# A Two-step Estimation Approach for Logistic Varying Coefficient Modeling of Longitudinal Data

Jun Dong, Jason P. Estes, Gang Li and Damla Şentürk*

University of California, Los Angeles
*email:* dsenturk@ucla.edu*

**Abstract**

Varying coefficient models are useful for modeling longitudinal data and have been extensively studied in the past decade. Motivated by commonly encountered dichotomous outcomes in medical and health cohort studies, we propose a two-step method to estimate the regression coefficient functions in a logistic varying coefficient model where the outcome is binary. The model depicts time-varying covariate effects without imposing stringent parametric assumptions. The proposed estimation is simple and can be conveniently implemented using existing statistical packages such as SAS and R. We study asymptotic properties of the proposed estimators which lead to asymptotic inference and also develop bootstrap inferential procedures to test whether the coefficient functions are indeed time-varying or are equal to zero. The proposed methodology is illustrated with the analysis of a smoking cessation data set. Simulations are used to evaluate the performance of the proposed method compared to an alternative estimation method based on local maximum likelihood and two parametric modeling approaches: generalized estimating equations (GEE) for logistic regression and the generalized linear mixed models (GLMM).

*Key words and phrases:* Generalized estimating equations, Generalized linear mixed models, Logistic regression, Longitudinal binary data, Smoothing, Time-varying effects

# 1 Introduction

Longitudinal data arise frequently from medical and health cohort studies where the subjects are measured repeatedly over time. Our working example is the smoking cessation data described in Shoptaw, Fuller, Yang, Frosch, Nahom, Jarvik, Rawson and Ling (2002). Follow-up data was collected on 175 participants for 12 weeks in a clinical trial to evaluate two behavioral methods for optimizing smoking cessation outcomes in methadone maintained cigarette smokers. At each visit, samples of breath were measured for carbon monoxide level and a binary outcome representing smoking status was recorded along with many covariates including age, gender and behavioral treatment. Hence, the data is of the form $[\{t_{ij}, X_i(t_{ij}), Y_i(t_{ij})\},\ i = 1, \ldots, n,\ j = 1, \ldots, T_i]$, where $X_i(t_{ij}) = \{X_{i1}(t_{ij}), \ldots, X_{id}(t_{ij})\}^{\mathrm{T}}$ and $Y_i(t_{ij})$ denote the vector of $d$ covariates and the binary response variable for subject $i$, respectively, measured at time $t_{ij}$. Of interest is to assess the potentially time-varying effects of behavioral treatments on the outcomes adjusting for potential risk factors.

Many parametric models have been proposed to analyze longitudinal binary data (Pendergast, Gange, Newton, Windstorm, Palta and Fisher (1996)). A commonly used approach is the method of Liang and Zeger (1986) based on generalized estimating equations (GEE). The GEE approach models the marginal distributions using a generalized linear model and assumes a common correlation matrix for the repeated measurements within each subject. The regression parameters are estimated by solving generalized estimating equations and the method provides consistent estimates of the regression coefficients even if the correlation matrix is misspecified. Another important approach is generalized linear mixed models (GLMM)

$$Y_i(t_{ij})\big|\{X_i(t_{ij}), \beta_i\} \sim \text{Bernoulli}\{\pi_i(t_{ij})\}, \qquad \log\left\{\frac{\pi_i(t_{ij})}{1-\pi_i(t_{ij})}\right\} = X_i(t_{ij})\alpha + Z_i(t_{ij})\beta_i; \quad (1)$$

see McCullagh and Nelder (1989, sec. 14.5), Breslow and Clayton (1993) and the references therein. In (1), $\alpha$ is a vector of unknown fixed effects and $\beta$ is a vector of unknown random

effects coefficients with an unknown covariance matrix $D$. Wolfinger and O'Connell (1993) used iterative reweighted likelihood to fit the GLMM model, and the method has been made available in SAS as a macro called *GLIMMIX*.

Both the GEE and GLMM methods assume constant covariate effects over time. This is a rather stringent assumption that may not always hold in applications. As illustrated by our simulation study in Section 5, both methods could lead to misleading results when the covariate effects are indeed time-varying. Therefore, it is important to check whether or not all the covariate effects are time invariant. Furthermore, even if the covariate effects do not change over time, parametric approaches involving a small number of parameters do not work well when there is a large number of repeated measurements, as the pattern of covariate effects over time may not be fully captured by only a few parameters. Reported simulations in Section 5 show that confidence intervals based on GLMM can have lower coverage probability than the nominal level for finite samples due to underestimation of the variance of the estimators.

In contrast to GEE and GLMM, logistic varying coefficient models for longitudinal binary data have been proposed to allow regression coefficient functions to change over time,

$$Y_i(t)\big|X_i(t) \sim \text{Bernoulli}\left\{\pi_i(t)\right\}, \qquad \log\left\{\frac{\pi_i(t)}{1-\pi_i(t)}\right\} = X_i(t)^{\mathrm{T}}\beta(t), \tag{2}$$

without assuming any parametric form (Cleveland, Grosse and Shyu (1991); Hastie and Tibshirani, (1993); Cai, Fan and Li (2000)). In (2), $\text{corr}\{Y_i(s), Y_{i'}(t)\} = \gamma(s,t)\mathrm{I}_{(i=i')}$, where $\beta(t)$ is a vector of $d$ regression coefficient functions, $\pi_i(t) = \Pr\{Y_i(t) = 1|X_i(t)\}$, and $\gamma(s,t)$ is an unknown bivariate correlation function. In this model, the observations from different subjects are independent and the repeated measurements from the same subject are correlated. The use of this model is two-fold. First, it can be used to check whether or not the effect of a covariate changes over time by plotting the corresponding coefficient function. Secondly, it provides a useful alternative to the GEE and GLMM methods for analyzing longitudinal binary data when the constant covariate effects assumption is not valid.

Recently several works have been proposed for estimation in generalized varying coefficient models. Zhang (2004) extended the GLMM model by representing the covariate effects via smooth but otherwise arbitrary functions of time. They use random effects to model the correlation among and within subjects, and use the double penalized quasi-likelihood method for estimation. However as mentioned in the paper, this approach does not perform well for binary outcomes and may require an additional bias correction step. Qu and Li (2006) proposed an efficient estimation procedure for generalized varying coefficient models for longitudinal data via an integrated quadratic inference function and penalized splines approach. This approach can easily take into account correlation within subjects; however it is still parametric in nature although the dimension of the parameter space is high. Şentürk, Dalrymple, Mohammed, Kaysen and Nguyen (2013) and Estes, Nguyen, Dalrymple, Mu and Şentürk (2014) consider extensions of the local maximum likelihood approach of Cai, Fan and Li (2000) for estimation in generalized varying coefficient models for i.i.d. data to modeling longitudinal data. This extension is shown to be useful in applications where follow-up in longitudinal studies are truncated by death. For estimation in a generalized varying coefficient model from unsynchronized longitudinal data where response and predictors may not be collected at the same time points, Şentürk, Dalrymple, Mohammed, Kaysen and Nguyen (2013) proposed a nonparametric moments approach, while Cao, Zeng, Fine (2014) proposed kernel weighted estimating equations.

As a novel departure from existing literature, we propose a two-step procedure to estimate the coefficient functions in a logistic varying coefficient model. The first step involves fitting a standard logistic regression at each of the observation time point $t_{ij}$. In the second step an estimate of each regression coefficient function is obtained by smoothing the raw estimates from the first step based on a nonparametric regression method. Thus a major advantage of the proposal is that our estimators can be easily obtained using existing statistical softwares.

We point out that our approach is similar to that used by Fan and Zhang (2000) for varying coefficient models with continuous response, referred to by the authors as the functional linear model. However, there is a fundamental difference between a functional linear model and a logistic varying coefficient model in that the raw estimates are unbiased for the linear model, but biased for the logistic regression model for finite samples. The bias for the latter model has to be handled with care when developing the large sample properties of the proposed two-step (TS) estimators. In addition to establishing the asymptotic properties of the TS estimators leading to asymptotic confidence intervals, we also develop bootstrap inferential procedures to test whether the coefficient functions are indeed time-varying or are equal to zero. While the first hypothesis evaluates whether the logistic varying coefficient model reduces to a parametric form, the second can be used in identifying significant predictors.

This paper is organized as follows. The two-step estimation procedure is described in detail in Section 2. In Section 3, the asymptotic properties of the proposed estimators are studied, and statistical inference procedures are discussed. In Section 4, we apply the proposed method to the smoking cessation data described earlier. In Section 5, we present simulation studies to assess and compare the performance of the proposed TS estimation with the local maximum likelihood (LML) approach of Şentürk, Dalrymple, Mohammed, Kaysen and Nguyen (2013) and Estes, Nguyen, Dalrymple, Mu and Şentürk (2014), the GEE method of Liang and Zeger (1986) and the GLMM model implemented in Wolfinger and O'Connell (1993). We conclude with a discussion section and collect technical proofs in an appendix.

## 2 The Proposed Two-step Estimation Procedure

In this section, we derive a two-step estimate for the coefficient function $\beta(t)$. In the first step, a raw estimate of $\beta(t)$ at each design time point is obtained by fitting a standard logistic regression. In the second step, a final estimate of $\beta(t)$ is obtained by smoothing the raw

estimates using a nonparametric curve estimation method. Throughout this paper, we let $\mathcal{D} = [\{t_{ij}, X_i(t_{ij})\},\ i = 1, \ldots, n,\ j = 1, \ldots, T_i]$, which contains the design time points and the covariate information. The range of time is $[0, D]$ for some specified $D$. Note that under model (2), we have $\mathrm{Cov}\{Y_i(t), Y_i(t)|\mathcal{D}\} = \mathrm{Var}\{Y_i(t)|\mathcal{D}\} = \pi_i(t)\{1 - \pi_i(t)\}$ and $\mathrm{Cov}\{Y_i(s), Y_i(t)|\mathcal{D}\} = \gamma(s,t)\,[\mathrm{Var}\{Y_i(s)|\mathcal{D}\} * \mathrm{Var}\{Y_i(t)|\mathcal{D}\}]^{1/2}$, where $\gamma(t,t) = 1$.

### 2.1 Step I: Obtaining the Raw Estimates

Let $A = \{t_j, j = 1, \ldots, T\}$ be the collection of distinct time points among $\{t_{ij}, i = 1, \ldots, n, j = 1, \ldots, T_i\}$. For any $t_j \in A$, let $N_j = \{i_1, \ldots, i_{n_j}\}$ denote the collection of subject indices of all $Y_i(t_{ij})$ observed at $t_j$, where $n_j$ is the number of subjects observed at $t_j$. Then, under model (2), we have at the time $t_j$,

$$Y_i(t_j)\big|X_i(t_j) \sim \text{Bernoulli}\{\pi_i(t_j)\}, \quad \log\left\{\frac{\pi_i(t_j)}{1-\pi_i(t_j)}\right\} = X_i(t_j)^{\mathrm{T}}\beta(t_j), \quad \text{for all } i \in N_j. \tag{3}$$

The raw estimate $b(t_j) = \{b_1(t_j), \ldots, b_d(t_j)\}^{\mathrm{T}}$ is defined as the maximum likelihood estimate of $\beta(t_j) = \{\beta_1(t_j), \ldots, \beta_d(t_j)\}^{\mathrm{T}}$ from the standard logistic regression model (3).

### 2.2 Step II: Refining the Raw Estimates

For the $r$-th component of the coefficient vector, we obtain a refined estimate by smoothing the raw estimates $[\{t_j, b_r(t_j)\}, j = 1, \ldots, T]$, $r = 1, \ldots, d$. For example, the local polynomial smoothing method (Fan and Gijbels (1996)) yields the following linear estimator for the $q^{th}$ derivative of $\beta(t)$, which is assumed to be $(p+1)$-times continuously differentiable for some $p \geq q$:

$$\widehat{\beta_r^{(q)}}(t) = \sum_{j=1}^{T} \omega_{q,p+1}(t_j, t) b_r(t_j) \qquad \text{for } r = 1, \ldots, d, \quad 0 \leq q \leq p+1. \tag{4}$$

The weight functions $\omega_{q,p+1}(t_j, t)$ in (4) are induced by the local polynomial fitting and are defined in the assumptions section given at beginning of the Appendix. Note that the raw estimates of the coefficient functions are defined only at the design time points. However, the

refined estimate $\widehat{\beta_r^{(q)}}(t)$ are defined for all $t \in [0, D]$. Furthermore, it aggregates the information around time $t$.

