derive representations of the varying coefficient functions of interest via the moments of the underlying covariate and response processes in Section 2.1, leading to our first direct approach to estimation of the varying coefficient functions. We outline in Section 2.3 a second alternative approach based on reconstruction of the predictor processes at the observation time points for the response. Both of the proposed approaches rest on ideas of borrowing strength from the entire data and dimension reduction. Hence, unlike binning, every observation contributes to the estimation of the varying coefficient functions, and no observation is discarded. Furthermore, although standard approaches can predict the mean response only at the original sparse observation times, the proposed methods lead to subject-specific predictions of the mean response trajectories for the entire study period as outlined in Section 3. We illustrate the proposed method with the aforementioned CDS data in Section 4. Section 5 reports on simulation studies of the accuracy of the proposed estimation and prediction procedures, including comparisons with local maximum likelihood estimators coupled with binning. We conclude with a brief discussion in Section 6.

## 2. Estimation in generalized varying coefficient models

Consider the generalized response $Y_i(t)$ and the longitudinal predictor $X_i(t)$ for $i = 1, \ldots, n$ subjects in model (1). We assume the observed covariate and response trajectories to be square integrable realizations of the random smooth processes $X$ and $Y$. The unsynchronized, infrequent and irregular nature of the CDS data, as described in the previous section, is characterized by subject-specific and variable-specific random observation times and a small total number of repeated measurements. To accommodate these data characteristics, we assume that the longitudinal response $(Y_i)$ and the covariate $(X_i)$ trajectories for subject $i$ are observed at *distinct* time points $T_{ij} \in [0, T]$ and $S_{ik} \in [0, T]$, for $j = 1, \ldots, N_i$ and $k = 1, \ldots, M_i$, respectively. $N_i$ and $M_i$ denote the $i$th subject's total number of repeated measurements for the response and covariate, respectively. We also assume additive measurement error on the longitudinal covariate, that is, $X_{ik} = X_i(S_{ik}) + \epsilon_{ik}$, where $\epsilon_{ik}$ are mean zero finite variance independent and identically distributed measurement errors. $Y_{ij} = Y_i(T_{ij})$ denotes the repeated measurements on the generalized response.

### 2.1. *Moment representations of the varying coefficient functions*

In this section, we outline our first approach for estimation in generalized varying coefficient models based on the moment representations for the time-varying coefficient functions of interest. We consider an expansion of $X_i$ about its mean function, where the variation about its mean is relatively small, that is, $X_i(t) = \mu_X(t) + \delta Z_i(t)$, with $\mu_X(t) \equiv E\{X(t)\}$, where $Z_i$ is a mean zero, bounded variance stochastic process and $\delta > 0$ is an unknown small constant. Assuming that the function $g$ is continuously differentiable, $g'$ does not vanish and $\inf_{s \in D} g'(s) > 0$, where $D$ is the range of $\beta_0(t) + \beta_1(t)\mu_X(t)$; it holds by Taylor's expansion that

$$g\{\beta_0(t) + \beta_1(t)X(t)\} = g\{\beta_0(t) + \beta_1(t)\mu_X(t)\} + \delta Z(t)\beta_1(t)g'\{\beta_0(t) + \beta_1(t)\mu_X(t)\} + O_p(\delta^2).$$

It follows that $\mu_Y(t) \equiv E\{Y(t)\} = g\{\beta_0(t) + \beta_1(t)\mu_X(t)\} + O(\delta^2)$ because $E\{Z(t)\} = 0$ and

$$\begin{aligned} G_{YX}(t,t) \equiv \operatorname{cov}\{Y(t), X(t)\} &= \operatorname{cov}[g\{\beta_0(t) + \beta_1(t)X(t)\}, X(t)] \\ &= \beta_1(t)g'\{\beta_0(t) + \beta_1(t)\mu_X(t)\}\operatorname{cov}\{X(t), X(t)\} + O(\delta^3) \\ &= \beta_1(t)g'\{\beta_0(t) + \beta_1(t)\mu_X(t)\}G_{XX}(t,t) + O(\delta^3), \end{aligned}$$

where $G_{XX}(t,t) \equiv \operatorname{cov}\{X(t), X(t)\}$. Hence, we obtain the following approximations to the time-varying coefficients of interest,

$$\beta_1(t) \approx \frac{G_{YX}(t,t)}{g'[g^{-1}\{\mu_Y(t)\}]G_{XX}(t,t)} \quad \text{and} \quad \beta_0(t) \approx g^{-1}\{\mu_Y(t)\} - \beta_1(t)\mu_X(t). \tag{2}$$

The approximations in (2) imply plug-in estimators for the targeted varying coefficient functions $\beta_1(t)$ and $\beta_0(t)$, respectively, both up to terms of order $O(\delta^2)$. Note that the derived expressions in (2) do not depend on $\delta$; therefore, we need not estimate $\delta$. Nevertheless, we study the sensitivity of the proposed estimators to different values of $\delta$ (i.e., different variance of $X$) via Monte Carlo simulations in Section 5.

Note that the derived expressions for the varying coefficient functions given in (2) depend only on population quantities. Hence, even though the data are unsynchronized and infrequent at the subject level, we can still effectively estimate population moments when the data from all subjects are pooled together as will be described in the next section.

### 2.2. *Estimation of the moments of the underlying stochastic processes*

We obtain the population moments through smoothing, which also allows for pooling of information from all subjects. In a first step, we obtain the mean functions of the longitudinal trajectories by smoothing the aggregated data $(S_{ik}, X_{ik})$ and $(T_{ij}, Y_{ij})$ for $i = 1, \ldots, n$, $k = 1, \ldots, M_i$ and $j = 1, \ldots, N_i$, with local linear fitting. This yields $\hat{\mu}_X(t)$ and $\hat{\mu}_Y(t)$, respectively. Next, we compute the raw covariances between $(Y, X)$ and $(X, X)$ as $G_{YX,i}(T_{ij}, S_{ik}) = \{Y_{ij} - \hat{\mu}_Y(T_{ij})\}\{X_{ik} - \hat{\mu}_X(S_{ik})\}$ and $G_{XX,i}(S_{ik}, S_{i\ell}) = \{X_{ik} - \hat{\mu}_X(S_{ik})\}\{X_{i\ell} - \hat{\mu}_X(S_{i\ell})\}$, respectively. To obtain the final smooth

estimates of the covariances, $\widehat{G}_{YX}$ and $\widehat{G}_{XX}$, we feed the raw estimates, $G_{YX,i}$ and $G_{XX,i}$, into a two-dimensional local least squares algorithm. Senturk and Mueller [28] give explicit expressions of the local least squares estimators. For a computationally efficient bandwidth choice in the proposed one-dimensional and two-dimensional smoothing, we adopt the generalized cross-validation algorithm of Liu and Mueller [31].

To eliminate the effects of covariate measurement error on the autocovariance $\widehat{G}_{XX}$, we exclude the diagonal raw covariance elements $G_{XX,i}(S_{ik}, S_{ik})$, $i = 1, \ldots, n$ and $k = 1, \ldots, M_i$, in the two-dimensional smoothing step. This is because the measurement error on the longitudinal predictor variables only affects the variance terms along the diagonal. We can find more details on this phenomenon in [25, 26]. To guarantee the nonnegative definiteness of the estimated autocovariance matrix, we exclude the negative estimates of the eigenvalues and corresponding eigenfunctions from the functional principal component decomposition $G_{XX}(s,t) = \sum_{\ell=1}^{\infty} \rho_\ell \phi_\ell(s)\phi_\ell(t)$, where $\phi_\ell$ denotes the eigenfunctions with nonincreasing eigenvalues $\rho_\ell$. Here, we would employ a nonparametric functional principal component analysis step on the smooth estimate of the autocovariance surface by a standard discretization procedure to estimate the eigenfunctions and eigenvalues. Once these quantities are estimated, we give the final autocovariance estimator as $\widehat{G}_{XX}(s,t) = \sum_{\ell=1,\hat{\rho}_\ell>0}^{L} \hat{\rho}_\ell \hat{\phi}_\ell(s)\hat{\phi}_\ell(t)$, where the total number $L$ of eigen-components included can be chosen by various criteria, including the AIC or fraction of variance explained. (For more details, see Appendix A.3 of Senturk and Mueller [28].)

We can plug the estimated mean and covariance functions into the varying coefficient function representations given in (2) to obtain

$$\hat{\beta}_1(t) = \frac{\widehat{G}_{YX}(t,t)}{g'[g^{-1}\{\hat{\mu}_Y(t)\}]\widehat{G}_{XX}(t,t)} \quad \text{and} \quad \hat{\beta}_0(t) = g^{-1}\{\hat{\mu}_Y(t)\} - \hat{\beta}_1(t)\hat{\mu}_X(t). \tag{3}$$

We will refer to the preceding estimators as the moment estimators of the varying coefficient functions throughout the manuscript. The uniform consistency of the moment estimators, up to terms of order $O(\delta^2)$ (Section 2.1), follows from the uniform consistency of the moment estimators described earlier. More specifically, Yao *et al.* [25, 26] and Hall *et al.* [27] have shown the uniform consistency of the proposed mean and covariance function estimates for continuous longitudinal observations and for generalized longitudinal observations, respectively. We study the sensitivity of the finite sample performance of the moment estimators to different $\delta$ values in the simulation studies of Section 5.