**Remark 1.** *We note that the raw estimate $b(t_j)$ of $\beta(t_j)$ usually has a finite sample bias that may not be negligible when $n_j$ is small. This bias will be carried over to the refined estimate obtained in the second step and needs to be handled with care when studying the asymptotic properties of the two-step estimator. In practice, one may also run into situations where, for some time point $t_j$, the sample size $n_j$ is smaller than the number of covariates $d$. In such a case, it is impossible to fit a logistic regression at time $t_j$. Similar to the approach by Fan and Zhang (2000) for functional linear models, one could leave $b(t_j)$ missing. This is equivalent to treating observations at these $t_j$'s as if they were not in the data at all. This potentially reduces the bias compared to including them in the calculation. Another possible solution is to increase the sample size by including observations from the neighbors. For instance, one could include observations at $t_{j-1}$ and $t_{j+1}$ to fit the logistic regression at $t_j$. A third approach is to impute the missing observations in the data via getting information from the neighboring time points. As indicated in Fan and Zhang (2000), the bias created by the second and third methods are negligible as long as $\beta(t)$ is smooth and the time window is small.*

**Remark 2.** *In step 2 we define our estimator (4) by smoothing each component separately without utilizing the covariance structure between different components. One could potentially improve our estimator by incorporating the covariance information that is determined by the correlation function $\gamma(s, t)$. However, because the bivariate function $\gamma(s, t)$ is unknown, the efficiency gain could be hard to realize if $\gamma(s, t)$ is not accurately estimated. We choose to use (4) for its simplicity and computational convenience. In addition, the fact that each component is smoothed separately allows the estimation to adapt to the different degrees of smoothness of the varying coefficient regression functions. This is a big advantage of the proposed TS algorithm where bandwidths for smoothing in the second step can be chosen by plotting the raw*

*estimates from the first step or by automatic bandwidth selection algorithms. We utilize plots of the raw estimates in the analysis of the smoking cessation data in Section 4 and utilize the rule-of-thumb bandwidth selection criteria of Ruppert, Sheather and Wand (1995) in the simulation studies presented in Section 5.*

# 3 Asymptotic Properties and Inference

In this section, we investigate the asymptotic bias, variance and normality of the proposed TS estimators. A bootstrap method is also proposed to construct global confidence bands, which enables one to perform hypothesis testing about the coefficient functions. We assume the outcomes at each time point are missing completely at random hereafter.

## 3.1 Asymptotic Properties

Denote the response vector and the design matrix for the logistic regression model (3) at $t_j$ by $\widetilde{Y}_j = \{Y_{i_1}(t_j), Y_{i_2}(t_j), \cdots, Y_{i_{n_j}}(t_j)\}^{\mathrm{T}}$, and $\widetilde{X}_j = \{X_{i_1}(t_j), X_{i_2}(t_j), \ldots, X_{i_{n_j}}(t_j)\}^{\mathrm{T}}$ respectively. The following lemma gives the asymptotic properties of the raw estimators.

**Lemma 1.** *Assume that condition (A4) in the Appendix holds. Assume further that given $\mathcal{D}$,*

*(N1) The covariates are uniformly bounded, i.e., there exists an $M_0$ such that $|X_{ijr}| \leq M_0$, for all $i, j$, and $r$.*

*(N2) Let $I_j = \widetilde{X}_j^{\mathrm{T}} W_j \widetilde{X}_j$ be the Fisher information matrix where $W_j = diag[\pi_{i_1}(t_j)\{1 - \pi_{i_1}(t_j)\}, \ldots, \pi_{i_{n_j}}(t_j)\{1 - \pi_{i_{n_j}}(t_j)\}]$ is the covariance matrix of $\widetilde{Y}_j$. Further let $\lambda_{1,n_j}$ and $\lambda_{\ell,n_j}$ be respectively the smallest and the largest eigenvalue of $I_j$. There exists a random variable $M_1$ such that, with probability 1, $\lambda_{\ell,n_j}/\lambda_{1,n_j} < M_1$, for all $n_j, j$ and $E(M_1) < \infty$.*

*Let $b(t_j)$ be the raw estimate of $\beta(t_j)$ defined in Section 2.1. Then*

$$E\{b(t_j) - \beta(t_j)|\mathcal{D}\} = o\left(n_j^{-1}\right), \quad Cov\{b(t_j)|\mathcal{D}\} = I_j^{-1}\{1 + o(1)\} \quad \text{and}$$

$$Cov\{b(t_j), b(t_k)|\mathcal{D}\} = I_j^{-1} I_{jk} I_k^{-1} \gamma(t_j, t_k)\{1 + o(1)\}, \tag{5}$$

*as* $n_j \to \infty$ *and* $n_k \to \infty$, *where* $I_{jk} = \widetilde{X}_j^{\mathrm{T}} W_j^{1/2} M_{jk} W_k^{1/2} \widetilde{X}_k$. *The* $n_j \times n_k$ *matrix* $M_{jk}$ *is defined as follows: If the* $a^{th}$ *entry of* $\widetilde{Y}_j$ *and the* $b^{th}$ *entry of* $\widetilde{Y}_k$ *come from the same subject, then the* $(a, b)^{th}$ *entry of* $M_{jk}$ *is equal to 1, and is 0 otherwise.*

Note that

$$\begin{aligned} \mathrm{E}\left\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\right\} &= \sum_{j=1}^{T} \omega_{q,p+1}(t_j, t)\mathrm{E}\left\{b_r(t_j)|\mathcal{D}\right\}, \quad \text{and} \\ \mathrm{Var}\left\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\right\} &= \sum_j \sum_k \omega_{q,p+1}(t_j, t)\omega_{q,p+1}(t_k, t)\mathrm{Cov}\left\{b_r(t_j), b_r(t_k)|\mathcal{D}\right\}. \end{aligned} \tag{6}$$

The following theorem gives the asymptotic bias of $\widehat{\beta_r^{(q)}}(t)$.

**Theorem 1.** *Assume that the conditions (A1)-(A6) in the Appendix and the conditions (N1) and (N2) of Lemma 1 hold. Then*

$$\begin{aligned} Bias\left\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\right\} &= \frac{q!\beta_r^{(p+1)}(t)h^{p-q+1}}{(p+1)!} B_{p+1}(K_{q,p+1})\left\{1 + o_p(1)\right\} + O(1/n_\wedge) \\ &= O(h^{p-q+1}) + O(1/n_\wedge), \end{aligned}$$

*as* $T \to \infty$ *and* $n_\wedge = \min\{n_1, \ldots, n_T\} \to \infty$, *for* $r = 1, \ldots, d$ *and* $0 \le q \le p + 1$, *where* $h$ *is the bandwidth for local polynomial smoothing and* $B_{p+1}(K_{q,p+1})$ *is as defined in the Appendix before the proof of Lemma 1.*

We note that the asymptotic bias comes from two sources. The first term is from the smoothing step, which goes to 0 when the bandwidth tends to 0. The second term is from the logistic regression in the first step, since the MLE in ordinary logistic regression is biased. It goes to 0 when the sample sizes go to $\infty$.

The variance of $\widehat{\beta^{(q)}}(t)$ in (6) can be further simplified under more assumptions on the model. First, assume condition (A4) holds and let $\Omega_j = \mathrm{E}[\pi_i(t_j)\{1 - \pi_i(t_j)\}X_i(t_j)X_i(t_j)^{\mathrm{T}}]$, and $\Omega_{jk} = \mathrm{E}[\sqrt{\pi_i(t_j)\{1 - \pi_i(t_j)\}}\sqrt{\pi_i(t_k)\{1 - \pi_i(t_k)\}}X_i(t_j)X_i(t_k)^{\mathrm{T}}]$. Then, for any given time

$t_j$ and $\beta(t_j)$, $I_j = \widetilde{X}_j^{\mathrm{T}} W_j \widetilde{X}_j = \sum_{k=1}^{n_j} \pi_{i_k}(t_j)\{1-\pi_{i_k}(t_j)\} X_{i_k}(t_j) X_{i_k}(t_j)^{\mathrm{T}}$, where $\pi_{i_k}(t_j)\{1-\pi_{i_k}(t_j)\} = \{e^{X_{i_k}(t_j)^{\mathrm{T}}\beta(t_j)}\}/\{1+e^{X_{i_k}(t_j)^{\mathrm{T}}\beta(t_j)}\}^2$, depends on $X_{i_k}(t_j)$ only. Therefore, $I_j$ is a sum of i.i.d. random matrices with $\mathrm{E}(I_j)$=$n_j\Omega_j$. This fact, combined with Lemma 1, implies that

$$\mathrm{Cov}\{b(t_j), b(t_k)|\mathcal{D}\} = I_j^{-1} I_{jk} I_k^{-1} \gamma(t_j,t_k)\{1+o(1)\} = \gamma(t_j,t_k)\frac{n_{jk}}{n_j n_k}\Omega_j^{-1}\Omega_{jk}\Omega_k^{-1}\{1+o_p(1)\}$$

and $\mathrm{Var}\{b(t_j)|\mathcal{D}\} = (\Omega_j^{-1}/n_j)\{1+o_p(1)\}$, with probability 1, where $n_{jk}$ is the number of subjects in $N_j \cap N_k$. Plugging the above equations into (6) gives

$$\begin{aligned}\mathrm{Var}\left\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\right\} =& \Big\{\sum_{j\neq k}\frac{n_{jk}}{n_j n_k}\gamma(t_j,t_k)\omega_{q,p+1}(t_j,t)\omega_{q,p+1}(t_k,t)(\Omega_j^{-1}\Omega_{jk}\Omega_k^{-1})^{(rr)} \\ &+ \sum_j \frac{1}{n_j}\omega_{q,p+1}^2(t_j,t)(\Omega_j^{-1})^{(rr)}\Big\}\{1+o_p(1)\},\end{aligned} \tag{7}$$

where $M^{(rr)}$ denotes the $(r,r)^{th}$ element of a matrix $M$. In general, we can not simplify the formula in (7) without further assumptions. This is because $\Omega_j$ depends on $j$ through $\beta(t_j)$ and $\widetilde{X}_j$, which makes the summation very hard to compute. If the covariates $X_i(t_j)$ and coefficient functions $\beta(t)$ satisfy conditions (A7) and (A8), that is, they are time-invariant, then $\Omega_j = \Omega_k = \Omega_{jk} = \Omega_1$. In this case, $\mathrm{Cov}\{b(t_j), b(t_k)|\mathcal{D}\} = \gamma(t_j,t_k)\{n_{jk}/(n_j n_k)\}\Omega_1^{-1}\{1+o_p(1)\}$ and $\mathrm{Cov}\{b_r(t_j), b_r(t_k)|\mathcal{D}\} = \gamma(t_j,t_k)\{n_{jk}/(n_j n_k)\}\omega^{(rr)}\{1+o_p(1)\}$ where $\omega^{rr} = (\Omega_1^{-1})^{(rr)}$ denotes the $(r,r)^{th}$ element of $\Omega_1^{-1}$.

We will derive the asymptotic variance for two specific situations: $n_{ij}$ is either small or large, as in Fan and Zhang (2000). Let $I_t = \{j : |t_j - t| \leq h\}$ be the indices of the local neighborhood. In some situations, $n_{jk}$ may be much smaller than $n_j$ or $n_k$ for all $j \neq k, j, k \in I_t$ and $n_j, j \in I_t$ are about the same proportion as $n$. Results for this situation are summarized in the following theorem.

**Theorem 2.** *Let conditions (A1)-(A8), (N1) and (N2) hold. Assume*

$$n_{jk}/(n_j n_k) = \begin{cases} o\{1/(nTh^{2q+1})\}, & j \neq k, \\ 1/(cn) + o\{1/(nTh^{2q+1})\}, & j = k \end{cases}$$

*holds uniformly for all* $j, k \in I_t$ *for some constant* $0 < c < 1$*, then when* $h \to 0$ *and* $nTh^{2q+1} \to \infty$ *as* $n, T \to \infty$*,*

$$Var\Big\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\Big\} = \frac{\omega^{rr} q!^2}{cnTh^{2q+1} f(t)} V(K_{q,p+1})\{1 + o_p(1)\},$$

*where* $V(\cdot)$ *is as defined in the Appendix before the proof of Lemma 1 and* $f(\cdot)$ *denotes the density of* $t$*.*

The proof of Theorem 2 is similar to the proof of Theorem 2 of Fan and Zhang (2000) except that $\gamma(t,t)$ is 1 and therefore is not included in the above result. Recall that they define $\gamma(s,t)$ as the covariance function of the process, and we define it as the correlation function.

In some other situations, $n_j$, $n_k$ and $n_{jk}$ may be about the same as $n$. An extreme case is a dataset with no missing values, in which $n_j = n$ for all $j = 1, \ldots, T$. Let $\gamma_{\alpha,\beta}(s,t)$ denote $\partial^{\alpha+\beta}\gamma(s,t)/\partial s^\alpha \partial t^\beta$ for any integers $\alpha, \beta = 0, 1, \ldots, p+1$.

**Theorem 3.** *Let conditions (A1)-(A8), (N1) and (N2) hold. Assume* $n_{jk}/(n_j n_k) = 1/n + o(1/n)$ *holds uniformly for all* $j = 1, \ldots, T$*. Then when* $h \to 0$ *and* $n, T \to \infty$*,*

$$Var\Big\{\widehat{\beta_r^{(q)}}(t)|\mathcal{D}\Big\} = \frac{\omega^{rr}}{n}\Big\{\gamma_{q,q}(t,t) + \frac{2q!\gamma_{q,p+1}(t,t)h^{p-q+1}}{(p+1)!} B_{p+1}(K_{q,p+1})\Big\} + o_p\Big(\frac{h^{p-q+1}}{n}\Big),$$

*where* $B_{p+1}(\cdot)$ *is as defined in the Appendix before the proof of Lemma 1.*

The proof of Theorem 3 is straight forward by applying Lemma 3 in Fan and Zhang (2000), but with $\sigma^2(t) = 0$. This lemma is applicable because our $\gamma(s,t)$ satisfies the requirements of $\gamma_0(s,t)$ in their paper.

Furthermore, the next theorem gives asymptotic normality of $\widehat{\beta_r^{(q)}}(t)$. First, define $b = (b_1^{\mathrm{T}}, b_2^{\mathrm{T}}, \ldots, b_T^{\mathrm{T}})^{\mathrm{T}}$ and $\beta = (\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}}, \ldots, \beta_T^{\mathrm{T}})^{\mathrm{T}}$, to be the vectors of the raw estimators and the true coefficients across time. For $r \in \{1, \ldots, d\}$, define a $T \times dT$ matrix $P^{(r)}$, whose $\{k, (k-1)d+r\}$th elements for $k \in \{1, \ldots, T\}$ are equal to 1, and all other elements are equal to 0. The operator $P^{(r)}$ extracts the $r^{th}$ row of $b$ and $\beta$, i.e. $P^{(r)}b = \{b_r(t_1), \ldots, b_r(t_T)\}^{\mathrm{T}}$. Define $dT \times dT$ block

diagonal matrix $\bar{B} = \text{Diag}\{I_0(\beta_1)^{-1}, \ldots, I_0(\beta_T)^{-1}\}$ where $I_0(\beta_j)$ is the Fisher information matrix for $\beta_j$ unconditional on $\mathcal{D}$ for $j = 1, \ldots, T$, i.e.

$$I_0(\beta_j) = \text{E}\left\{\pi_{1j}(1-\pi_{1j})X_1(t_j)^{\text{T}}X_1(t_j)\right\}. \tag{8}$$

Further let $\Sigma_i$ be the matrix

$$\begin{bmatrix} X_{i1}X_{i1}^{\text{T}}\pi_{i1}(1-\pi_{i1}) & \cdots & \cdots \\ X_{i2}X_{i1}^{\text{T}}\sqrt{\pi_{i1}(1-\pi_{i1})}\sqrt{\pi_{i2}(1-\pi_{i2})}\gamma(t_1,t_2) & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ X_{iT}X_{i1}^{\text{T}}\sqrt{\pi_{i1}(1-\pi_{i1})}\sqrt{\pi_{iT}(1-\pi_{iT})}\gamma(t_1,t_T) & \cdots & X_{iT}X_{iT}^{\text{T}}\pi_{iT}(1-\pi_{iT}) \end{bmatrix}$$

and $\Sigma = \text{E}(\Sigma_i)$ with respect to $[X_{ij} = \{X_{i1}(t_j), \ldots, X_{id}(t_j)\}^{\text{T}},\ j = 1, \ldots, T]$. The matrix $\Sigma$ is well defined because under condition (A4), $\text{E}(\Sigma_i) = \text{E}(\Sigma_{i'})$.

**Theorem 4.** *Let conditions (A1)-(A4), (A6), (N1) and (N2) hold. Then conditional on $\mathcal{D}$, it holds that*

$$\sqrt{n}(b-\beta) \xrightarrow{d} \bar{B} * N(0, \Sigma),$$

*as $T$ is fixed and $n \to \infty$. For fixed $T$, let $\omega_T(t)$ be the vector of weight functions, $\omega_T(t) = \{\omega_{q,p+1}(t_1,t), \ldots, \omega_{q,p+1}(t_T,t)\}^{\text{T}}$ where $\widehat{\beta_r^{(q)}}(t) = \omega_T(t)P^{(r)}b$ by (4). Then it holds that*

$$\sqrt{n}\left\{\widehat{\beta_r^{(q)}}(t) - \omega_T(t)P^{(r)}\beta\right\} \xrightarrow{d} \omega_T(t)P^{(r)}\bar{B} * N(0, \Sigma),$$

*as $T$ is fixed and $n \to \infty$. Or equivalently,*

$$V_T^{-\frac{1}{2}}\sqrt{n}\left\{\widehat{\beta_r^{(q)}}(t) - \omega_T(t)P^{(r)}\beta\right\} \xrightarrow{d} N(0, I_T),$$

*as $n \to \infty$ for fixed $T$ where $V_T = \omega_T(t)P^{(r)}\bar{B}\Sigma\left\{\omega_T(t)P^{(r)}\bar{B}\right\}^{\text{T}}$.*

Theorem 4 shows that for any fixed $T$, the distribution of our final estimate $\widehat{\beta_r^{(q)}}(t)$ for $\beta_r^{(q)}(t)$ is approximately normal for sufficiently large $n$. However, to construct a confidence interval for $\beta_r^{(q)}(t)$, the difference between $\omega_T(t)P^{(r)}\beta$ and $\beta_r^{(q)}(t)$ must go to zero at a rate

faster than $(V_T/n)^{1/2}$, since

$$V_T^{-\frac{1}{2}}\sqrt{n}\left\{\widehat{\beta_r^{(q)}}(t)-\beta_r^{(q)}(t)\right\}=V_T^{-\frac{1}{2}}\sqrt{n}\left\{\widehat{\beta_r^{(q)}}(t)-\omega_T(t)P^{(r)}\beta\right\}+V_T^{-\frac{1}{2}}\sqrt{n}\left\{\omega_T(t)P^{(r)}\beta-\beta_r^{(q)}(t)\right\}.$$

The following proposition gives conditions under which this requirement is satisfied. For simplicity, we only consider the case $n_j = n$ for $j = 1, \ldots, T$.

**Proposition 1.** *Assume that the conditions in Theorem 4 hold and $\sqrt{n}h^{p-q+1}/T \to 0$, then*

$$V_T^{-\frac{1}{2}}\sqrt{n}\left\{\omega_T(t)P^{(r)}\beta-\beta_r^{(q)}(t)\right\}=o_p(1)I_T.$$

**Remark 3.** *As an example, lets consider the case $p = 1$ and $q = 0$, the local linear smoothing. It is easy to verify that if $h \propto T^{\epsilon-1}$ for $\epsilon \in (0, 1)$ and $n \propto T^{\delta}$ for $\delta \in (0, 6-4\epsilon)$, then $n \to \infty$, $h \to 0$, $Th \to \infty$ and $\sqrt{n}h^{p-q+1}/T \to 0$ as $T \to \infty$, which are needed for Theorem 4 and Proposition 1 to hold. For instance, if $\epsilon = 4/5$, then $h = O(T^{-1/5})$. In addition, $\delta$ should be between 0 and 2.8, which could be easily satisfied in practice since $n$ is usually much bigger than $T$.*

### 3.2 Statistical Inference: The Proposed Asymptotic Confidence Intervals and the Bootstrap Confidence Bands

In practice, the variance of $\widehat{\beta_r^{(q)}}(t)$ can be estimated using equation (6). $\mathrm{Cov}\{b(t_j), b(t_k)\}$ is estimated by the first term in the second and the third equations of (1) by replacing $W_j, W_k$ and $\gamma(t_j, t_k)$ with their estimates accordingly. Here we estimate $\gamma(t_j, t_k)$ by the Pearson's sample correlation, denoted by $\hat{\gamma}(t_j, t_k)$, with data $\{Y_i(t_j), Y_i(t_k)\}$ for all $i \in N_{jk}$. We estimate $W_j$ by $\widehat{W}_j = diag[\hat{\pi}_{i_1}(t_j)\{1-\hat{\pi}_{i_1}(t_j)\}, \ldots, \hat{\pi}_{i_{n_j}}(t_j)\{1-\hat{\pi}_{i_{n_j}}(t_j)\}]$, where $\hat{\pi}_{i_k}(t_j) = \{e^{X_{i_k}(t_j)^{\mathrm{T}}\widehat{\beta}(t_j)}\}/\{1+e^{X_{i_k}(t_j)^{\mathrm{T}}\widehat{\beta}(t_j)}\}$. Then $\hat{I}_j = \widetilde{X}_j^{\mathrm{T}}\widehat{W}_j\widetilde{X}_j$ and $\hat{I}_{jk} = \widetilde{X}_j^{\mathrm{T}}\widehat{W}_j^{\frac{1}{2}}M_{jk}\widehat{W}_k^{\frac{1}{2}}\widetilde{X}_k$. In (6), $\mathrm{Var}\{b_r(t_j)\}$ is estimated by the $(r,r)^{th}$ element of $\hat{I}_j^{-1}$, and $\mathrm{Cov}\{b_r(t_j), b_r(t_k)\}$ by the $(r,r)^{th}$ element of $\hat{\gamma}(t_j, t_k)\hat{I}_j^{-1}\hat{I}_{jk}\hat{I}_k^{-1}$. Finally, the variance estimator for $\beta_r^{(q)}(t)$ is given by

$$\widehat{\mathrm{Var}}\left\{\widehat{\beta_r^{(q)}}(t)\right\}=2\sum_{j<k}\omega_{q,p+1}(t_j,t)\omega_{q,p+1}(t_k,t)\widehat{\mathrm{Cov}}\{b_r(t_j),b_r(t_k)\}+\sum_{j=1}^{T}\omega_{q,p+1}^2(t_j,t)\widehat{\mathrm{Var}}\{b_r(t_j)\}. \tag{9}$$

The asymptotic results suggest that a 95% confidence interval of $\beta_r^{(q)}(t)$ be given by $\widehat{\beta_r^{(q)}}(t) \pm 1.96[\widehat{\text{Var}}\{\widehat{\beta_r^{(q)}}(t)\}]^{1/2}$, where the variance estimator is from (9).

Next we propose a global confidence band for the estimated curve $\widehat{\beta_r^{(q)}}(t), t \in [t_1, t_T]$ via bootstrap. We want to find two curves $L(t)$ and $U(t)$, $t \in [t_1, t_T]$, such that, in the nominal confidence level 0.95,

$$\text{P}\left\{L(t) \leq \beta_r^{(q)}(t) \leq U(t), t \in [t_1, t_T]\right\} = 0.95. \tag{10}$$

We consider a confidence band that is symmetric about the estimated curve. Therefore, $\{L(t), U(t)\} = \widehat{\beta_r^{(q)}}(t) \pm C_{0.95}[\widehat{\text{Var}}\{\widehat{\beta_r^{(q)}}(t)\}]^{1/2}$, where $C_{0.95}$ is an unknown constant that satisfies equation (10). With the confidence band taking the form above, equation (10) is equivalent to

$$\text{P}\left(\text{sup}_{t\in[t_1,t_T]} \frac{\mid \widehat{\beta_r^{(q)}}(t) - \beta_r^{(q)}(t) \mid}{\sqrt{\widehat{\text{Var}}\{\widehat{\beta_r^{(q)}}(t)\}}} < C_{0.95}\right) = 0.95.$$

We can estimate $C_{0.95}$ with a bootstrap $95^{th}$ percentile of the distribution of the supremum in the equation above. The algorithm is as following:

1. Resample the subjects with replacement from the original data, say $B$ times. For simplicity, the size of each resample is the same as the original data.

2. For the $k^{th}$ resample, $k \in 1, \ldots, B$, calculate the value

$$C^{(k)} = \text{sup}_{t\in[t_1,t_T]} \frac{\mid \widehat{\beta_r^{(q)}}^{(k)}(t) - \widehat{\beta_r^{(q)}}(t) \mid}{\sqrt{\widehat{\text{Var}}\{\widehat{\beta_r^{(q)}}^{(k)}(t)\}}},$$

   where the superscript $k$ indicates it is for the $k^{th}$ resample.

3. Estimate $C_{0.95}$ by the sample $95^{th}$ percentile of the $B$ values $C^{(k)}, k = 1, \ldots, B$, denoted by $\widehat{C_{0.95}}$.

Therefore, our bootstrap confidence band for $\beta_r^{(q)}(t), t \in [t_1, t_T]$ is given by $\widehat{\beta_r^{(q)}}(t) \pm \widehat{C_{0.95}}$ $[\widehat{\text{Var}}\{\widehat{\beta_r^{(q)}}(t)\}]^{1/2}$.

Finally, the bootstrap confidence band can be used to test hypotheses about $\beta_r(t)$. A typical null hypothesis is $H_0 : \beta_r^{(q)}(t) = f(t)$, for all $t \in [t_1, t_T]$, where $f(t)$ is a known function defined in the specific interval. When $f(t) \equiv 0$, we can test whether the $r^{th}$ covariate is insignificant throughout this interval, which in turn provides a way of variable selection in modeling. We reject the null hypothesis if the curve $f(t)$ is not completely inside the confidence band.

Another null hypothesis of interest is $H_0 : \beta_r^{(q)}(t) \equiv C^*$, for all $t \in [t_1, t_T]$, where $C^*$ is an unknown constant. With this null hypothesis, we can test whether the correlation of the $r^{th}$ covariate with the response variable is time-invariant, which in turn provides a way to simplify a fully nonparametric model into a semiparametric model, or even a fully parametric model. In the latter case, GEE or GLMM may be more efficient. We reject the null hypothesis if there does not exist a horizontal line completely inside the confidence band. Note that this test is expected to be conservative because the significance level is usually less than $\alpha$. The reason is clear from the testing procedure. When the null hypothesis is true, confidence bands at nominal confidence level 95% for the line $f(t) \equiv C^*$, for all $t \in [t_1, t_T]$, has a probability of 0.95 to cover $f(t)$. For those that do not cover $f(t)$, they may cover another constant line such as $f(t) + 0.01$. In this case, the test will still accept $H_0$. This results an acceptance rate higher than 0.95 for $H_0$, which implies that the significance level is less than 0.05.