Note that instead of applying a binning procedure to each subject trajectory to synchronize the response and covariate measurements, the proposed method uses smoothing in the estimation of the population quantities. Although regularization via smoothing of each subject's trajectory may be appropriate for densely measured longitudinal data, the proposed approach with regularization applied at the estimation of the population moments is much more suitable for infrequent and unsynchronized designs. In this way, every observation on each subject, whether it be unsynchronized or infrequent, contributes to the estimation via the connection between the population moments and the varying coefficient functions given in (2).

For the analysis with a fixed duration, when another time index is considered for the varying coefficient model, other than the duration of the study, such as age, we have subject-specific supports for the observed data: $\{a_i; X_i(S_{ik}); Y_i(T_{ij})\}$ for $j = 1, \ldots, N_i$ and $k = 1, \ldots, M_i$, $S_{ik}, T_{ij} \in [a_i + T_0, a_i + T_1]$ and $T_0 < T_1$. We will refer to this case as the fixed duration case throughout the manuscript. As we will describe in more details in the data analysis of Section 4, for the CDS data, the protein inflammation marker and hospitalization measurement times per subject are randomly scattered within the [100, 550] day interval of interest, after the initiation of dialysis. Hence, in this setup, $a_i$ refers to age at initiation of dialysis for subject $i$, $T_0 = 100$ and $T_1 = 550$. Because each subject is observed within 100 to 550 days from their baseline age (age at initiation of dialysis, marking the beginning of the study), the raw covariances are only available in a band around the diagonal of length twice the duration of the study (450 days). (Section 4 provides details for this.) Hence, we need to perform the smoothing only in this region around the diagonal. In addition, note that (2) involves only the diagonal values of the autocovariance and cross-covariance surfaces of the underlying covariate and response processes. Hence, we can estimate the proposed age-varying coefficient models on the basis of an analysis with a fixed duration time, such as the analysis of the CDS data considered here. We study in detail the properties of the proposed estimation procedures under both time indices, the duration of the study and subject age (similar to CDS data), in simulation studies of Section 5.

### 2.3. *Reconstruction of the predictor processes and local maximum likelihood estimation*

Because traditional estimation procedures for modeling longitudinal data can only handle synchronized data, our second proposed estimation approach uses the functional data analysis framework and the estimated population moments in Section 2.2 to reconstruct the predictor processes at the observation times of the response. Once the predictor measurements are obtained, synchronized with the response for each subject, we utilize an extension of the local maximum likelihood estimators of Cai *et al.* [13] originally proposed for independent and identically distributed data to longitudinal data to obtain estimators of the varying coefficient functions.

The starting point in reconstructing the predictor trajectories will be the Karhunen–Loéve expansion for the observed process for subject $i$,

$$X_{ik} = \mu_X(S_{ik}) + \sum_{\ell=1}^{\infty} \xi_{i\ell}\phi_\ell(S_{ik}) + \epsilon_{ik},$$

where $\xi_{i\ell}$ is the $\ell$th functional principal component score playing the role of random effects with $E(\xi_\ell) = 0$ and $\text{var}(\xi_\ell) = \rho_\ell$, $\epsilon_{ik}$ being the zero-mean finite variance measurement error introduced before, $S_{ik} \in [0, T]$ and $k = 1, \ldots, M_i$. In the previous section, we outline the estimation of the mean function $\mu_X(t)$ and the autocovariance operator $G_{XX}$ from noise-contaminated infrequent and irregular observed data. We can recover the eigenfunctions $\phi_\ell(t)$ through a functional principal component step applied to the discretization of the smooth autocovariance estimator $\widehat{G}_{XX}$. Following the works in [25, 26], Senturk and Mueller [28] proposed to recover $\xi_{i\ell}$ from infrequent observations on the longitudinal predictor using Gaussian assumptions on all eigen-scores and measurement error of the longitudinal predictor based on the conditional expectation $E(\xi_{i\ell}|U_i, M_i, S_i)$. Here, $U_i$ is the $M_i \times 1$ observation vector $U_i \equiv (X_{i1}, \ldots, X_{iM_i})^{\mathrm{T}}$ with $X_{ik} = X_i(S_{ik}) + \epsilon_{ik}$, and $M_i$ and $S_i = (S_{i1}, \ldots, S_{iM_i})$ are the total number of repeated measurements and the vector of observation time points for subject $i$, respectively. We refer readers to [28] for explicit expressions of $\hat{\xi}_{i\ell}$. Next, putting together all estimated model components, we reconstruct the predictor process at the observation time points of the response: $\widetilde{X}_{ij} \equiv \hat{\mu}_X(T_{ij}) + \sum_{\ell=1}^{L} \hat{\xi}_{i\ell}\hat{\phi}_\ell(T_{ij})$, $T_{ij} \in [0, T]$ and $j = 1, \ldots, N_i$. Here, we can choose the number $L$ of eigen-components included by various criteria, including AIC and the fraction of variance explained.

For reconstruction in the analysis with a fixed duration, such as our analysis of the CDS data, we estimate the eigenfunctions and eigenvalues of the covariate process via pooling all covariate observations on the common observation period $[T_0, T_1]$. For this, we use predictor trajectories shifted from subject-specific supports $S_{ik} \in [a_i + T_0, a_i + T_1]$ to the common observation period $S_{ik} - a_i \in [T_0, T_1]$, for example within the [100, 550] day interval after the initiation of dialysis for the CDS data. (The data analysis of Section 4 provides details.) Once the predictor processes are reconstructed at the response observation times $T_{ij} - a_i \in [T_0, T_1]$ within the common observation period, we then shift them back to subject-specific supports $T_{ij} \in [a_i + T_0, a_i + T_1]$ by adding $a_i$. In summary, we reconstruct on subject-specific intervals where the subject was originally observed, without extrapolation.

Using the synchronized data $\left(T_{ij}, \widetilde{X}_{ij}, Y_{ij}\right)$ for $j = 1, \ldots, N_i$, we utilize a local maximum likelihood procedure for estimation of the varying coefficient functions. Assuming that $\beta_0(t)$ and $\beta_1(t)$ have continuous second derivatives, we approximate each function locally by $\beta_0 \approx a_0 + a_1(t - t_0)$ and $\beta_1(t) \approx b_0 + b_1(t - t_0)$ for $t$ in a neighborhood of the fixed time point $t_0$. Local maximum likelihood estimators aim to maximize the local log-likelihood,

$$\ell_n(\mathbf{a}, \mathbf{b}) = \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \ell\left(g\left[a_0 + a_1(T_{ij} - t_0) + \{b_0 + b_1(T_{ij} - t_0)\}\widetilde{X}_{ij}\right], Y_{ij}\right) K_h(T_{ij} - t_0), \quad (4)$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ denotes a kernel function, $h$ is the bandwidth, $\mathbf{a} \equiv (a_0, a_1)^{\mathrm{T}}$, $\mathbf{b} \equiv (b_0, b_1)^{\mathrm{T}}$ and $\ell(\cdot, \cdot)$ denotes the log-likelihood function. Maximizing the local log-likelihood $\ell_n(\mathbf{a}, \mathbf{b})$ results in the local maximum likelihood estimators for the varying coefficient functions $\hat{\beta}_0(t_0) = \hat{a}_0$ and $\hat{\beta}_0(t_0) = \hat{b}_0$. We refer to these second set of estimators obtained for the varying coefficient functions as reconstruction estimators throughout the manuscript. We can implement the maximization using the Newton–Raphson algorithm, with the $r + 1$ iteration update given by

$$\left(\hat{\mathbf{a}}_{r+1}, \hat{\mathbf{b}}_{r+1}\right)^{\mathrm{T}} = \left(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r\right)^{\mathrm{T}} - \left\{\ell_n''\left(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r\right)\right\}^{-1} \ell_n'\left(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r\right), \quad (5)$$

where $\ell_n'(\cdot,\cdot)$ and $\ell_n''(\cdot,\cdot)$ denote the gradient and Hessian matrix of the log-likelihood, respectively. The Appendix gives the explicit forms of the terms involved in the updating step (5) for the Bernoulli and Poisson distributed responses.

Note that the proposed reconstruction for the predictor process is quite different from preprocessing steps such as binning or smoothing a single subject's trajectory. Whereas the latter uses only information from that particular subject, the proposed approach uses the pooled information from all subjects. The functional data analysis framework provides a unique opportunity for synchronizing the response and predictor measurements using the estimated underlying population quantities, moments of the predictor process. Simulation studies and data analysis give comparisons of the two proposed approaches based on functional data analysis, along with a binning coupled with local maximum likelihood approach, outlined in the following.

### 2.4. *Binning and local maximum likelihood estimation*

In this section, we outline an equidistant binning procedure to synchronize the data followed by local maximum likelihood for estimation of the varying coefficient functions, as a baseline method in comparisons with the proposed estimators. For the equidistant binning, we select the maximum number of equidistant bins per subject such that each bin contains at least one repeated measurement on the covariate and the response. In applications, we would select the maximum number of bins from a preliminary set of total number of bins that is determined by the distributions of the subject-specific total number of repetitions for the covariate and the response, $M_i$ and $N_i$, respectively. More specifically, for subject $i$, the binning yields synchronized data $(t_{ib}, X_{ib}, Y_{ib})$ for $b = 1, \ldots, B_i$ bins, where the time $t_{ib}$ denotes the midpoint of the $b$th bin, $X_{ib}$ is the average of the covariate observations $X_{ik}$ and $Y_{ib}$ is the sum of the response observations $Y_{ij}$ falling in bin $b$. We use the sum of the binary and count response values within a bin to yield binomial and Poisson distributed response values, respectively, after binning.