## 4 Application to Smoking Cessation Data

In this section, we illustrate the proposed method using the smoking cessation data described in the Introduction. The main objective of this clinical trial is to evaluate and compare two behavioral methods, relapse prevention (RP) and contingency management (CM), alone and in combination, for optimizing smoking cessation outcomes using nicotine replacement therapy in methadone maintained cigarette smokers. All 175 participants received nicotine transdermal therapy and were randomly assigned to receive one of the four behavioral treatments (none,

RP, CM, RP+CM) for a period of 12 weeks. The participants were scheduled to visit back on every Monday, Wednesday and Friday. At every visit, measures were taken, including samples of breath (analyzed for carbon monoxide - CO reading) and urine, and weekly self-reported number of cigarettes smoked. Some participants didn't complete all the 36 visits, nevertheless many covariates were measured for each participant.

The dichotomous response variable of interest is smoking status determined from the CO reading, where smokers are coded as 1 (smoking status=1) and non-smokers as 0 (smoking status=0). The following subset of covariates are considered in our analysis: gender (2 categories), ethnicity (3 categories), treatment group assignments (4 categories), baseline CO reading, baseline urine opiate result (2 categories dirty or clean), baseline urine cocaine result (2 categories dirty or clean), baseline cotinine reading, age, number of cigarettes smoked per day, number of years smoked, depth of inhalation (3 categories), and number of times making serious attempt to quit. These covariates are all baseline measures, which means they are time-invariant. We treat categorical variables as class variables. That is, each category (except the reference level) has its own coefficient function. Among the 175 participants, only one subject is found to have a 0 (not at all) for the variable INHALE. It is modified to value 1 to reduce the categories to 3 for INHALE. The only two Asian subjects are dropped from the data to reduce the variable ETHNICITY to 3 categories. The rationale for these reductions in categories is that if a category has too few observations, the coefficient function corresponding to this category will have a sample size that is too small for a logistic regression model. This may result in an unstable raw estimator in the first step, and make the final estimator questionable. Hence, there are 17 coefficient functions to be estimated, including the intercept and all non-reference levels of the categorical variables. Using the notation of our model, we have $T = 36, n = 173, d = 17$ for this example. We utilize local linear regression as the smoothing method in step two where the bandwidths are selected visually by plotting the raw estimates

from step one separately for each varying coefficient function. The selected 17 bandwidths were between 12 and 17.

Figure 1 shows the percentage of nonmissing outcomes during each visit of the study. In the first 3 weeks, most individuals (over 90%) are observed at the scheduled visits. In the next several weeks, this percentage drops to about 70%. Figure 1 also descriptively illustrates the effect of behavioral treatments. It plots the percentage of smokers (smoking status=1) by the 4 treatment groups along the 36 time points. The CM-only and RP+CM groups are significantly below the reference group ("none"), by having almost no overlap. The RP-only group is also below the reference group, but they overlap during the middle of the 12 week period. RP+CM group is also slightly below CM-only group with some overlap. It can be seen that both treatments are helping, but CM is much more effective.

The refined estimators of the coefficient functions, along with their 95% bootstrap confidence bands and 95% point-wise confidence bands for all the covariates are presented in Figures 2-4. It is observed that the treatment effects of CM-only and CM+RP are significantly different from 0. In particular, the 95% bootstrap confidence band of the CM+RP treatment is almost completely below the zero line. This indicates a strong negative effect of the CM+RP treatment on the probability of being a smoker. The estimated curve for RP-only treatment is generally below the zero line, except in the middle. But the 95% bootstrap confidence band covers the entire zero line, indicating that it is not significant. These results are consistent with the findings of Shoptaw, Fuller, Yang, Frosch, Nahom, Jarvik, Rawson and Ling (2002) and visual findings from Figure 1.

The effect of the baseline CO reading is significant in the first 5 weeks of the study. This is likely because it is more difficult for heavier smokers entering the study to quit smoking, and this effect became weaker and weaker along time until there was no effect. All other covariates are non-significant since the 95% bootstrap confidence bands cover the entire zero line. Similar

to the baseline CO reading, the baseline cotinine reading also has a consistently positive effect, although it is not significant. Men have higher probability of being smokers than women, as the estimated curve is mostly above the zero line. There is no difference among the different ethnicities. Age has a slight negative effect. It may reflect a stronger mind to quit smoking among older participants. As expected, cigarettes per day reported at baseline positively predicts smoking. The effect has become stronger at the end of the study, which may indicate a relapse. Number of years smoked has a positive effect only for the second half of the study, also reflecting a relapse. It reflects the fact that it is harder to change long standing behavior patterns. The number of attempts to quit smoking has a negative effect on smoking status. People who are more committed to quit smoking by themselves are less likely to be smokers in the study. Inhaling deeply when smoking has a constant positive relationship on smoking status, compared to inhaling somewhat. Inhaling very deeply has no obvious relationship, possibly because of the small sample size in the group that inhales very deeply (24) compared to those inhaling deeply (110). The relationship between smoking status and clean urine opiate is positive, while the relationship to clean urine cocaine is negative. Intuitively, both should be negative. This result may be due to the collinearity between the two. The Pearson's sample correlation is 0.266 with p-value 0.0004.

Overall for the smoking cessation data, the proposed two-step method and the logistic varying coefficient modeling were very effective in describing the results. They not only confirm the finding of Shoptaw, Fuller, Yang, Frosch, Nahom, Jarvik, Rawson and Ling (2002) in a more general model, but also evaluate the effects of many other covariates and lead to intuitive interpretations. We are also able to study the change of effect along time, which distinguishes varying coefficient models from many others.

# 5 Simulation Studies

We conduct simulation studies to evaluate the finite sample performance of the proposed methodology including the TS estimation, asymptotic pointwise confidence intervals and the bootstrap confidence bands. We also include comparisons with LML, the parametric GEE method of Liang and Zeger (1986) and GLMM of Wolfinger and O'Connell (1993) with a random y-intercept. Smoothing in the TS is carried out via local linear regression. For component-wise bandwidth selection of the proposed TS method, we utilize the automatic rule-of-thumb bandwidth selector of Ruppert, Sheather and Wand (1995), separately for each varying coefficient function. LML maximizes the local likelihood and selects a single global bandwidth for all varying coefficient functions. We utilize leave-one-subject out cross-validation for selection of the global bandwidth similar to Cai, Fan and Li (2000). For more details on the LML method, we refer the readers to Şentürk, Dalrymple, Mohammed, Kaysen and Nguyen (2013) and Estes, Nguyen, Dalrymple, Mu and Şentürk (2014). While all four methods are used for comparisons via integrated mean squared error (IMSE), coverage of the point-wise asymptotic confidence intervals are compared for TS, GEE and GLMM.

## 5.1 Finite Sample Performance Comparisons

We utilize two simulation models. In model 1, there are 3 coefficient functions for the two covariates $X_1$, $X_2$, and the y-intercept. The covariate $X_1$ is a time-invariant discrete uniform variable taking on values in $\{0.5, 1, 1.5\}$. The covariate $X_2$ is generated from a Uniform$(0, 0.5)$ distribution. The sample size is 175 as in the smoking cessation data and results are reported based on 500 Monte Carlo runs. The times $\{t_j, j = 1, \ldots, 36\}$ are also from the smoking cessation data. We assume the correlation structure among repeated measurements to be AR-1(0.65), that is $\text{corr}\{Y_i(t_{j_1}), Y_i(t_{j_2})\} = 0.65^{|j_1 - j_2|}$. The algorithm described in Park, Park and Shin (1996) is adopted to generate correlated binary data. The varying coefficient functions

are $\beta_0(t) = 2\sin\{2\pi(t-1)/81\}$, $\beta_1(t) = \{\log_{10}(t) - 1\}/4$, and $\beta_2(t) = 1/(20t) - 2$. Model 2 differs from Model 1 only in the specification of the regression coefficient functions and the correlation structure. The coefficients are assumed to be time-invariant, $\beta_0 = 0.6, \beta_1 = 0.2$, $\beta_2 = -0.1$, and the correlation structure is AR-1(0.42).

The median of the selected bandwidths across the 500 Monte Carlo runs were (8.5, 15.0, 17.2) and (15.3, 15.3, 15.7) for $\{\beta_0(t), \beta_1(t), \beta_2(t)\}$ for the TS method in the two simulation models, respectively. The median of the selected global bandwidths for LML were 9 and 30 in the two simulation models, respectively. The results from the two models are reported in Tables 1 - 2 and in Figure 5-7. Figure 5 displays the true coefficient functions (solid gray) and their TS estimates (solid black) together with the proposed 95% bootstrap confidence bands (dashed black) from the sample with the median IMSE value among 500 Monte Carlo runs. Note that the true coefficient functions fall inside the bootstrap confidence bands, and that the automatic bandwidth selection may lead to under smoothing at times, as displayed for the estimation of $\beta_1(t)$. Nevertheless, the TS method, selecting different bandwidths for each coefficient function separately, is more effective in targeting varying coefficient functions of varying degrees of smoothness compared to the LML method with a global bandwidth. This can be observed in the estimated integrated mean square errors (IMSE) reported in Table 2. Since the median global bandwidth selected by LML is 9 in the first simulation model with coefficient functions of varying degrees of smoothness, LML performs better in estimation of $\beta_0(t)$ which requires a lower bandwidth, but undersmooths $\beta_1(t)$ and $\beta_2(t)$, leading to higher mean IMSE values, compared to the TS method. Note also that when the covariate effects change over time (Model 1), GLMM and GEE have a much larger mean IMSE, compared to TS and LML, due to modeling bias. It can be 26 times as big as the IMSE from TS. Figure 6 also plots the mean estimated mean square error (MSE) over time in Model 1. Estimated MSE from GEE and GLMM is much higher than those from TS and LML except for $\beta_1(t)$. As

expected, GEE and GLMM perform much better than TS and LML when the assumption for GEE and GLMM models for constant covariate effects is met in simulation Model 2 (Figure 7). This is because parametric modeling is much more efficient under constant covariate effects.

Table 1 compares the coverage probabilities of point-wise confidence intervals at nominal levels 95% and 90% for the two simulation models at eight time points from the total 36 points. It is observed that the coverage probability of the proposed TS method is reasonably close to the nominal level for both models. In contrast, the coverage probability of GLMM and GEE can be very low ($< 5\%$) when the true regression coefficients change over time (Model 1). Even for Model 2 where the covariate effects do not change over time, the coverage probability of GLMM for $\beta_2(t)$ is still lower than the nominal level. Hence the proposed method provides an efficient tool to check whether the varying coefficient functions are constants and can accommodate varying coefficient functions of varying degrees of smoothness. For cases where covariate effects are constant, the parametric modeling approaches are preferred.

### 5.2 Performance of the Proposed Bootstrap Confidence Bands

We conduct further simulations to study the performance of the bootstrap confidence bands described in Section 3.2. Results are shown in Tables 3 and 4. While Table 3 reports on coverage rates of the proposed bootstrap confidence bands, Tables 4 reports results from a hypotheses testing setup, utilizing the relationship between hypotheses testing and confidence bands (or confidence interval in non-functional situations). Results are reported from 200 Monte Carlo runs where each run is based on 500 bootstrap samples at sample size $n = 175$. Component-wise bandwidths are selected based on the automatic rule-of-thumb bandwidth selection of Ruppert, Sheather and Wand (1995) in each Monte Carlo run and fits to bootstrap samples utilize the same bandwidths as those selected for the Monte Carlo runs. We use two settings where the first setting is the same as Model 1 described above and the second setting differs from Model 1 by utilizing time-invariant coefficient functions, $\beta_0(t) = -1$, $\beta_1(t) = 0$

and $\beta_2(t) = 2$.

The coverage rates reported in Table 3 are pretty close to the nominal levels in both settings, where $\beta_2(t)$ is less covered than $\beta_0(t)$ and $\beta_1(t)$. This may be due to the fact that $\beta_2(t)$, being the most smooth function of the three, may be under smoothed in some runs because of the under smoothing tendency of the automatic bandwidth selectors. Table 4 gives the estimated rejection proportions (in %) for two hypotheses tests: 1. $H_0(a) : \beta_r(t)$ does not change over time; 2. $H_0(b) : \beta_r(t) = 0$, for all $t \in [t_0, t_T]$. The testing procedure is based on the proposed bootstrap confidence bands. In the first setting, the powers for rejecting $H_0(a)$ and $H_0(b)$ are satisfying for $\beta_0(t)$ and $\beta_2(t)$ where they are all at 100%. The powers for $\beta_1(t)$ are much smaller than those for the other two coefficient functions. This is because $\beta_1(t)$ is much more similar to a constant function, more specifically a constant function at 0. Note also that the powers for rejecting $H_0(a)$ are consistently smaller than those for rejecting $H_0(b)$, since $H_0(b)$ is a special case for $H_0(a)$. For the second setting, reported proportions for $H_0(a)$ at all varying coefficient functions and for $H_0(b)$ at $\beta_1(t)$ are estimated significance levels since the null hypotheses are true in these cases. For $H_0(b)$, while the significance levels for $\beta_1(t)$ are close to the nominal levels, the reported values for the other two coefficient functions show that the powers are 1 for rejecting $H_0(b)$ when the constants are other than 0. For $H_0(a)$, the estimated significance levels are consistently less than the nominal level as discussed in Section 3.2. These findings imply that the proposed bootstrap confidence bands are very effective in identifying whether $H_0(a)$ is true and the unknown constant.

## 6 Discussion

In this paper, we proposed a TS estimation procedure for logistic varying coefficient modeling of longitudinal binary data. The basic idea behind the proposal as well as its implementation are simple. We also evaluated the asymptotic properties of the proposed estimators and found

them to be asymptotically unbiased. We established the asymptotic variance under two specific situations and proved that the estimators are asymptotically normal, leading to the proposed asymptotic and finite sample inference procedures. We applied the proposed methodology to smoking cessation data. The main results are consistent with findings from previous studies. Moreover, we evaluated many other covariates and have provided reasonable interpretations of the results. The estimators give intuitively consistent inferences and the bootstrap confidence intervals are effective in identifying significant predictors.

Simulation studies indicate that TS and LML perform better than GEE and GLMM models when their parametric assumptions do not hold. Unlike LML, TS is able to target coefficient functions with varying degrees of smoothness, via component-wise bandwidth selections. In addition, TS also allows for visual selection of component-wise bandwidths via plotting of the raw varying coefficient function estimates. When the underlying model reduces to a parametric form with time-invariant coefficient functions, parametric models GEE and GLMM lead to more efficient estimation as expected. The efficacy of the proposed bootstrap confidence bands are shown via simulation studies where the implied tests have very high power in many cases. While the first hypothesis of constant coefficient functions tests whether the logistic varying coefficient model reduces to a semi-parametric or a parametric model, the second hypothesis of coefficient functions being equal to zero, allows us to perform model selection.

The proposed methodology can easily be extended to be applicable to other forms of longitudinal data. For example longitudinal categorical data can be modeled in a similar way, as long as an appropriate marginal model (e.g. the *proportional odds* model of Agresti (2002)) is selected for cross-sectional modeling in the first step. A second extension can be to spatial correlated longitudinal data, such as that encountered in progression detection of glaucoma in the visual field (Gardiner and Crabb (2002)). Spatial correlation can be taken into account in the proposed TS method by applying a higher dimensional smoothing procedure in the second

step.

## Acknowledgements

This publication was made possible by National Institute of Health grants CA016042 (GL), 8UL1TR000124-02 (GL), 1P01CA163200-01A1 (GL) CA78314-03 (GL) and the National Institute of Diabetes and Digestive and Kidney Diseases grant R01 DK092232 (DS). The authors thank Professor Jianqing Fan for his helpful discussion and Professor Xiaoyan Wei for providing the smoking cessation data.

## Appendix: Proofs

The following technical conditions are needed.

(A1) The time points $t_1, t_2, \ldots, t_T$ are a random sample from a probability density $f$ and $t$ is a continuous point of $f$ in the interior of the support of $f$.

(A2) The function $\beta_r(t)$ is $(p+1)$-times continuously differentiable for some $p$.

(A3) The kernel function K is a bounded symmetric probability density function with a bounded support.

(A4) The covariates $X_i(t_j), i = 1, \ldots, n$ are independently and identically distributed as $X_1(t_j)$ with $\mathrm{E}\{X_1(t_j)X_1(t_j)^{\mathrm{T}}\}$ positive definite for $j = 1, \ldots, T$.

(A5) $h \to 0$ and $Th \to \infty$ as $T \to \infty$.

(A6) $\min\{n_1, n_2, \ldots, n_T\} \to \infty$ as $n \to \infty$, while $T$ is fixed or $T \to \infty$.

(A7) The covariates $X_i(t_j)$ satisfy condition (A4) and they are time-invariant. That is, $X_i(t_j) = X_i(t_1)$ for all $j = 1, \ldots, T$.

(A8) All the true coefficient functions are time-invariant. That is, $\beta_r(t) = \beta_r$ for all $r = 1, \ldots, d$ and $t \in [\,0, D]$.

We define further notations. Let $C_j = \{1, t_j - t, \ldots, (t_j - t)^p\}^{\mathrm{T}}$, $j = 1, 2, \ldots, T$ and $K_h(t) = K(t/h)/h$ be a kernel function with a bandwidth $h$. Let $C = (C_1, C_2, \ldots, C_T)$ and $W = \mathrm{diag}(W_1, \ldots, W_T)$ with $W_j = K_h(t_j - t)$. Then the weights in (4) are defined as $\omega_{q,p+1}(t_j, t) = q! e_{q+1,p+1}^{\mathrm{T}} (C^{\mathrm{T}} W C)^{-1} C_j W_j$, $j = 1, 2, \ldots, T$, where $e_{q+1,p+1}$ denotes a $(p+1)$-dimensional unit vector with one at its $(q+1)^{th}$ entry, and zero elsewhere. More specifically, the local linear weights are given by $\omega_{0,2}(t_j, t)$, $j = 1, 2, \ldots, T$ with $q = 0$ and $p = 1$. Let $K_{q,p+1}$ be the equivalent kernel of $\omega_{q,p+1}$, which is defined by $K_{q,p+1}(t) = e_{q+1,p+1}^{\mathrm{T}} S^{-1} (1, t, \ldots, t^p)^{\mathrm{T}} K(t)$, where $S = (s_{ij})$, $i, j = 0, 1, \ldots, p$, and $s_{ij} = \int K(u) u^{i+j} du$. Recall that $K(t)$ is the original kernel function. Furthermore, define $B_{p+1}(K) = \int K(u) u^{p+1} du$, and $V(K) = \int K^2(u) du$.

**Proof of Lemma 1:** For $t_j \in A$, let $\beta_j = \beta(t_j)$ and $b_j = b(t_j)$. Let $l(\theta)$ be the log-likelihood defined for the logistic regression at $t_j$. Refer to McCullagh and Nelder (1989) for details. Here $\theta$ is the parameter vector of interest in the logistic model. Therefore, the true value of $\theta$ is $\beta_j$, and it is estimated by $b_j$. The first part of the Lemma on asymptotic bias follows from equation (4.18) of McCullagh and Nelder (1989). Refer to Ferguson (1996) (page 119) for part of the deduction below. First, expand $\dot{l}(\theta)$ at $\beta_j$ as $\dot{l}(\theta) = \dot{l}(\beta_j) + \int_0^1 \ddot{l}\{\beta_j + \lambda(\theta - \beta_j)\} \mathrm{d}\lambda (\theta - \beta_j)$. Now let $\theta = b_j$. Because $b_j$ is the MLE of $\beta_j$, it is a strongly consistent sequence satisfying $\dot{l}(b_j) = 0$. Hence $\dot{l}(\beta_j) = n_j B_{n_j} (b_j - \beta_j)$, where $B_{n_j} = -\int_0^1 (1/n_j) \ddot{l}\{\beta_j + \lambda(b_j - \beta_j)\} \mathrm{d}\lambda$. Recall the Fisher information for this logistic regression is $I_j = \widetilde{X}_j^{\mathrm{T}} W_j \widetilde{X_j}$ and note that $\dot{l}(\beta_j) - I_j(b_j - \beta_j) - (n_j B_{n_j} - I_j)(b_j - \beta_j) = 0$. This implies that

$$b_j - \beta_j = I_j^{-1}\{\dot{l}(\beta_j) - (n_j B_{n_j} - I_j)(b_j - \beta_j)\} = \sqrt{n_j} I_j^{-1} \left\{ \frac{1}{\sqrt{n_j}} \dot{l}(\beta_j) - (B_{n_j} - \frac{1}{n_j} I_j) \sqrt{n_j} (b_j - \beta_j) \right\}. \tag{A.1}$$

Note that under condition (A4), $I_j = \widetilde{X}_j W_j^{\mathrm{T}} \widetilde{X_j} = \sum_{i=1}^{n_j} \pi_{ij}(1-\pi_{ij}) X_i(t_j)^{\mathrm{T}} X_i(t_j) \sim O(n_j)$. Conditional on $\mathcal{D}$, and under assumptions (N1) and (N2), the normed $b_j$ is asymptotically normal, i.e. $I_j^{1/2}(b_j - \beta_j) \xrightarrow{d} N(0, I)$ as $n_j \to \infty$. We refer readers to Gourieroux and Monfort (1981), where the result is shown in the proof of Proposition 4. Define the first term in equation (A.1) as $A_{n_j}$. By condition (A4) and the Strong Law of Large Numbers (SLLN), we have $A_{n_j} = \dot{l}(\beta_j)/\sqrt{n_j} = O_p(1)$. Also $\mathrm{E}(A_{n_j}) = \mathrm{E}\{\dot{l}(\beta_j)\}/\sqrt{n_j} = 0$. Define the second term in equation (A.1) as $C_{n_j}$. From the part (3) of the proof for Theorem (4) given below, we have $C_{n_j}|\mathcal{D} \xrightarrow{d} 0$, or $C_{n_j} = o_p(1)$. Then $b_j - \beta_j = \sqrt{n_j} I_j^{-1}\left(A_{n_j} - C_{n_j}\right)$, and

$$
\begin{aligned}
\mathrm{Cov}(b_j|\mathcal{D}) &= \mathrm{E}[\{b_j - \mathrm{E}(b_j)\}\{b_j - \mathrm{E}(b_j)\}^{\mathrm{T}}] \\
&= \mathrm{E}\left\{(b_j - \beta_j) - \mathrm{E}(b_j - \beta_j)\right\}\left\{(b_j - \beta_j) - \mathrm{E}(b_j - \beta_j)\right\}^{\mathrm{T}} \\
&= \mathrm{E}\left\{(b_j - \beta_j)(b_j - \beta_j)^{\mathrm{T}}\right\} - \mathrm{E}(b_j - \beta_j)\mathrm{E}(b_j - \beta_j)^{\mathrm{T}} \\
&= \mathrm{E}\left\{n_j I_j^{-1}(A_{n_j} - C_{n_j})(A_{n_j} - C_{n_j})^{\mathrm{T}} I_j^{-1}\right\} + o\left(\frac{1}{n_j}\right) o\left(\frac{1}{n_j}\right)^{\mathrm{T}} \\
&= n_j I_j^{-1}\mathrm{E}(A_{n_j} A_{n_j}^{\mathrm{T}}) I_j^{-1} - n_j I_j^{-1}\mathrm{E}(A_{n_j} C_{n_j}^{\mathrm{T}} + C_{n_j} A_{n_j}^{\mathrm{T}}) I_j^{-1} \\
&\quad + n_j I_j^{-1}\mathrm{E}(C_{n_j} C_{n_j}^{\mathrm{T}}) I_j^{-1} + o\left(\frac{1}{n_j^2}\right).
\end{aligned}
$$

This, combined with $A_{n_j} = O_p(1)$, $C_{n_j} = o_p(1)$ and the following result from McCullagh and Nelder (1989), $A_{n_j} = \dot{l}(\beta_j)/\sqrt{n_j} = \widetilde{X}_j^{\mathrm{T}}\{\widetilde{Y}_j - \mathrm{E}(\widetilde{Y}_j)\}/\sqrt{n_j}$, implies that

$$
\begin{aligned}
\mathrm{Cov}(b_j|\mathcal{D}) &= I_j^{-1}\widetilde{X}_j^{\mathrm{T}}\mathrm{E}[\{\widetilde{Y}_j - \mathrm{E}(\widetilde{Y}_j)\}\{\widetilde{Y}_j - \mathrm{E}(\widetilde{Y}_j)\}^{\mathrm{T}}]\widetilde{X}_j I_j^{-1}\{1 + o(1)\} \\
&= I_j^{-1}\widetilde{X}_j^{\mathrm{T}} W_j \widetilde{X}_j I_j^{-1}\{1 + o(1)\} = I_j^{-1}\{1 + o(1)\}, \\
\text{and} \quad \mathrm{Cov}(b_j, b_k|\mathcal{D}) &= \mathrm{E}[\{(b_j - \beta_j) - \mathrm{E}(b_j - \beta_j)\}\{(b_k - \beta_k) - \mathrm{E}(b_k - \beta_k)\}^{\mathrm{T}}] \\
&= \mathrm{E}(b_j - \beta_j)(b_k - \beta_k)^{\mathrm{T}} - \mathrm{E}(b_j - \beta_j)\mathrm{E}(b_k - \beta_k)^{\mathrm{T}} \\
&= \mathrm{E}\left\{\sqrt{n_j}\sqrt{n_k} I_j^{-1}(A_{n_j} - C_{n_j})(A_{n_k} - C_{n_k})^{\mathrm{T}} I_k^{-1}\right\} - o(n_j^{-1}) o(n_k^{-1}) \\
&= \sqrt{n_j}\sqrt{n_k} I_j^{-1}\mathrm{E}(A_{n_j} A_{n_k}^{\mathrm{T}}) I_k^{-1} - \sqrt{n_j}\sqrt{n_k} I_j^{-1}\mathrm{E}(A_{n_j} C_{n_k}^{\mathrm{T}} + C_{n_j} A_{n_k}^{\mathrm{T}}) I_k^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&+\sqrt{n_j}\sqrt{n_k}I_j^{-1}\mathrm{E}(C_{n_j}C_{n_k}^{\mathrm{T}})I_k^{-1} - o\{(n_jn_k)^{-1}\} \\
= \;& I_j^{-1}\widetilde{X}_j^{\mathrm{T}}\mathrm{E}\left[\{\widetilde{Y}_j - \mathrm{E}(\widetilde{Y}_j)\}\{\widetilde{Y}_k - \mathrm{E}(\widetilde{Y}_j)\}^{\mathrm{T}}\right]\widetilde{X}_kI_k^{-1} - \sqrt{n_j}\sqrt{n_k}I_j^{-1}\mathrm{E}\{O_p(1)o_p(1)\}I_k^{-1} \\
&+\sqrt{n_j}\sqrt{n_k}I_j^{-1}\mathrm{E}\{o_p(1)o_p(1)\}I_k^{-1} - o\{(n_jn_k)^{-1}\} \\
= \;& I_j^{-1}\widetilde{X}_j^{\mathrm{T}}W_j^{\frac{1}{2}}M_{jk}W_k^{\frac{1}{2}}\widetilde{X}_kI_k^{-1}\gamma(t_j,t_k) + o\{(n_jn_k)^{-1}\} = I_j^{-1}I_{jk}I_k^{-1}\gamma(t_j,t_k)\{1+o(1)\}.
\end{aligned}
$$

This completes the proof.