The synchronized data obtained from binning are then fed into a local maximum likelihood procedure for estimation of the varying coefficient functions. For binned data $(t_{ib}, X_{ib}, Y_{ib})$ with possibly reduced total number of subjects $i = 1, \ldots, n_B$ $(n_B \leqslant n)$ and possibly reduced repetitions per subject $b = 1, \ldots, B_i$ $(B_i \leqslant \min\{N_i, M_i\})$, local maximum likelihood estimators aim to maximize the local log-likelihood,

$$\ell_{n_B}(\mathbf{a}, \mathbf{b}) = \frac{1}{\sum_{i=1}^{n_B} B_i} \sum_{i=1}^{n_B} \sum_{b=1}^{B_i} \ell\left(g\left[a_0 + a_1(t_{ib} - t_0) + \{b_0 + b_1(t_{ib} - t_0)\}X_{ib}\right], Y_{ib}\right) K_h(t_{ib} - t_0),$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ denotes a kernel function, $h$ is the bandwidth, $\mathbf{a} \equiv (a_0, a_1)^{\mathrm{T}}$, $\mathbf{b} \equiv (b_0, b_1)^{\mathrm{T}}$ and $\ell(\cdot,\cdot)$ denotes the log-likelihood function. Maximizing the local log-likelihood $\ell_{n_B}(\mathbf{a}, \mathbf{b})$ with similar computations as in Section 2.3, we obtain what we will refer to as binning estimators for the varying coefficient functions $\hat{\beta}_0(t_0) = \hat{a}_0$ and $\hat{\beta}_0(t_0) = \hat{b}_0$.

## 3. Prediction of subject-specific mean response trajectories

Although the traditional estimation methods, such as the binning estimator outlined earlier, can only predict the mean response at the bin midpoints $t_{ib}$, we will utilize functional data analysis techniques to provide smooth predicted mean response trajectories throughout the duration of the study based on the two proposed estimation procedures. We begin by the Karhunen–Loéve expansion of the covariate trajectory of a new subject, $X^*(t) = \mu_X(t) + \sum_{\ell=1}^{\infty} \xi_\ell^* \phi_\ell(t)$, where $\xi_\ell^* = \int_0^T \{X^*(t) - \mu_X(t)\}\phi_\ell(t)\mathrm{d}t$ is the $\ell$th functional principal component score with $E\left(\xi_\ell^*\right) = 0$ and $\mathrm{var}\left(\xi_\ell^*\right) = \rho_\ell$. On the basis of the generalized varying coefficient model (1), we will obtain the predicted response trajectories through

$$E\{Y^*(t)|X^*(t)\} = g\{\eta^*(t)\} = g\left\{\beta_0(t) + \beta_1(t)\mu_X(t) + \beta_1(t)\sum_{\ell=1}^{\infty} \xi_\ell^* \phi_\ell(t)\right\}, \tag{6}$$

where $\eta^*(t) \equiv \beta_0(t) + \beta_1(t)X^*(t)$ and $Y^*(t)$ denotes the generalized response trajectory of a new subject.

We estimate the eigenfunctions $\phi_\ell(t)$ and eigen-scores $\xi_\ell^*$ as outlined in Section 2.3 on the basis of the conditional expectation $E\left(\xi_\ell^* | U^*, M^*, S^*\right)$, where $U^*$ is the $M^* \times 1$ observation vector $U^* \equiv$

$\left(X_1^*, \ldots, X_{M^*}^*\right)^{\mathrm{T}}$ with $X_k^* = X^*\left(S_k^*\right) + \epsilon_j^*$ and $M^*$ and $S^* = \left(S_1^*, \ldots, S_{M^*}^*\right)$ are the total number of repeated measurements and the vector of observation time points of the new subject, respectively. Thus, using this plug-in estimate for (6) enables us to predict individual mean response trajectories in the generalized varying coefficient model by

$$\hat{Y}_L^*(t) = g\{\hat{\eta}^*(t)\} = g\left\{\hat{\beta}_0(t) + \hat{\beta}_1(t)\hat{\mu}_X(t) + \hat{\beta}_1(t)\sum_{\ell=1}^{L}\hat{\xi}_\ell^*\hat{\phi}_\ell(t)\right\}. \tag{7}$$

Here, $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ refer to the moment and reconstruction estimators of the varying coefficient functions proposed, respectively, leading to predictions from the two estimation proposals. Note that the resulting predictions are for $t \in [0, T]$, that is, the trajectory on the entire time domain, not just for the original infrequent observation times $S^*$.

For prediction in analysis with a fixed duration, such as our analysis of the CDS data, we begin by predicting subject-specific predictor trajectories on the interval $[T_0, T_1]$ by using all predictor trajectories shifted from subject-specific supports $[a_i + T_0, a_i + T_1]$ to the common observation period $[T_0, T_1]$, similar the reconstruction procedure described in Section 2.3. We then shift back the predicted subject-specific trajectories to subject-specific supports, for example, to $[a_i + T_0, a_i + T_1]$ for subject $i$. Hence, even though the varying coefficient functions are estimated on $[\min(a_i) + T_0, \max(a_i) + T_1]$, we only need the estimated functions from $[a_i + T_0, a_i + T_1]$ to obtain our predicted response trajectories on $[a_i + T_0, a_i + T_1]$ for subject $i$. In this way, we do not extrapolate in prediction of the fixed duration case; we predict on subject-specific time intervals where the subject was originally observed.

## 4. Application to data from the comprehensive dialysis study

### *4.1. Description of Comprehensive Dialysis Study data and analysis cohort*

Longitudinal CRP levels were available for a subset of 266 participants in the CDS study (with one to five measurements per subject), and we obtain their infection-related hospitalization data from the United States Renal Data System database. The median time to the first serum collection is about 6 months after the initiation of dialysis, and then we took samples approximately every 3 months (median) thereafter. The minimum time to first serum sample collection is about 3.4 months; therefore, CRP measurements are mostly between [100, 550] days from the initiation of the dialysis. To avoid boundary bias (due to extremely few data points at the boundaries), we consider longitudinal measurements on a subset of 228 patients with CRP measurements in this [100, 550] day period (approximately 1.2 years). This analysis cohort yields 729 total longitudinal CRP measurements (with 47, 35, 30, 58 and 58 subjects with one to five measurements, respectively). The median baseline age is 63.6 years (standard deviation 10.0). There are 304 total hospitalizations, of which 88 were infection related. The total number of hospitalizations per person ranges between 0 and 16. Because hospitalization events are stochastic, times of infection-related hospitalization statuses (0 or 1) and CRP measurement times are unsynchronized, in addition to being highly irregular and infrequent.

As mentioned earlier, a preliminary binning step to synchronize the data leads to 51% loss in the subject sample size because nearly half of the subjects are missing a response measurement. In terms of the total number of measurements, this represents about 69% data loss. In our analysis in the following, we compare the two proposed estimators, moments and reconstruction, as well as binning estimators. We note that in the application of the binning, we search for the maximum number of bins formed for each subject such that each bin contains at least one covariate and response measurement, as outlined in Section 2.4. We search for the maximum bin number for each subject starting from five (and decreasing to zero) because the maximum total number of CRP measurements per subject and hence the maximum number of bins is five. Nearly half of the subjects have zero bins, with no response measurements, and the subject-specific total number of bins ranges from zero to three. In contrast, individuals with CRP measurements but without any hospitalizations still contribute to the proposed estimation procedures, and in particular, we use them to estimate the autocovariance $G_{XX}(t, t)$ in the proposed methods.

### *4.2. Data analysis results*

To illustrate the proposed methods, we consider estimation of the age-varying association between longitudinal infection-related hospitalization status and a protein inflammation marker, serum CRP levels.

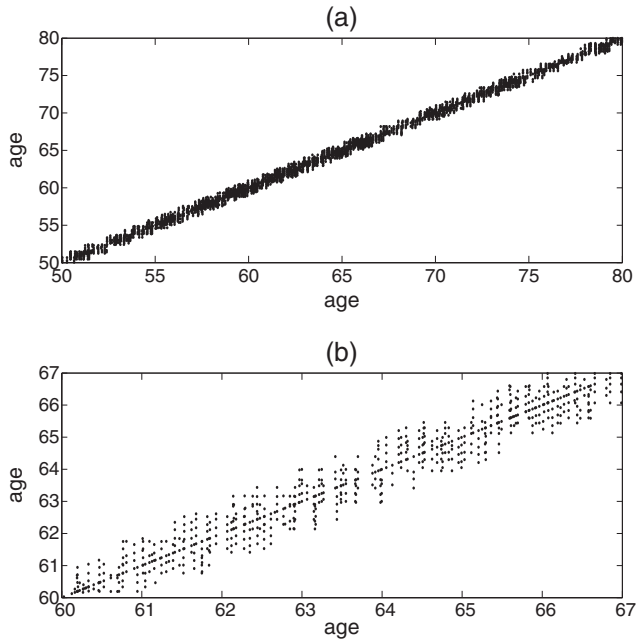


**Figure 1.** (a) Support $(S_{ik}, S_{i\ell})$, $k, \ell = 1, \ldots, M_i$, $i = 1, \ldots, n$, of the autocovariance for the covariate process. (b) A closer view of the support for ages between 60 and 67 years.

Because CRP has a skewed distribution, we take the logarithm transformation for our analysis. Figure 1 gives the autocovariance support in age scale of log(CRP), where both the autocovariance and cross-covariance are estimated within about a 1.2-year band ([100, 550] days from the initiation of dialysis) around the diagonal. Figure 2(a, b) gives the estimated cross-sectional mean functions for the response and the covariate, where the infection probability and mean log(CRP) concentration are increasing slightly with age.

Figure 2 also gives the estimated age-varying coefficient functions from the three estimators: moments, reconstruction and binning. More specifically, the middle row contains comparisons of the moment (solid) and binning (dash-dotted) fits along with moment-based bootstrap confidence intervals (dotted), whereas the last row contains comparisons of the reconstruction (solid) and binning (dash-dotted) fits along with reconstruction-based bootstrap confidence intervals (dotted). We utilize percentile pointwise confidence intervals from 500 bootstrap samples obtained by resampling from subjects with repetition. Both moment and reconstruction approaches yield slightly more positive age-varying slope functions for the relationship between infection-related hospitalization status and log(CRP) over parts of the 60 to 80-year age range than the binning approach that hovers around zero over age. This observed positive association between infection-related hospitalization and inflammation markers, particularly CRP, is consistent with well-established findings in the literature on infection and CRP ([32] and references therein). Although small sample sizes lead to wider bootstrap confidence intervals, as expected, we found the estimated slope functions based on the moment and reconstruction methods to be significant with the 90% confidence intervals, especially in the mid age range of 63–69 and (60–64, 72–74) years.

Next, we consider in more detail the predicted subject-specific mean response trajectories, corresponding to the predicted probability of having an infection-related hospitalization $P\{Y(t)\} = 1$, obtained for all subjects with the predicted subject's data left out, provided from moment and reconstruction estimates separately in Figure 3. Predictions based on the reconstruction estimates seem to vary less in general compared with predictions based on moment estimates, especially in regions around ages 55, 66 and 78 years as these are the values where the reconstruction estimates of the slope function in Figure 3(f) cross 0. In addition, distinct concave and convex patterns in the 100 to 550 days from the initiation of dialysis are evident for subject-specific trajectories in both sets of predictions, with estimated infection-related hospitalization probabilities roughly above ('high' risk) or below ('low' risk) the mean probability of infection of 0.3. Figure 4 displays the predicted response trajectories for these high-risk and low-risk groups on the basis of moment estimators with their corresponding log(CRP) trajectories used in the predictions. The observed concave and convex age-varying subject-specific infection

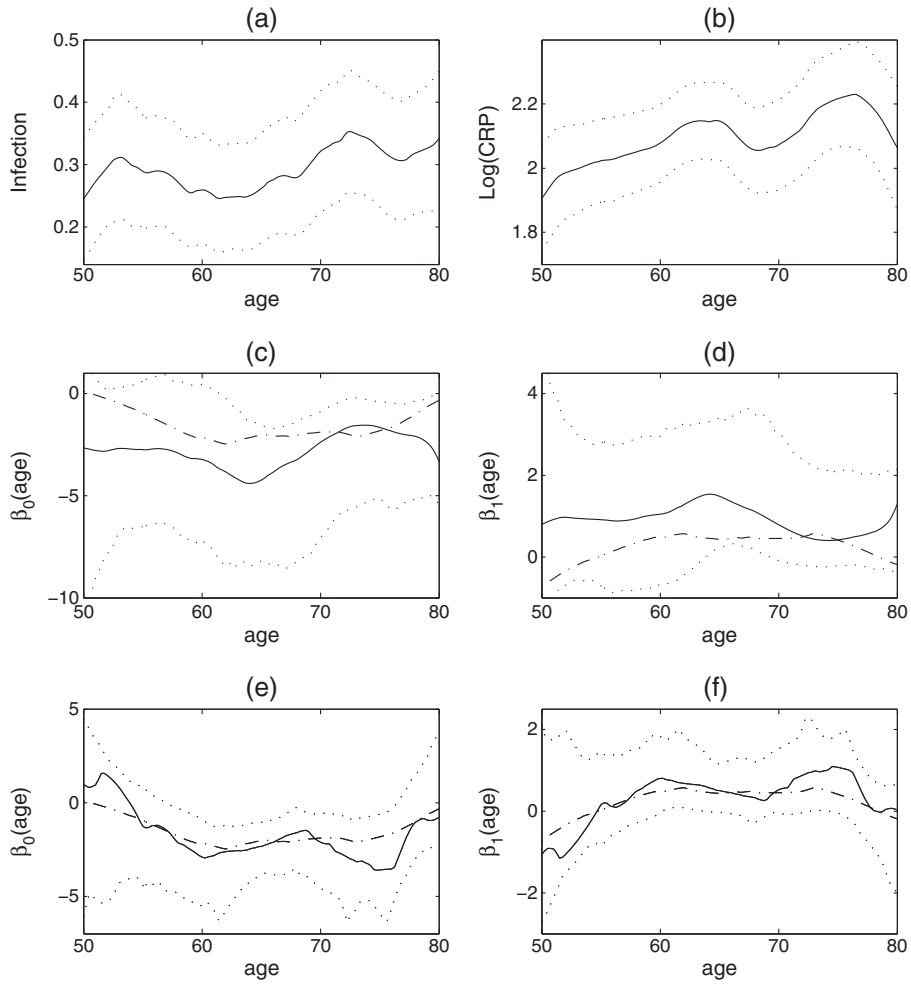


**Figure 2.** (a, b) The smoothed cross-sectional estimate of the mean function $\hat{\mu}_Y(age)$ (solid) for the presence of an infection-related hospitalization (a) and $\hat{\mu}_X(age)$ (solid) of log(CRP) (b) along with $\pm 2$ sliding window standard deviation error bars (dotted). (c, d) Estimated varying coefficient functions $\beta_0(age)$ (c) and $\beta_1(age)$, the slope function of log(CRP) (d) from the proposed moment fits (solid) along with moment-based 90% bootstrap confidence intervals (dotted) for the Comprehensive Dialysis Study data. Also displayed are the estimated functions from the binning fits (dash-dotted). (e, f) Estimated varying coefficient functions from proposed reconstruction fits (solid) and binning fits (dash-dotted) along with reconstruction-based 90% bootstrap confidence intervals (dotted). CRP, C-reactive protein.

trajectories/patterns correspond to the subject-specific covariate log(CRP) trajectories (Figure 4(b, e)), explaining the patterns of the predicted mean response trajectories. That is, the high-risk and low-risk groups for infection-related hospitalization also correspond to patients with higher (mean CRP = 15.5) and lower (mean CRP = 5.0) CRP concentrations, respectively, roughly above and below log(CRP) = 2. We also plot the estimated mean trajectories of log(CRP) for these two groups, where similar concave and convex patterns exist for the mean log(CRP) trajectories within 1.2 years from the initiation of dialysis (Figure 4(c, f)).

## 5. Simulation studies

### *5.1. Simulation design*

We carry out three simulation studies to evaluate the performance of the proposed estimators for both binary and count responses. We study the properties of the proposed estimation procedures under three cases: (a) unsynchronized Bernoulli response; (b) unsynchronized Poisson response; and (c)

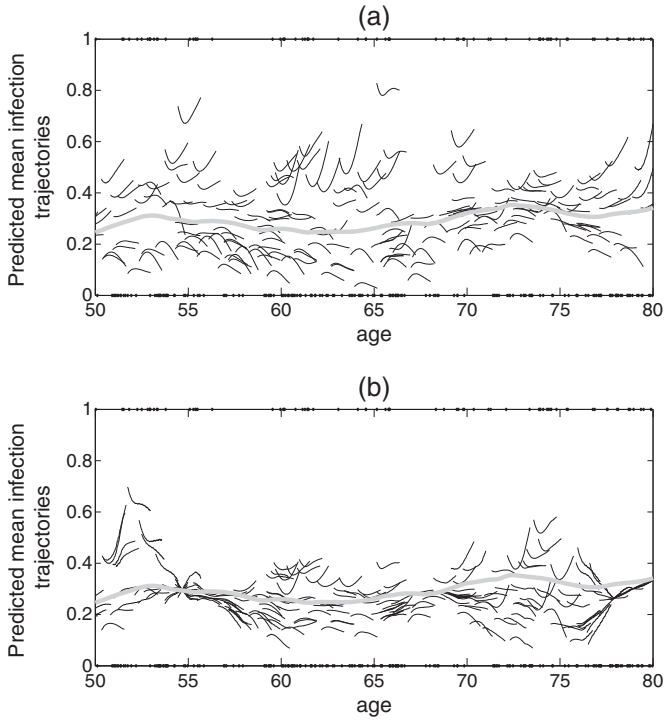


**Figure 3.** (a, b) Observed values (dots) for the presence of an infection-related hospitalization and predicted subject-specific mean response curves (solid) based on moment (a) and reconstruction estimates (b). Also displayed (thick solid gray) is the smoothed estimate of the mean function $\hat{\mu}_Y(age)$ of infection.