**Proof of Theorem 1:** Suppose the conditions of the theorem hold. Then

$$
\begin{aligned}
&\mathrm{E}\left\{\hat{\beta}_r^{(q)}(t)|\mathcal{D}\right\} \\
=&\sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)\mathrm{E}\left\{b_r(t_j)\right\} = \sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)\left\{\beta_r(t_j)+O(1/n_j)\right\} \\
=&\sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)\beta_r(t_j) + \left\{\sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)\right\}O(1/n_\wedge) \\
=&\sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)\left[\sum_{k=0}^{p+1}\beta_r^{(k)}(t)\frac{(t_j-t)^k}{k!} + o\left\{(t_j-t)^{p+1}\right\}\right] + O(1/n_\wedge) \\
=&\sum_{k=0}^{p+1}\left\{\frac{\beta_r^{(k)}(t)}{k!}\sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)(t_j-t)^k\right\} + \sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)o\left\{(t_j-t)^{p+1}\right\} + O(1/n_\wedge) \\
=&\beta_r^{(q)}(t) + \left\{\frac{1}{(p+1)!}\beta_r^{(p+1)}(t) + o_p(1)\right\}\sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)(t_j-t)^{p+1} + O(1/n_\wedge) \\
=&\beta_r^{(q)}(t) + \frac{q!\beta_r^{(p+1)}(t)h^{p-q+1}}{(p+1)!}B_{p+1}(K_{q,p+1})\left\{1+o_p(1)\right\} + O(1/n_\wedge),
\end{aligned}
$$

where we used Lemma 2 of Fan and Zhang (2000) in the third and the last two equalities. The conclusion of the Theorem follows immediately.

**Proof of Theorem 4:** Under mild conditions, the MLE $b_j$ of $\beta_j$ exists and is strongly consistent. We will show the asymptotic normality of the vector $(b-\beta)$ as $n_\wedge \to \infty$. Without loss of generality, let's consider a simple case: $n_j = n, j = 1,\ldots,T$. From the proof of Lemma 1, we have $b_j - \beta_j = \sqrt{n_j}I_j^{-1}\left(A_{n_j} - C_{n_j}\right)$. Then we can write the vector $(b-\beta)$ as

$$\begin{bmatrix} b_1-\beta_1 \\ b_2-\beta_2 \\ \vdots \\ b_T-\beta_T \end{bmatrix} = \mathrm{Diag}\left(\sqrt{n}I_1^{-1},\ldots,\sqrt{n}I_T^{-1}\right) \begin{bmatrix} A_{n_1}-C_{n_1} \\ A_{n_2}-C_{n_2} \\ \vdots \\ A_{n_T}-C_{n_T} \end{bmatrix} = B_{(n)}(A_{(n)}-C_{(n)}),$$

where $A_{(n)} = (A_{n_1},\ldots,A_{n_T})^{\mathrm{T}}$, $B_{(n)} = \mathrm{Diag}\left(\sqrt{n}I_1^{-1},\ldots,\sqrt{n}I_T^{-1}\right)$ and $C_{(n)} = (C_{n_1},\ldots,C_{n_T})^{\mathrm{T}}$. In the following, we would like to prove that, conditional on $\mathcal{D}$, $\sqrt{n}(b-\beta) = \sqrt{n}B_{(n)}(A_{(n)}-C_{(n)})$ is asymptotically normal when $n \to \infty$ and $T$ is fixed.

1. From the notations above, we have $\sqrt{n}B_{(n)} = n * \mathrm{Diag}\left(I_1^{-1},\ldots,I_T^{-1}\right)$ and its inverse $(\sqrt{n}B_{(n)})^{-1} = \mathrm{Diag}\left\{I_1,\ldots,I_T\right\}/n$. Under the condition (A4) and by SLLN,

$$\begin{aligned} \frac{1}{n}I_j &= \frac{1}{n}\widetilde{X}_j^{\mathrm{T}}W_j\widetilde{X}_j = \frac{1}{n}\sum_{i=1}^{n}\pi_{ij}(1-\pi_{ij})X_i(t_j)^{\mathrm{T}}X_i(t_j) \\ &\xrightarrow{a.s.} \mathrm{E}\left\{\pi_{1j}(1-\pi_{1j})X_1(t_j)^{\mathrm{T}}X_1(t_j)\right\} = I_0(\beta_j). \end{aligned}$$

   Therefore, with probability one, $n^{-1}I_j|\mathcal{D} \to I_0(\beta_j)$. And with probability one, $\sqrt{n}B_{(n)}|\mathcal{D} \to \bar{B}$, a constant matrix.

2. $$A_{(n)} = \begin{bmatrix} A_{n_1} \\ A_{n_2} \\ \vdots \\ A_{n_T} \end{bmatrix} = \frac{1}{\sqrt{n}} \begin{bmatrix} \dot{l}(\beta_1) \\ \dot{l}(\beta_2) \\ \vdots \\ \dot{l}(\beta_T) \end{bmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \begin{bmatrix} X_i(t_1)\{Y_i(t_1)-\pi_{i1}\} \\ X_i(t_2)\{Y_i(t_2)-\pi_{i2}\} \\ \vdots \\ X_i(t_T)\{Y_i(t_T)-\pi_{iT}\} \end{bmatrix} = \sum_{i=1}^{n} Z_i.$$

   To prove that $A_{(n)}|\mathcal{D} \xrightarrow{d} N(0,\Sigma)$, we need to show that the following Lindeberg conditions hold. Conditional on $\mathcal{D}$, $\sum_{i=1}^{n}\mathrm{E}\|Z_i\|^2 1\{\|Z_i\| > \epsilon\} \to 0$, every $\epsilon > 0$, and $\sum_{i=1}^{n}\mathrm{Cov}(Z_i) \to \Sigma$.

   *Proof:* $\|Z_i\| = [\sum_{j=1}^{T}\|X_i(t_j)\|^2\{Y_i(t_j)-\pi_{ij}\}^2]^{1/2}/\sqrt{n}$. Conditional on $\mathcal{D}$, and for all $\epsilon > 0$,

$$\sum_{i=1}^{n}\mathrm{E}\|Z_i\|^2 1\{\|Z_i\| > \epsilon\} \leqslant \sum_{i=1}^{n}\mathrm{E}\frac{\|Z_i\|^{2+\delta}}{\epsilon^{\delta}} = \frac{1}{n^{1+\frac{\delta}{2}}\epsilon^{\delta}}\sum_{i=1}^{n}\mathrm{E}\left[\sum_{j=1}^{T}\|X_i(t_j)\|^2\{Y_i(t_j)-\pi_{ij}\}^2\right]^{\frac{2+\delta}{2}}.$$

Now let $h\{X_i(t_1),\ldots,X_i(t_T)\} = \mathrm{E}[\sum_{j=1}^{T}\|X_i(t_j)\|^2\{Y_i(t_j)-\pi_{ij}\}^2|\mathcal{D}]^{(2+\delta)/2}$. Since $|Y_i(t_j)-\pi_{ij}|\leq 1$, and by the assumption (N1),

$$Eh\{X_i(t_1),\ldots,X_i(t_T)\} \leq E\left(\sum_{j=1}^{T}\|X_i(t_j)\|^2\right)^{\frac{2+\delta}{2}} \leq \left(T\times d\times M_0^2\right)^{\frac{2+\delta}{2}} < \infty.$$

By the condition (A4) and SLLN, $[\sum_{i=1}^{n} h\{X_i(t_1),\ldots,X_i(t_T)\}]/n \xrightarrow{a.s.} \mathrm{E}h\{X_i(t_1),\ldots,X_i(t_T)\}$. That is, with probability one and conditional on $\mathcal{D}$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} h\{X_i(t_1),\ldots,X_i(t_T)\} &\rightarrow \mathrm{E}h\{X_i(t_1),\ldots,X_i(t_T)\},\\
\sum_{i=1}^{n}\mathrm{E}\|Z_i\|^2 1\{\|Z_i\|>\epsilon\} &= \frac{1}{n^{\frac{\delta}{2}}\epsilon^{\delta}} * \frac{1}{n}\sum_{i=1}^{n} h\{X_i(t_1),\ldots,X_i(t_T)\} \rightarrow 0,\\
\text{and}\quad \mathrm{Cov}Z_i &= \frac{1}{n}\mathrm{Cov}\begin{bmatrix} X_i(t_1)\{Y_i(t_1)-\pi_{i1}\}\\ X_i(t_2)\{Y_i(t_2)-\pi_{i2}\}\\ \vdots\\ X_i(t_T)(Y_i(t_T)-\pi_{iT}\}\end{bmatrix} = \frac{1}{n}\Sigma_i.
\end{aligned}$$

By the condition (A4) and SLLN, $\sum_{i=1}^{n}\mathrm{Cov}Z_i = (\sum_{i=1}^{n}\Sigma_i)/n \xrightarrow{a.s.} \Sigma$. With probability one and conditional on $\mathcal{D}$, $\sum_{i=1}^{n}\mathrm{Cov}Z_i \rightarrow \Sigma$. It is obvious that $\mathrm{E}(Z_i|\mathcal{D}) = 0$. Therefore, the multivariate Lindeberg-Feller Central Limit Theorem from Van der Vaart (1989) applies. We have shown that $A_{(n)}|\mathcal{D}$ is asymptotically normal with distribution $\mathrm{N}(0,\Sigma)$.

3. In this part, we want to prove $C_{(n)}|\mathcal{D} \xrightarrow{d} 0$, where

$$C_{(n)} = \begin{bmatrix} (B_{n1}-I_1/n_1)\sqrt{n_1}(b_1-\beta_1)\\ (B_{n2}-I_2/n_2)\sqrt{n_2}(b_2-\beta_2)\\ \vdots\\ (B_{nT}-I_T/n_T)\sqrt{n_T}(b_T-\beta_T)\end{bmatrix}.$$

Let $C_{nj} = (B_{nj}-I_j/n_j)\sqrt{n_j}(b_j-\beta_j)$.

(a) By Ferguson (1996), $B_{n_j} \xrightarrow{a.s.} I_0(\beta_j)$. This is also the same matrix as in part 1 of this proof: $I_0(\beta_j) = \mathrm{E}\left\{\pi_{1j}(1-\pi_{1j})X_1(t_j)^{\mathrm{T}}X_1(t_j)\right\}$. Therefore, $B_{n_j} \xrightarrow{a.s.} I_0(\beta_j)$ and $I_j/n \xrightarrow{a.s.} I_0(\beta_j)$ together imply that $B_{n_j} - I_j/n \xrightarrow{a.s.} 0$ and $B_{n_j} - n^{-1}I_j|\mathcal{D} \xrightarrow{a.s.} 0$.

(b) $\sqrt{n_j}(b_j - \beta_j) = \sqrt{n_j} I_j^{-1/2} * I_j^{1/2}(b_j - \beta_j)$. We have $I_j^{1/2}(b_j - \beta_j)|\mathcal{D} \xrightarrow{d} N(0, I)$ from the proof of Lemma 1. Also we have $n^{-1} I_j|\mathcal{D} \to I_0(\beta_j)$ from (a). Therefore, $\sqrt{n_j}(b_j - \beta_j)|\mathcal{D} \xrightarrow{d} N\{0, I_0(\beta_j)\}$.

Combined the results above, we have $C_{(n)}|\mathcal{D} \xrightarrow{d} 0$.

Therefore $A_{(n)} - C_{(n)}|\mathcal{D} \xrightarrow{d} N(0, \Sigma)$. Conditional on $\mathcal{D}$, $\sqrt{n}(b-\beta) = \sqrt{n} B_{(n)} * (A_{(n)} - C_{(n)}) \xrightarrow{d} \bar{B} * N(0, \Sigma)$ as $n \to \infty$. This means $b$ is asymptotic multivariate normal as $n \to \infty$. Note that the smoothing coefficients $\{\omega_{q,p+1}(t_j, t), j = 1, 2, \ldots, T\}$ only depend on $t, \{t_1, \ldots, t_T\}$ and the specification of kernel function $K$ and bandwidth $h$. When $T$ and $\{t_1, \ldots, t_T\}$ are fixed, they don't change as $n \to \infty$. Therefore, our linear smoother (linear combination of the raw estimates $b_{1r}, \ldots, b_{Tr}$ for the $r^{th}$ component of $\beta_t$) is asymptotically normal. Explicitly,

$$\sqrt{n}\left\{\widehat{\beta_r^{(q)}}(t) - \omega_T(t) P^{(r)} \beta\right\} = \omega_T(t) P^{(r)} \sqrt{n}(b - \beta) \xrightarrow{d} \omega_T(t) P^{(r)} \bar{B} * N(0, \Sigma),$$

as $n \to \infty$. This completes the proof.

**Proof of Proposition 1:** To prove the proposition, we first need to study the order of $V_T$. Define $I_0(t_j, t_k) = \mathrm{E}\{\sqrt{\pi_{1j}(1-\pi_{1j})}\sqrt{\pi_{1k}(1-\pi_{1k})} X_1(t_j)^{\mathrm{T}} X_1(t_k)\}$, where the expected value is taken with respected to the predictors $X_i(t_j)'s$. More specifically, $I_0(t_j, t_j) = \mathrm{E}\{\pi_{1j}(1 - \pi_{1j}) X_1(t_j)^{\mathrm{T}} X_1(t_j)\}$ is the Fisher information matrix $I_0(\beta_j)$ defined in (8). Also, $I_0(t_j, t_k) = I_0(t_k, t_j)$. With these notations, the matrix $\Sigma$ can be written as

$$\Sigma = \begin{bmatrix} I_0(\beta_1) & I_0(t_1, t_2) & \cdots & \cdots \\ I_0(t_2, t_1) & I_0(\beta_2) & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ I_0(t_T, t_1) & I_0(t_T, t_2) & \cdots & I_0(\beta_T) \end{bmatrix}.$$

Thus

$$\bar{B}\Sigma\bar{B}^{\mathrm{T}} = \begin{bmatrix} I_0(\beta_1)^{-1} & I_0(\beta_1)^{-1} I_0(t_1, t_2) I_0(\beta_2)^{-1} & \cdots & \cdots \\ I_0(\beta_2)^{-1} I_0(t_2, t_1) I_0(\beta_1)^{-1} & I_0(\beta_2)^{-1} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ I_0(\beta_T)^{-1} I_0(t_T, t_1) I_0(\beta_1)^{-1} & I_0(\beta_T)^{-1} I_0(t_T, t_2) I_0(\beta_2)^{-1} & \cdots & I_0(\beta_T) \end{bmatrix},$$

$$P^{(r)}\bar{B}\Sigma\bar{B}^{\mathrm{T}}P^{(r)^{\mathrm{T}}} = \begin{bmatrix} \{I_0(\beta_1)^{-1}\}^{(rr)} & \cdots & \cdots & \cdots \\ \{I_0(\beta_2)^{-1}I_0(t_2,t_1)I_0(\beta_1)^{-1}\}^{(rr)} & \{I_0(\beta_2)^{-1}\}^{(rr)} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \{I_0(\beta_T)^{-1}I_0(t_T,t_1)I_0(\beta_1)^{-1}\}^{(rr)} & \cdots & \cdots & \{I_0(\beta_T)\}^{(rr)} \end{bmatrix}.$$

Therefore $V_T = \omega_T(t)P^{(r)}\bar{B}\Sigma\bar{B}^{\mathrm{T}}P^{(r)^{\mathrm{T}}}\omega_T(t)^{\mathrm{T}} = \sum_{j\neq k}\omega_{q,p+1}(t_j,t)\omega_{q,p+1}(t_k,t)\{I_0(\beta_j)^{-1}I_0(t_j,t_k)$ $I_0(\beta_k)^{-1}\}^{(rr)} + \sum_{j=1}^{T}\omega_{q,p+1}^2(t_j,t)\{I_0(\beta_j)^{-1}\}^{(rr)} = V_T(1) + V_T(2)$. Let $\Phi(t_j,t_k) = \omega_{q,p+1}(t_j,t)$ $\omega_{q,p+1}(t_k,t)\{I_0(\beta_j)^{-1}I_0(t_j,t_k)I_0(\beta_k)^{-1}\}^{(rr)}$. Recall that $t_1,t_2,\ldots,t_T$ are i.i.d. from a probability density $f$ under condition (A1). We hereby assume the following regularity conditions: $\mathrm{E}\{I_0(\beta_j)^{-1}\}^{(rr)} < \infty$, $\theta = \mathrm{E}\Phi(t_j,t_k) < \infty$ and $\zeta = \mathrm{E}\{\Phi(t_j,t_k)\}^2 < \infty$. Notice that $\Phi(t_j,t_k)$ is symmetric in its arguments $(t_j,t_k)$. In the notation of Hoeffding (1948), $V_T(1)$ is proportional to a U-Statistic satisfying all conditions in Theorem 7.1. By this theorem, we have $\sqrt{T}[V_T(1)/\{T(T-1)\} - \theta] \xrightarrow{d} N(0,4\zeta)$. Thus $V_T(1) = O_p(T^2)$ since $\theta > 0$ and $\zeta > 0$ in general. As a direct result from SLLN, $V_T(2) = \sum_{j=1}^{T}\omega_{q,p+1}^2(t_j,t)\{I_0(t_j,t_j)^{-1}\}^{(rr)} = O_p(T)$. Therefore, $V_T = V_T(1) + V_T(2) = O_p(T^2)$. Furthermore, from the proof of Theorem 1, we have $\sum_{j=1}^{T}\omega_{q,p+1}(t_j,t)\beta_r(t_j) = \beta_r^{(q)}(t) + O_p(h^{p-q+1}) + O_p(1/n)$. Or equivalently, $\omega_T(t)P^{(r)}\beta - \beta_r^{(q)}(t) = O_p(h^{p-q+1}) + O_p(1/n)$. Therefore, we have $V_T^{-1/2}\sqrt{n}\{\omega_T(t)P^{(r)}\beta - \beta_r^{(q)}(t)\} = O_p(\sqrt{n}h^{p-q+1}/T) + O_p\{1/(\sqrt{n}T)\}$. This completes the proof.

## References


Agresti, A. (2002). *Categorical Data Analysis*. John Wiley, New York.

Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Cao, H., Zeng, D. and Fine, J. P. (2014). Regression analysis of sparse asynchronous longitudinal data. Technical report.

Cai, Z., Fan, J., Li and R. Z. (2000). Efficient estimation and inferences for varying coefficient

models. *Journal of the American Statistical Association* **95**, 888-902.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. In *Statistical Models in S* (Chamber J.M. and Hastie T.J. eds), 309-376. Wadsworth and Brooks, Pacific Grove.

Estes, J., Nguyen, D. V., Dalrymple, L. S., Mu, Y. and Şentürk, D. (2014). Cardiovascular event risk dynamics over time in older patients on dialysis: A generalized multiple-index varying coefficient model approach. *Biometrics*, in press.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Application*, Chapman and Hall, London.

Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data, *J. Roy. Statist. Soc. Ser. B* **62**, 303-322.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*, Chapman and Hall, London.

Gardiner, S. K. and Crabb, D. P. (2002). Frequency of testing for detecting visual field progression. *Br J Ophthalmol* **86**, 560-564.

Gourieroux, C. and Monfort, A. (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics* **17**, 83-97.

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

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *Annals of Mathematical Statistics* **19**, 293-325.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13-22.

McCullagh, P. and Nelder J. A. (1989). *Gerneralized Linear Models*, Chapman and Hall, London.

Park, C. G., Park, T. and Shin, D. W. (1996). A simple method for generating correlated binary variates, *The American Statistician* **50**, 306-310.

Pendergast, J. F., Gange, S. J. Newton, M. A., Lindstrom, M. J. Palta, M. and Fisher, M. R. (1996). A Survey of Methods for Analyzing Clustered Binary Response Data, *International Statistical Review* **64**, 89-118.

Qu, A. and Li, R. (2006). Nonparametric modeling and inference functions for longitudinal data, *Biometrics* **62**, 379-391.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression, *J. Amer. Statist. Assoc.* **90**, 1257-70.

Şentürk, D., Dalrymple, L. S., Mohammed, S. M., Kaysen, G. A. and Nguyen, D. V. (2013). Modeling time varying effects with generalized and unsynchronized longitudinal data. *Statistics in Medicine* **32**, 2971-2987.

Shoptaw, S., Fuller, E.R., Yang, X., Frosch, D., Nahom, D., Jarvik, M.E., Rawson, R.A. and Ling, W. (2002). Smoking Cessation in Methadone Maintenance. *Addiction* **97**, 1317-1328.

Van der Vaart A.W. (1989). *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *J. Statist. Comput. Simul.* **48**, 233-243.

Zhang, D. (2004). Generalized Linear Mixed Models with Varying Coefficients for Longitudinal Data. *Biometrics* **60**, 8-15.

Table 1: Comparison of coverage rates (in %) of the pointwise confidence intervals based on TS, GLMM and GEE at 8 selected time points at nominal confidence level $1-\alpha = 95\%$ and 90%.

| | | Model I | | | | | | Model II | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 95% | | | 90% | | | 95% | | | 90% | |
| | t | TS | GLMM | GEE | TS | GLMM | GEE | TS | GLMM | GEE | TS | GLMM | GEE |
| $\beta_0(t)$ | 1 | 94.4 | 97.0 | 95.8 | 89.8 | 93.2 | 90.0 | 96.0 | 94.4 | 96.0 | 90.8 | 89.4 | 90.4 |
| | 12 | 93.6 | 0.0 | 0.0 | 89.4 | 0.0 | 0.0 | 96.0 | | | 92.8 | | |
| | 24 | 94.8 | 0.0 | 0.0 | 87.2 | 0.0 | 0.0 | 98.4 | | | 94.8 | | |
| | 36 | 95.0 | 0.0 | 0.0 | 87.6 | 0.0 | 0.0 | 97.6 | | | 91.8 | | |
| | 47 | 94.2 | 0.0 | 0.0 | 88.0 | 0.0 | 0.0 | 97.2 | | | 94.2 | | |
| | 59 | 93.4 | 0.0 | 0.0 | 87.2 | 0.0 | 0.0 | 97.2 | | | 93.0 | | |
| | 71 | 94.6 | 0.0 | 0.0 | 88.6 | 0.0 | 0.0 | 98.0 | | | 95.2 | | |
| | 82 | 96.6 | 97.0 | 95.8 | 93.0 | 93.2 | 90.0 | 96.6 | | | 91.2 | | |
| $\beta_1(t)$ | 1 | 93.0 | 0.2 | 0.4 | 87.2 | 0.2 | 0.2 | 93.6 | 94.2 | 94.4 | 89.2 | 90.0 | 89.6 |
| | 12 | 93.2 | 79.8 | 86.8 | 87.4 | 69.2 | 75.0 | 96.8 | | | 93.2 | | |
| | 24 | 94.0 | 95.0 | 93.6 | 87.8 | 90.6 | 90.4 | 98.0 | | | 94.4 | | |
| | 36 | 95.4 | 88.8 | 85.6 | 89.6 | 81.0 | 76.2 | 97.8 | | | 94.8 | | |
| | 47 | 94.8 | 78.4 | 73.2 | 89.2 | 68.8 | 59.4 | 98.4 | | | 95.2 | | |
| | 59 | 95.6 | 66.8 | 56.4 | 90.4 | 53.8 | 45.0 | 97.0 | | | 93.6 | | |
| | 71 | 95.2 | 55.0 | 45.2 | 89.8 | 43.2 | 35.4 | 96.6 | | | 94.2 | | |
| | 82 | 94.8 | 46.6 | 38.2 | 88.8 | 35.6 | 28.4 | 94.4 | | | 89.4 | | |
| $\beta_2(t)$ | 1 | 94.2 | 0.0 | 0.0 | 88.8 | 0.0 | 0.0 | 96.0 | 85.6 | 95.2 | 91.2 | 76.6 | 90.0 |
| | 12 | 92.6 | 0.0 | 0.0 | 87.8 | 0.0 | 0.0 | 97.8 | | | 92.0 | | |
| | 24 | 94.4 | 0.0 | 0.0 | 88.2 | 0.0 | 0.0 | 98.2 | | | 95.0 | | |
| | 36 | 94.0 | 76.4 | 33.0 | 90.0 | 62.6 | 21.2 | 97.4 | | | 93.6 | | |
| | 47 | 95.2 | 62.8 | 40.2 | 90.2 | 47.8 | 28.6 | 97.8 | | | 94.8 | | |
| | 59 | 94.4 | 0.0 | 0.0 | 87.2 | 0.0 | 0.0 | 98.0 | | | 93.8 | | |
| | 71 | 95.0 | 0.0 | 0.0 | 89.2 | 0.0 | 0.0 | 97.2 | | | 93.8 | | |
| | 82 | 94.4 | 0.0 | 0.0 | 89.0 | 0.0 | 0.0 | 96.0 | | | 89.8 | | |

Table 2: Comparison of the mean estimated integrated mean square error (IMSE) over 500 Monte Carlo runs. Estimated IMSE is taken to be the sum of the estimated MSE across all 36 time points.