unsynchronized Bernoulli response with fixed study length resulting in a diagonal support for an age-varying coefficient model (analogous to the CDS analysis described earlier). Thus, cases (a) and (b) will allow us to more thoroughly assess the performance of the proposed methods generally, whereas case (c) is designed to mimic the characteristics of the CDS data. In all three scenarios, we compare the proposed estimation algorithms and moment and reconstruction estimators along with the baseline method of binning described in Section 2.4. We generate the covariate process $X$ according to $X_i(t) = \mu_X(t) + \xi_{i1}\phi_1(t) + \xi_{i2}\phi_2(t)$, where the functional principal component scores $\xi_{i1}$ and $\xi_{i2}$ are simulated from independent normals with means zero and variances equal to $\sigma^2$. To study the sensitivity of the moment estimators to different $\delta$ values (refer to Taylor's expansions in Section 2.1), which correspond to different levels of variation in the covariate $X$, we report results for three different variances $\sigma^2$ of 2, 4 and 6. We assume the predictor trajectories, which are simulated independently from a Gaussian distribution with zero mean and variance equal to 0.3, to be observed with measurement error. We base reported results for all simulations on 200 Monte Carlo runs. We give technical details of the three data cases as follows:

**Case (a):** We randomly choose the number of repeated measurements for $n = 100$ and 200 subjects between 5 and 15 with equal probabilities, independent for the response $(Y)$ and the covariate $(X)$ to create unsynchronized design. We randomly select the observation times $T_{ij}$ and $S_{ik}$ for each subject, independent for the covariate and the response from the time interval $[0, 10]$. The mean function and two eigenfunctions for the predictor process are $\mu_X(t) = 4\sin(\pi t/5)/\sqrt{5}$, $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$ and $\phi_2(t) = \sin(\pi t/10)/\sqrt{5}$, $0 \leqslant t \leqslant 10$, respectively. The varying coefficient functions are $\beta_0(t) = \sin(\pi t/5)$ and $\beta_1(t) = -\sin(\pi t/10)$. We simulate the response $Y_{ij}$ from a Bernoulli distribution with mean $E\{Y_{ij}|X_i(T_{ij})\} = g\{\beta_0(T_{ij}) + \beta_1(T_{ij})X_i(T_{ij})\}$, where $g(p) = e^p/(1+e^p)$.

**Case (b):** For the count response, we generate the sample size, the number of repetitions, the observation times and the predictor process in the same way as in case (a). The varying coefficient functions are $\beta_0(t) = t/5$ and $\beta_1(t) = \sin(\pi t/10)/3$. We simulate the response $Y_{ij}$ from a Poisson distribution with mean $E\{Y_{ij}|X_i(T_{ij})\} = g\{\beta_0(T_{ij}) + \beta_1(T_{ij})X_i(T_{ij})\}$, where $g(p) = e^p$.

**Case (c):** To simulate data similar to the analyzed CDS data in Section 4, we choose the number of repeated measurements for the response between zero and seven with probabilities

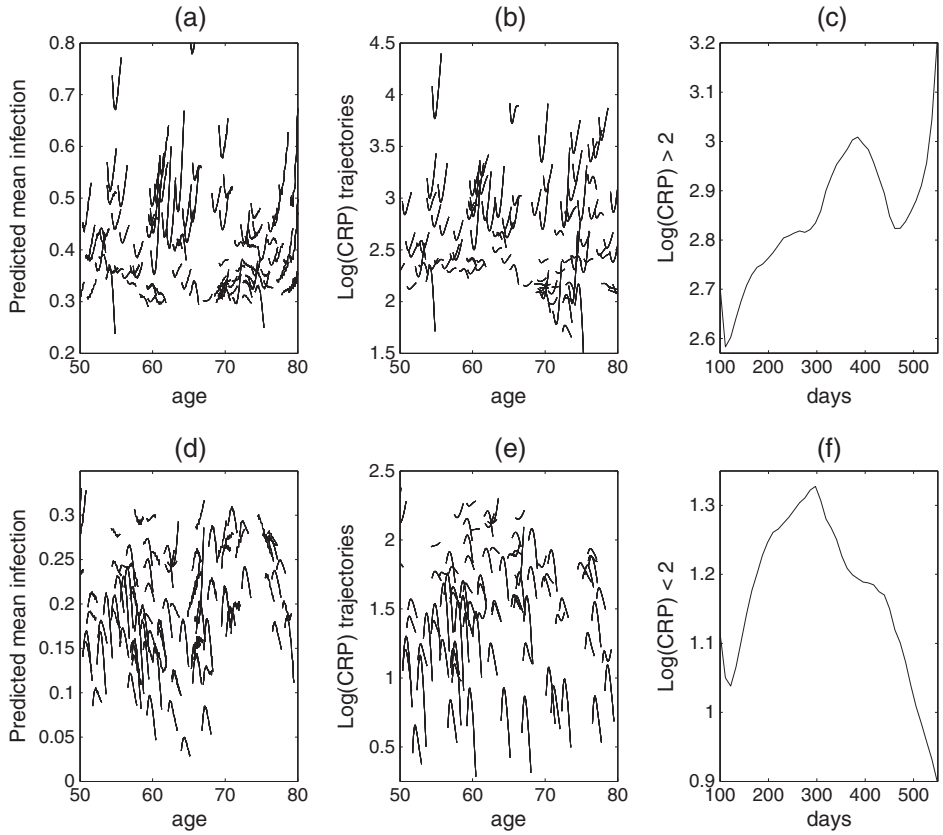


**Figure 4.** (a) Predicted subject-specific infection-related hospitalization probability trajectories based on moment estimates for the high-infection-risk and (d) lower-infection-risk groups. (b) Log(CRP) trajectories of the subjects with high infection probabilities corresponding to (a) and (e) and similarly for subjects with lower infection-related hospitalization probability corresponding to (d). (c) Smoothed estimate $\hat{\mu}_X$ of the mean log(CRP) trajectories for log(CRP) values higher than 2 and (f) lower than 2. CRP, C-reactive protein.

$[0.5, 0.15, 0.1, 0.1, 0.05, 0.05, 0.025, 0.025]$ and between one and five for the covariate with probabilities $[0.15, 0.15, 0.20, 0.25, 0.25]$ for $n = 200$ and 400 subjects, respectively. We randomly choose the baseline age $a_i$ of each subject from $[50, 79]$ where the repeated measurement times for that subject are then chosen randomly from the time interval $[a_i, a_i + 1.2]$, separate for the response and covariate processes. This leads to unsynchronized data, corresponding to a fixed (average) follow-up time of approximately 1.2 years, similar to diagonal support of the age-varying coefficient model for the CDS data. The mean function and the two eigenfunctions for the covariate process are $\mu_X(t) = \sin\{\pi(t-50)/30)\}\sqrt{15}$, $\phi_1(t) = -\cos\{\pi(t-50)/30\}/\sqrt{15}$ and $\phi_2(t) = \sin\{\pi(t-50)/30\}/\sqrt{15}$, $50 \leqslant t \leqslant 80$, respectively. The age-varying coefficient functions are $\beta_0(t) = -1.5\sin\{\pi(t-50)/30\}$ and $\beta_1(t) = -2\sin\{\pi(t-65)/30\}$. We simulate the response $Y_{ij}$ from a Bernoulli distribution with mean $E\{Y_{ij}|X_i(T_{ij})\} = g\{\beta_0(T_{ij}) + \beta_1(T_{ij})X_i(T_{ij})\}$, where $g(p) = e^p/(1+e^p)$.

To study the performance of the proposed estimation method for the varying coefficient functions, we use relative mean squared deviation error (ME):

$$\mathrm{ME}_0 = \frac{\int \left\{\beta_0(t) - \hat{\beta}_0(t)\right\}^2 \mathrm{d}t}{\int \beta_0^2(t)\mathrm{d}t} \quad \text{and} \quad \mathrm{ME}_1 = \frac{\int \left\{\beta_1(t) - \hat{\beta}_1(t)\right\}^2 \mathrm{d}t}{\int \beta_1^2(t)\mathrm{d}t}.$$

Overall ME will be the average of $\mathrm{ME}_0$ and $\mathrm{ME}_1$. For comparisons, we also obtain the ME values for the three estimation procedures, moments ($\mathrm{ME}_\mathrm{M}$), reconstruction ($\mathrm{ME}_\mathrm{R}$) and binning ($\mathrm{ME}_\mathrm{B}$), in the three simulation cases. In the implementation of the equidistant binning algorithm, for cases (a) and (b), we divide the study period $[0, 10]$ into equidistant bins, and for case (c), we divide the subject-specific observation interval $[\min_{k,j}\{S_{ik}, T_{ij}\}, \max_{k,j}\{S_{ik}, T_{ij}\}]$ into equidistant bins to synchronize the data prior to local maximum likelihood estimation.

In addition, we use relative mean squared prediction error ($\mathrm{PE}_i$) to study the proposed subject-specific mean response trajectories, where

$$\mathrm{PE_i} = \frac{\int [g\{\eta_i(t)\} - g\{\hat{\eta}_i(t)\}]^2 \mathrm{d}t}{\int g^2\{\eta_i(t)\}\mathrm{d}t},$$

with $\eta_i(t) = \beta_0(t) + \beta_1(t)X_i(t)$ and $\hat{\eta}_i(t) = \hat{\beta}_0 + \hat{\beta}_1\hat{\mu}_X(t) + \hat{\beta}_1 \sum_{\ell=1}^{L} \hat{\xi}_{i\ell}\hat{\phi}_\ell(t)$ based on moment ($\mathrm{PE_M}$) and reconstruction estimates ($\mathrm{PE_R}$). To examine how the estimated population moments affect the overall quality of the varying coefficient function moment estimators given in (3), we define the following similar relative deviation quantities:

$$\mathrm{ME}_{YX} = \frac{\int \left\{G_{YX}(t,t) - \widehat{G}_{YX}(t,t)\right\}^2 \mathrm{d}t}{\int G_{YX}^2(t,t)\mathrm{d}t}, \qquad \mathrm{ME}_{XX} = \frac{\int \left\{G_{XX}(t,t) - \widehat{G}_{XX}(t,t)\right\}^2 \mathrm{d}t}{\int G_{XX}^2(t,t)\mathrm{d}t},$$

$$\mathrm{ME}_{\mu_Y} = \frac{\int [g^{-1}\{\mu_Y(t)\} - g^{-1}\{\hat{\mu}_Y(t)\}]\mathrm{d}t}{\int [g^{-1}\{\mu_Y(t)\}]^2\mathrm{d}t} \quad \text{and} \quad \mathrm{ME}_{\mu_X} = \frac{\int \{\mu_X(t) - \hat{\mu}_X(t)\}^2\mathrm{d}t}{\int \mu_X^2(t)\mathrm{d}t}.$$

### 5.2. *Simulation results*

Figure 5 gives the cross-sectional medians and the 5% and 95% cross-sectional percentiles of the estimated varying coefficient functions from moment and binning methods for the simulation cases (a), (b) and (c) with $\sigma^2 = 4$. (In this plot, we omit percentiles of the reconstruction estimates that are close to the percentiles of the moment estimates.) For the proposed methods, the median varying coefficient estimates track the true coefficient functions more closely for all three cases. The (median) binning estimates deviate substantially more from the true underlying functions relative to the proposed estimates, especially in simulation cases (a) and (b) (Figure 5). Results are similar for other simulation studies with varying $\sigma^2$ (not shown).

Tables I and II provide a more detailed summary of the performance of the methods, with respect to the relative mean squared deviation error (ME) and subject-specific relative mean squared prediction error (PE). More specifically, provided in Tables I and II are the median, 25% and 75% percentiles of ME and PE for the proposed moment and reconstruction estimators over all three simulation cases. Also reported for comparison is the ratio of the MEs for the proposed methods over the binning approach,

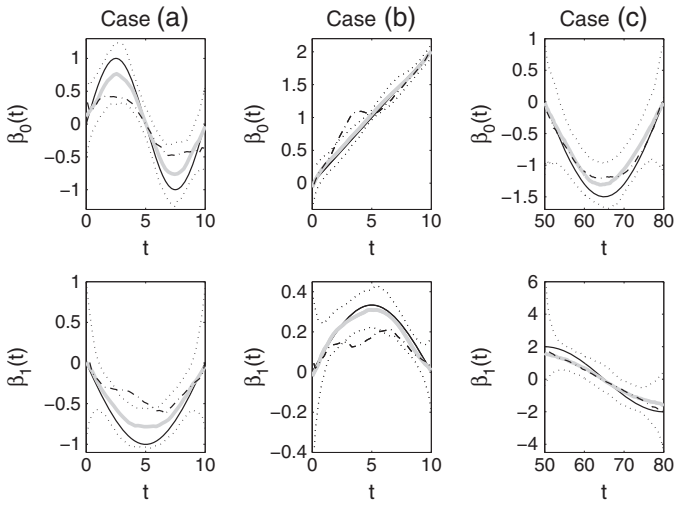


**Figure 5.** Simulation results on estimated varying coefficient functions $\beta_0(t)$ and $\beta_1(t)$ for the three simulation setups: case (a), unsynchronized binary response; case (b), unsynchronized count response; case (c), unsynchronized binary response with diagonal support analogous to the CDS data. Displayed results are from simulations with $\sigma^2 = 4$. The cross-sectional median curves of the proposed moment estimates (thick solid gray) along with 5% and 95% cross-sectional percentiles (dotted) are plotted along with the true varying coefficient functions (solid). Also displayed are the cross-sectional median curves from binning fits (dash-dotted). Percentiles presented are based on 200 Monte Carlo runs/data sets.

**Table I.** Relative mean squared deviation error based on moment ($ME_M$) and reconstruction ($ME_R$) estimates and ratios of two MEs (moment ($r_{ME,\,MB}$) and reconstruction ($r_{ME,\,RB}$) estimates over binning estimates) for the three simulation setups: case (a), general unsynchronized binary response; case (b), general unsynchronized count response; case (c), unsynchronized binary response with diagonal support analogous to the Comprehensive Dialysis Study data.

| | | | $ME_M$ | | | $ME_R$ | | | $r_{ME,\,MB}$ | | | $r_{ME,\,RB}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case | $n$ | $\sigma^2$ | Med | 25% | 75% | Med | 25% | 75% | Med | 25% | 75% | Med | 25% | 75% |
| (a) | 100 | 2 | 0.182 | 0.134 | 0.254 | 0.167 | 0.114 | 0.234 | 0.197 | 0.112 | 0.344 | 0.179 | 0.106 | 0.322 |
| (a) | 100 | 4 | 0.136 | 0.097 | 0.196 | 0.101 | 0.069 | 0.146 | 0.212 | 0.137 | 0.329 | 0.153 | 0.090 | 0.252 |
| (a) | 100 | 6 | 0.115 | 0.086 | 0.159 | 0.081 | 0.060 | 0.110 | 0.222 | 0.147 | 0.319 | 0.148 | 0.088 | 0.224 |
| (a) | 200 | 2 | 0.114 | 0.080 | 0.154 | 0.084 | 0.056 | 0.119 | 0.183 | 0.124 | 0.252 | 0.138 | 0.088 | 0.196 |
| (a) | 200 | 4 | 0.083 | 0.062 | 0.118 | 0.053 | 0.035 | 0.070 | 0.200 | 0.153 | 0.280 | 0.118 | 0.080 | 0.183 |
| (a) | 200 | 6 | 0.087 | 0.064 | 0.117 | 0.042 | 0.032 | 0.059 | 0.231 | 0.162 | 0.325 | 0.109 | 0.073 | 0.158 |
| (b) | 100 | 2 | 0.084 | 0.048 | 0.124 | 0.061 | 0.039 | 0.089 | 0.223 | 0.134 | 0.398 | 0.162 | 0.090 | 0.276 |
| (b) | 100 | 4 | 0.040 | 0.027 | 0.066 | 0.028 | 0.018 | 0.040 | 0.156 | 0.107 | 0.321 | 0.112 | 0.067 | 0.183 |
| (b) | 100 | 6 | 0.034 | 0.022 | 0.049 | 0.021 | 0.013 | 0.031 | 0.178 | 0.105 | 0.290 | 0.116 | 0.065 | 0.174 |
| (b) | 200 | 2 | 0.040 | 0.024 | 0.065 | 0.026 | 0.016 | 0.042 | 0.158 | 0.095 | 0.256 | 0.093 | 0.059 | 0.172 |
| (b) | 200 | 4 | 0.024 | 0.017 | 0.038 | 0.015 | 0.009 | 0.022 | 0.135 | 0.096 | 0.217 | 0.080 | 0.049 | 0.137 |
| (b) | 200 | 6 | 0.019 | 0.013 | 0.026 | 0.010 | 0.007 | 0.014 | 0.134 | 0.095 | 0.194 | 0.068 | 0.045 | 0.109 |
| (c) | 200 | 2 | 0.217 | 0.133 | 0.315 | 0.184 | 0.121 | 0.282 | 0.679 | 0.417 | 1.073 | 0.589 | 0.340 | 0.886 |
| (c) | 200 | 4 | 0.160 | 0.119 | 0.226 | 0.128 | 0.087 | 0.198 | 0.693 | 0.395 | 1.214 | 0.539 | 0.345 | 0.888 |
| (c) | 200 | 6 | 0.158 | 0.114 | 0.218 | 0.108 | 0.072 | 0.165 | 0.657 | 0.438 | 1.216 | 0.487 | 0.306 | 0.797 |
| (c) | 400 | 2 | 0.134 | 0.087 | 0.180 | 0.101 | 0.074 | 0.142 | 0.681 | 0.450 | 1.051 | 0.543 | 0.380 | 0.863 |
| (c) | 400 | 4 | 0.102 | 0.072 | 0.134 | 0.076 | 0.055 | 0.097 | 0.804 | 0.564 | 1.070 | 0.636 | 0.418 | 0.896 |
| (c) | 400 | 6 | 0.101 | 0.070 | 0.129 | 0.059 | 0.044 | 0.088 | 0.913 | 0.648 | 1.265 | 0.565 | 0.378 | 0.826 |

We present median, 25% and 75% percentiles of the deviation measures on the basis of 200 Monte Carlo runs/data sets.

**Table II.** Relative mean squared prediction error based on moment ($PE_M$) and reconstruction ($PE_R$) estimates for the three simulation setups: case (a), general unsynchronized binary response; case (b), general unsynchronized count response; case (c), unsynchronized binary response with diagonal support analogous to the Comprehensive Dialysis Study data.

| | | | $PE_M$ | | | $PE_R$ | | |
|---|---|---|---|---|---|---|---|---|
| Case | $n$ | $\sigma^2$ | Med | 25% | 75% | Med | 25% | 75% |
| (a) | 100 | 2 | 0.013 | 0.008 | 0.021 | 0.012 | 0.008 | 0.019 |
| (a) | 100 | 4 | 0.013 | 0.008 | 0.022 | 0.012 | 0.007 | 0.019 |
| (a) | 100 | 6 | 0.015 | 0.009 | 0.025 | 0.012 | 0.007 | 0.020 |
| (a) | 200 | 2 | 0.008 | 0.005 | 0.013 | 0.007 | 0.004 | 0.011 |
| (a) | 200 | 4 | 0.009 | 0.006 | 0.015 | 0.007 | 0.004 | 0.011 |
| (a) | 200 | 6 | 0.010 | 0.006 | 0.017 | 0.007 | 0.004 | 0.011 |
| (b) | 100 | 2 | 0.005 | 0.003 | 0.008 | 0.004 | 0.003 | 0.007 |
| (b) | 100 | 4 | 0.005 | 0.003 | 0.009 | 0.004 | 0.002 | 0.006 |
| (b) | 100 | 6 | 0.007 | 0.004 | 0.011 | 0.004 | 0.002 | 0.006 |
| (b) | 200 | 2 | 0.003 | 0.002 | 0.005 | 0.002 | 0.001 | 0.004 |
| (b) | 200 | 4 | 0.004 | 0.002 | 0.006 | 0.002 | 0.002 | 0.004 |
| (b) | 200 | 6 | 0.005 | 0.003 | 0.007 | 0.002 | 0.001 | 0.004 |
| (c) | 200 | 2 | 0.043 | 0.012 | 0.129 | 0.054 | 0.014 | 0.156 |
| (c) | 200 | 4 | 0.050 | 0.013 | 0.161 | 0.055 | 0.015 | 0.160 |
| (c) | 200 | 6 | 0.058 | 0.015 | 0.202 | 0.051 | 0.014 | 0.161 |
| (c) | 400 | 2 | 0.033 | 0.008 | 0.100 | 0.045 | 0.012 | 0.132 |
| (c) | 400 | 4 | 0.042 | 0.011 | 0.137 | 0.047 | 0.013 | 0.138 |
| (c) | 400 | 6 | 0.046 | 0.012 | 0.161 | 0.042 | 0.011 | 0.124 |

We estimate percentiles of the deviation measures presented from 200 Monte Carlo runs/data sets.

denoted by $r_{ME,\,MB}$ and $r_{ME,\,RB}$ for the moment and reconstruction methods, respectively. The ratios $r_{ME}$ reported in Table I are roughly fluctuating around 0.1 to 0.2 for general unsynchronized data, cases (a) and (b), and around 0.5 to 0.75 for unsynchronized data with diagonal support analogous to the CDS data (i.e., case (c)). Thus, the efficiency gain of the proposed methods over binning is about 80% to 90%

for general unsynchronized data cases and about 25% to 50% for special case (c) with age-varying coefficient model under fixed-study-length analysis similar to the analysis of CDS data. The relative mean squared prediction errors (PE in Table II) of the proposed prediction methods are quite small for all three simulation cases, showing clearly the efficacy of the proposed subject-specific predictions.

There are several clear conclusions that can be drawn. First, there are gains over the binning combined with the local likelihood approach in all three cases (a, b and c), which are consistent with sample size. Second, the gains of the proposed approaches over binning are greater in cases (a) and (b), where the time domain is common to all subjects. And this is expected because synchronizing the data through binning is expected to lead to more loss of information in the respective time index in cases (a) and (b). In cases (a) and (b), the time index for the varying coefficient model considered is the duration of the study, and the predictor and response measurements span the entire observation interval, leading to significant loss of information for synchronization methods such as binning. In contrast, in case (c), the time index of the varying coefficient model considered is age, and we observe the predictor and response measurements only on subject-specific subsets of the entire age domain, which leads to loss of less information in the age timescale when data are synchronized via binning.

As for the comparison between the two proposed methods, moment and reconstruction approaches, we note that the reconstruction method leads to more favorable results over the moment estimators in estimation of the varying coefficient functions and prediction for simulation cases (a) and (b); moment estimators lead to smaller relative prediction error for simulation case (c), $\sigma^2 = 2$ and 4. Nevertheless, the moment approach provides a more direct approach to estimation without need for data synchronization, leading to computational savings compared with the reconstruction approach, especially in simulation case (c). For example, in simulation case (c), whereas the moment approach estimates the mean and covariance processes of the observed data only along the diagonal, the reconstruction approach additionally reconstructs each predictor trajectory on the basis of the functional principle component decomposition of the shifted predictor trajectories. We note here that most of the computations involved in the proposed estimation algorithms (estimation of the moments of the predictor and response processes including the bivariate smoothing procedures and the choice of appropriate bandwidths and eigen-components, as well as reconstruction of the predictor trajectories) can be carried out with the publicly available software package PACE (http://anson.ucdavis.edu/∼ntyang/PACE; [20, 21, 25, 26]).

Because the small $\delta$ assumption needed in the derivation of the moment estimators cannot be checked in applications, we study the sensitivity of the moment varying coefficient function estimators and predicted response trajectories to the constant $\delta$, regulating the variance of the covariate process. Table I gives summaries of the error measures for three different variation levels of $X$, namely $\sigma^2 = 2, 4$ and 6. Table III gives more detailed summaries of the deviation of the separate components of the varying coefficient functions proposed in (3), specifically $\mathrm{ME}_{YX}$, $\mathrm{ME}_{XX}$, $\mathrm{ME}_{\mu_Y}$ and $\mathrm{ME}_{\mu_X}$, for case (a). (The results are similar for the other cases.) As smaller $\sigma^2$ values correspond to a larger proportion of the variation in the observed covariate trajectories because of measurement error, the covariance processes and hence the varying coefficient functions become more difficult to estimate. Hall *et al.* [27] has reported the same phenomenon in estimating eigenfunctions of the covariance processes. Hence, even though smaller $\sigma^2$, hence $\delta$, values correspond to smaller biases from the Taylor expansion approximations, we observe an increase in the overall ME values as $\sigma^2$ decreases. However, we show the relative mean squared prediction error to be quite robust to the variation in the variance of the covariate. We conclude that the exact value of $\sigma^2$ (or equivalently $\delta$) has a modest effect on the errors in estimation of the varying

**Table III.** Sensitivity of the proposed moment estimators to varying degrees of variation in the covariate process, $\sigma^2$, under simulation case (a) for general unsynchronized binary response.

| | | | $\mathrm{ME}_0$ | | | $\mathrm{ME}_1$ | | | $\mathrm{ME}_{YX}$ | $\mathrm{ME}_{XX}$ | $\mathrm{ME}_{\mu_Y}$ | $\mathrm{ME}_{\mu_X}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case | $n$ | $\sigma^2$ | Med | 25% | 75% | Med | 25% | 75% | Med | Med | Med | Med |
| (a) | 100 | 2 | 0.191 | 0.133 | 0.282 | 0.150 | 0.101 | 0.224 | 0.166 | 0.046 | 0.321 | 0.011 |
| (a) | 100 | 4 | 0.156 | 0.112 | 0.222 | 0.107 | 0.074 | 0.154 | 0.115 | 0.043 | 0.406 | 0.015 |
| (a) | 100 | 6 | 0.158 | 0.102 | 0.234 | 0.095 | 0.064 | 0.142 | 0.092 | 0.044 | 0.393 | 0.015 |

We report relative mean squared deviation error for both varying coefficient functions $\beta_0$ ($\mathrm{ME}_0$) and $\beta_1$ ($\mathrm{ME}_1$), along with deviation measures for the components of the varying coefficient function estimators, as defined in Section 5.1. We estimate the presented percentiles of the deviation measures from 200 Monte Carlo runs/data sets.

coefficient functions for both proposed methods (moments and reconstruction), whereas the individual predictions are relatively robust to fluctuations in $\sigma^2$.

## 6. Discussion

The works reported here fill an important methodological gap. For unsynchronized longitudinal data where the time-dependent response and covariate measurements within each individual are measured at distinct time points, no estimation method exists to estimate time-varying effects in a generalized regression relationship. Informal approaches, based on preprocessing steps that use information in single subject trajectories to synchronize the data to make standard estimation methods applicable, as demonstrated in this work, lead to severe loss of data and introduce further estimation bias in irregular and infrequent longitudinal designs. The proposed methods based on the functional data analysis framework resolve these challenges by offering new ways for pooling information from all subjects under challenging longitudinal data structures, characterized by unsynchronized, irregular and infrequent longitudinal measurements.

The proposed methodology was motivated by an age-varying generalized regression model between unsynchronized longitudinal infection-related hospitalization status and serum CRP, a marker of inflammation, in the CDS. One referee has suggested alternative modeling of the CDS data, where the index of the varying coefficient model can be set to time since dialysis (vintage) and baseline age can be included in the model as a cross-sectional covariate. Note that such a model with additional cross-sectional covariates can be readily implemented using the reconstruction estimation approach. The regression relationship as a function of vintage is also generally of interest in dialysis; however, we cannot feasibly address it with the current data as follow time from initiation of dialysis is short. In addition, in this particular application, our scientific interest is the age-varying trends between CRP concentration and infection-related hospitalizations. We finally note that the proposed model with the age index carries different interpretations than the one where the index is time since initiation of dialysis.

As mentioned earlier, the reconstruction method can readily accommodate additional cross-sectional covariates, whereas including additional longitudinal covariates in the model would require further study for both proposed estimation techniques. Developments needed to accommodate additional longitudinal covariates would involve considering separate Taylor's expansions for the additional covariates in the moment approach and would require joint reconstruction of multiple longitudinal predictor trajectories for the reconstruction method. We recognize these as topics that require further research.

It is also of interest to study the asymptotic distributions of the proposed estimators leading to asymptotic inference for the varying coefficient functions of interest. In addition, we can obtain confidence intervals for subject-specific mean response trajectories $g\{\eta^*(t)\}$ given in (6), building onto the proposed confidence intervals of Senturk and Nguyen [29] for $\eta^*(t)$ (the mean response trajectory in a varying coefficient model with continuous response). Senturk and Nguyen proposed asymptotic pointwise confidence intervals for $\eta^*(t)$ of the form $\hat{\eta}^*(t) \pm \Phi(1-\alpha/2)\sqrt{\hat{w}_t}$, where $\Phi(\cdot)$ denotes the Gaussian CDF and $\hat{w}_t$ denotes the estimated asymptotic variance, which is given in [29]. We can consider multiple extensions for generalized varying coefficient models. A naive approach would be to consider the transformed confidence bounds $[g\{\hat{\eta}^*(t) - \Phi(1-\alpha/2)\sqrt{\hat{w}_t}\}, g\{\hat{\eta}^*(t) + \Phi(1-\alpha/2)\sqrt{\hat{w}_t}\}]$, where $\hat{\eta}^*(t)$ is as given in (7). Another approach is to consider the confidence interval $g\{\hat{\eta}^*(t)\} \pm \Phi(1-\alpha/2)g'\{\hat{\eta}^*(t)\}\sqrt{\hat{w}_t}$, where $g'\{\hat{\eta}^*(t)\}$ is used to target $g'\{\eta^*(t)\}$. However, both proposals require further research to assess their properties.

## Appendix A. Newton–Raphson updates for local maximum likelihood estimators

The log-likelihood function $\ell(p_{ij}, Y_{ij})$ in (4) is equal to $y_{ij}\log(p_{ij}) + (1 - y_{ij})\log(1 - p_{ij})$ and $-p_{ij} + y_{ij}\log(p_{ij}) - \log(y_{ij})$ for the Bernoulli and Poisson distributions, respectively, where $p_{ij} = g\left[a_0 + a_1(T_{ij} - t_0) + \{b_0 + b_1(T_{ij} - t_0)\}\widetilde{X}_{ij}\right]$. In addition, for both Bernoulli and Poisson distributions, the Newton–Raphson update given in (5) will have the form

$$\left(\hat{\mathbf{a}}_{r+1}, \hat{\mathbf{b}}_{r+1}\right)^{\mathrm{T}} = \left(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r\right)^{\mathrm{T}} + \left\{\sum_{i=1}^{n} \mathcal{X}_i^{\mathrm{T}} W_{1i}\left(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r\right)\mathcal{X}_i\right\}^{-1} \sum_{i=1}^{n} \mathcal{X}_i^{\mathrm{T}} W_{2i}\widetilde{Y}_i\left(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r\right),$$

where $\mathcal{X}_i \equiv \{1, \ldots, 1; \widetilde{X}_{i1}, \ldots, \widetilde{X}_{iN_i}; (T_{i1} - t_0), \ldots, (T_{iN_i} - t_0); (T_{i1} - t_0)\widetilde{X}_{i1}, \ldots, (T_{iN_i} - t_0)\widetilde{X}_{iN_i}\}^{\mathrm{T}}$ is the predictor matrix of size $N_i \times 4$, $\hat{p}_{ij} \equiv g[\hat{a}_{r0} + \hat{a}_{r1}(T_{ij} - t_0) + \{\hat{b}_{r0} + \hat{b}_{r1}(T_{ij} - t_0)\}\widetilde{X}_{ij}]$, $W_{2i} \equiv \mathrm{diag}\{K_h(T_{i1} - t_0), \ldots, K_h(T_{iN_i} - t_0)\}$ and $\widetilde{Y}_i(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r) \equiv (Y_{i1} - \hat{p}_{i1}, \ldots, Y_{iN_i} - \hat{p}_{iN_i})^{\mathrm{T}}$. For the Bernoulli distribution, $W_{1i}(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r) \equiv \mathrm{diag}\{K_h(T_{i1} - t_0)\hat{p}_{i1}(1 - \hat{p}_{i1}), \ldots, K_h(T_{iN_i} - t_0)\hat{p}_{iN_i}(1 - \hat{p}_{iN_i})\}$, whereas for Poisson, $W_{1i}(\hat{\mathbf{a}}_r, \hat{\mathbf{b}}_r) \equiv \mathrm{diag}\{K_h(T_{i1} - t_0)\hat{p}_{i1}, \ldots, K_h(T_{iN_i} - t_0)\hat{p}_{iN_i}\}$.

## Acknowledgements

A referee suggested the second proposed approach, the reconstruction method. We are very thankful to the insightful comments received by three referees, associate editor and the editor in the review process. This publication was made possible by the National Institute of Diabetes and Digestive and Kidney Diseases grant DK092232 (DS, LSD and DVN) and grant UL1RR024146 from the National Center for Advancing Translational Sciences (LSD and DVN). We thank Barbara Grimes, Department of Biostatistics, University of California, San Francisco, and Yi Mu at UC Davis Department of Public Health Sciences. The interpretation and reporting of the data presented here are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the US government. The Institutional Review Board of the University of California Davis Health System approved this study.

## References

1. Gilbertson DT, Liu J, Xue JL. Projecting the number of patients with end stage renal disease in the United States to the year 2015. *Journal of the American Society of Nephrology* 2005; **16**:3736–3741.
2. Dalrymple LS, Johansen KL, Chertow GM, Cheng SC, Grimes B, Gold EB, Kaysen GA. Infection-related hospitalizations in older patients with ESRD. *American Journal of Kidney Disease* 2010; **56**:522–530.
3. Kutner NG, Johansen KL, Kaysen GA, Pederson S, Chen SC, Agodoa LY, Eggers PW, Chertow GM. The Comprehensive Dialysis Study (CDS): a USRDS special study. *Clinical Journal of the American Society of Nephrology* 2009; **4**:645–650.
4. Cleveland WS, Grosse E, Shyu WM. *Local Regression Models*. Wadsworth & Brooks: Pacific Grove, 1991; 309–376.
5. Hastie T, Tibshirani R. Varying coefficient models. *Journal of the Royal Statistical Society B* 1993; **55**:757–796.
6. Fan J, Zhang W. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics* 2000; **27**:715–731.
7. Fan J, Zhang W. Statistical methods with varying coefficient models. *Statistics and Its Interface* 2008; **1**:179–195.
8. Huang JZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* 2002; **89**:111–128.
9. Huang JZ, Wu CO, Zhou L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 2004; **14**:763–788.
10. Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 1998; **85**:809–822.
11. 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.
12. Wu CO, Chiang CT. Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica* 2000; **10**:433–456.
13. Cai Z, Fan J, Li RZ. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* 2000; **95**:888–902.
14. Qu A, Li R. Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics* 2006; **62**:379–391.
15. Zhang D. Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics* 2004; **60**:8–15.
16. Xiong X, Dubin JA. A binning method for analyzing mixed longitudinal data measured at distinct time points. *Statistics in Medicine* 2010; **29**:1919–1931.
17. Ramsay JO, Silverman BW. *Applied Functional Data Analysis*. Springer-Verlag: New York, 2002.
18. Ramsay JO, Silverman BW. *Functional Data Analysis*, 2nd ed. Springer: New York, 2005.
19. Rice JA. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* 2004; **14**:631–647.
20. Mueller HG. Functional modeling and classification of longitudinal data. *Scandinavian Journal of Statistics* 2005; **32**:223–240.
21. Mueller HG. *Functional Modeling of Longitudinal Data*. Chapman & Hall/CRC: New York, 2009; 223–252.
22. 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.
23. James G, Hastie TJ, Sugar CA. Principal component models for sparse functional data. *Biometrika* 2000; **87**:587–602.
24. Zhou L, Huang JZ, Carroll R. Joint modeling of paired sparse functional data using principle components. *Biometrika* 2008; **95**:601–619.
25. Yao F, Mueller HG, Wang JL. Functional linear regression analysis for longitudinal data. *Annals of Statistics* 2005; **3**:2873–2903.
26. Yao F, Mueller HG, Wang JL. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005; **100**:577–590.

27. Hall P, Mueller HG, Yao F. Modeling sparse generalized longitudinal observations via latent Gaussian processes. *Journal of the Royal Statistical Society Series B* 2008; **70**:703–723.
28. Senturk D, Mueller HG. Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association* 2010; **105**:1256–1264.
29. Senturk D, Nguyen DV. Varying coefficient models for sparse noise-contaminated longitudinal data. *Statistica Sinica* 2011; **21**:1831–1856.
30. Kim K, Senturk D, Li R. Recent history functional linear models for sparse longitudinal data. *Journal of Statistical Planning and Inference* 2011; **141**:1554–1566.
31. Liu B, Mueller HG. *Functional Data Analysis for Sparse Auction Data*. Wiley & Sons: New York, 2008; 269–290.
32. Kaysen GA. Biochemistry and biomarkers of inflamed patients: why look, what to assess. *Clinical Journal of the American Society of Nephrology* 2009; **4**:56–63.