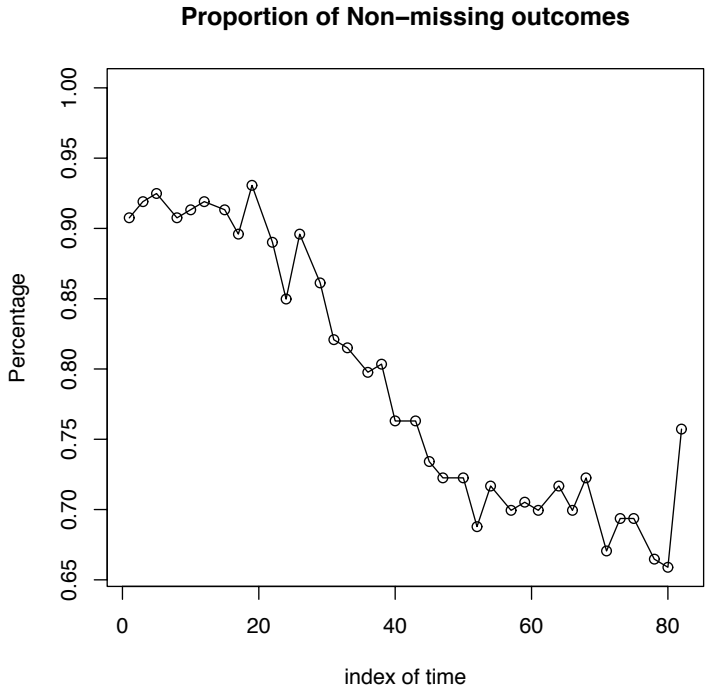
| | | IMSE | | | | IMSE Ratio | | |
|---|---|---|---|---|---|---|---|---|
| | | TS | GLMM | GEE | LML | GLMM/TS | GEE/TS | LML/TS |
| Model 1 | $\beta_0(t)$ | 2.5818 | 69.6154 | 69.5812 | 2.4231 | 26.96 | 26.95 | 0.94 |
| | $\beta_1(t)$ | 1.2855 | 0.6183 | 0.6427 | 1.5610 | 0.48 | 0.50 | 1.21 |
| | $\beta_2(t)$ | 7.4790 | 53.6650 | 53.2927 | 10.3667 | 7.18 | 7.13 | 1.39 |
| Model 2 | $\beta_0(t)$ | 2.9709 | 0.6714 | 0.5635 | 2.0040 | 0.23 | 0.19 | 0.67 |
| | $\beta_1(t)$ | 0.2523 | 0.0576 | 0.0495 | 0.1709 | 0.23 | 0.20 | 0.68 |
| | $\beta_2(t)$ | 0.0059 | 0.0018 | 0.0009 | 0.0037 | 0.31 | 0.15 | 0.62 |

Table 3: Bootstrap confidence bands: Coverage rates (in %) at nominal confidence level $1-\alpha$ = 95% and 90%. All simulations are based on 200 Monte Carlo runs. 1.96∗(standard error) is reported in parenthesis.

| | 95% | | | 90% | | |
|---|---|---|---|---|---|---|
| Setting | $\beta_0(t)$ | $\beta_1(t)$ | $\beta_2(t)$ | $\beta_0(t)$ | $\beta_1(t)$ | $\beta_2(t)$ |
| 1 | 92.5 | 92.0 | 89.0 | 85.0 | 87.5 | 80.0 |
| | (3.65) | (3.76) | (4.34) | (4.95) | (4.58) | (5.54) |
| 2 | 94.5 | 93.0 | 91.0 | 85.5 | 84.5 | 82.0 |
| | (3.16) | (3.54) | (3.97) | (4.88) | (5.02) | (5.32) |

Table 4: Hypotheses Testing: The estimated rejection ratio (in %) for the two hypotheses tests : $H_0(a) : \beta_r(t)$ does not change over time; $H_0(b) : \beta_r(t) = 0$, for all $t \in [t_0, t_T]$. Note that the superscript $^*$ indicates the empirical probability of a Type I error.

| | | $\alpha = 5\%$ | | | $\alpha = 10\%$ | | |
|---|---|---|---|---|---|---|---|
| Setting | $H_0$ | $\beta_0(t)$ | $\beta_1(t)$ | $\beta_2(t)$ | $\beta_0(t)$ | $\beta_1(t)$ | $\beta_2(t)$ |
| 1 | $(a)$ | 100 | 3.5 | 100 | 100 | 9.0 | 100 |
| | $(b)$ | 100 | 27.5 | 100 | 100 | 43.0 | 100 |
| 2 | $(a)$ | $1.5^*$ | $1.5^*$ | $2.5^*$ | $4.0^*$ | $3.5^*$ | $5.5^*$ |
| | $(b)$ | 100 | $7.0^*$ | 100 | 100 | $15.5^*$ | 100 |



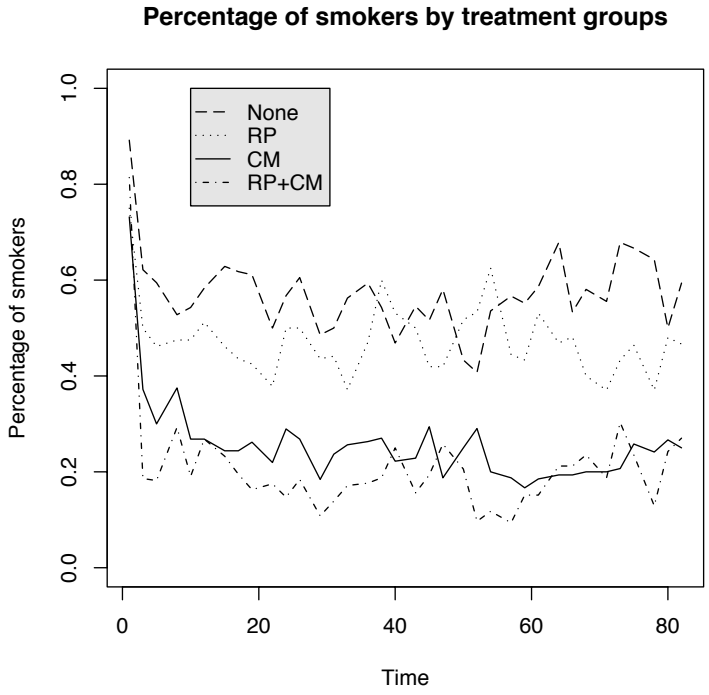




Figure 1: The Smoking Cessation data. Percentage of nonmissing outcomes during each visit of the study (top plot). Percentage of smokers (determined by CO readings) by 4 treatment groups (bottom plot). Index of time is plotted in days.

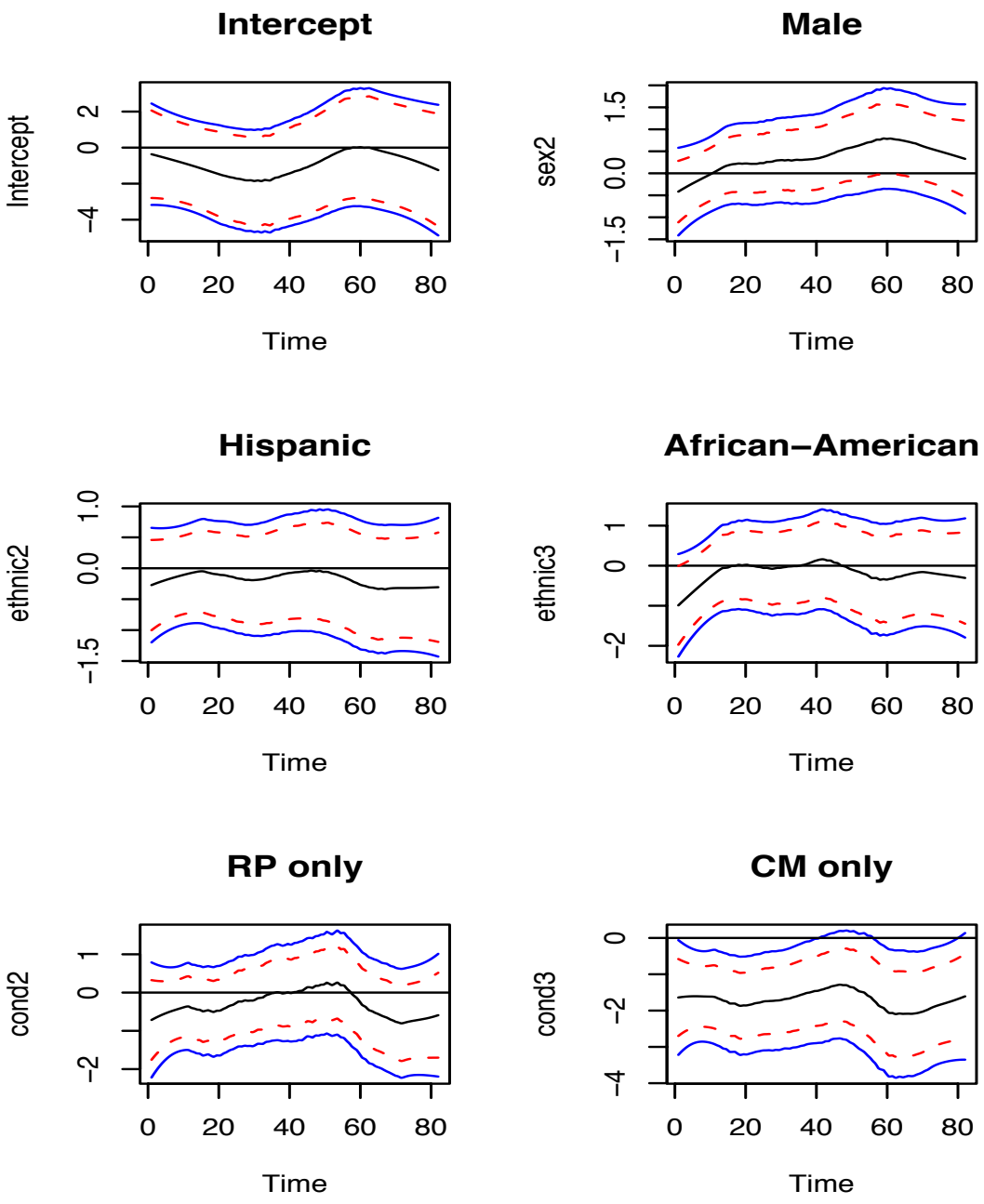



Figure 2: The Smoking Cessation data (part 1). Curve estimate (solid line in center), 95% bootstrap confidence band (solid lines) and 95% point-wise confidence intervals (dash lines).

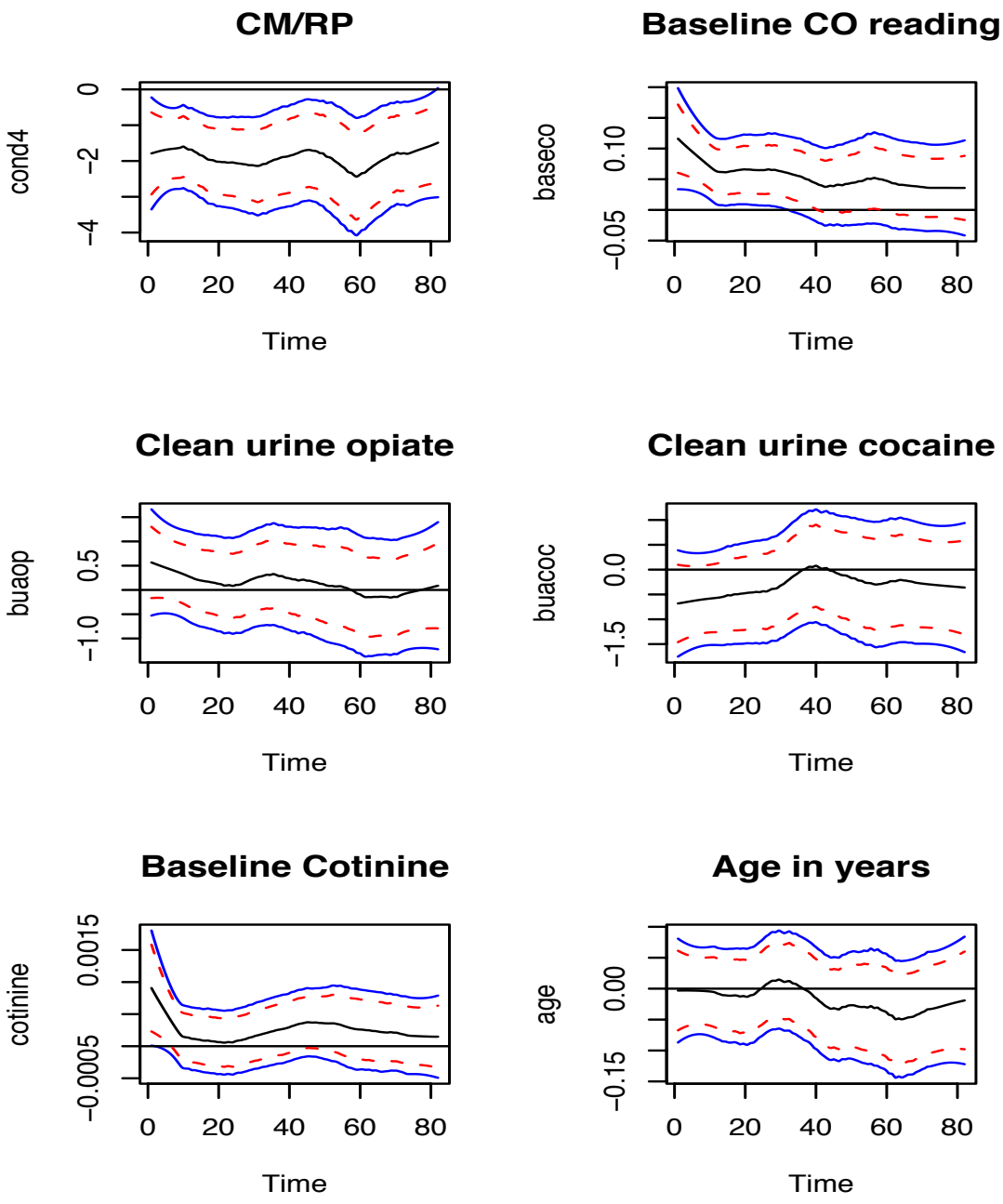



Figure 3: The Smoking Cessation data (part 2). Curve estimate (solid line in center), 95% bootstrap confidence band (solid lines) and 95% point-wise confidence intervals (dash lines).

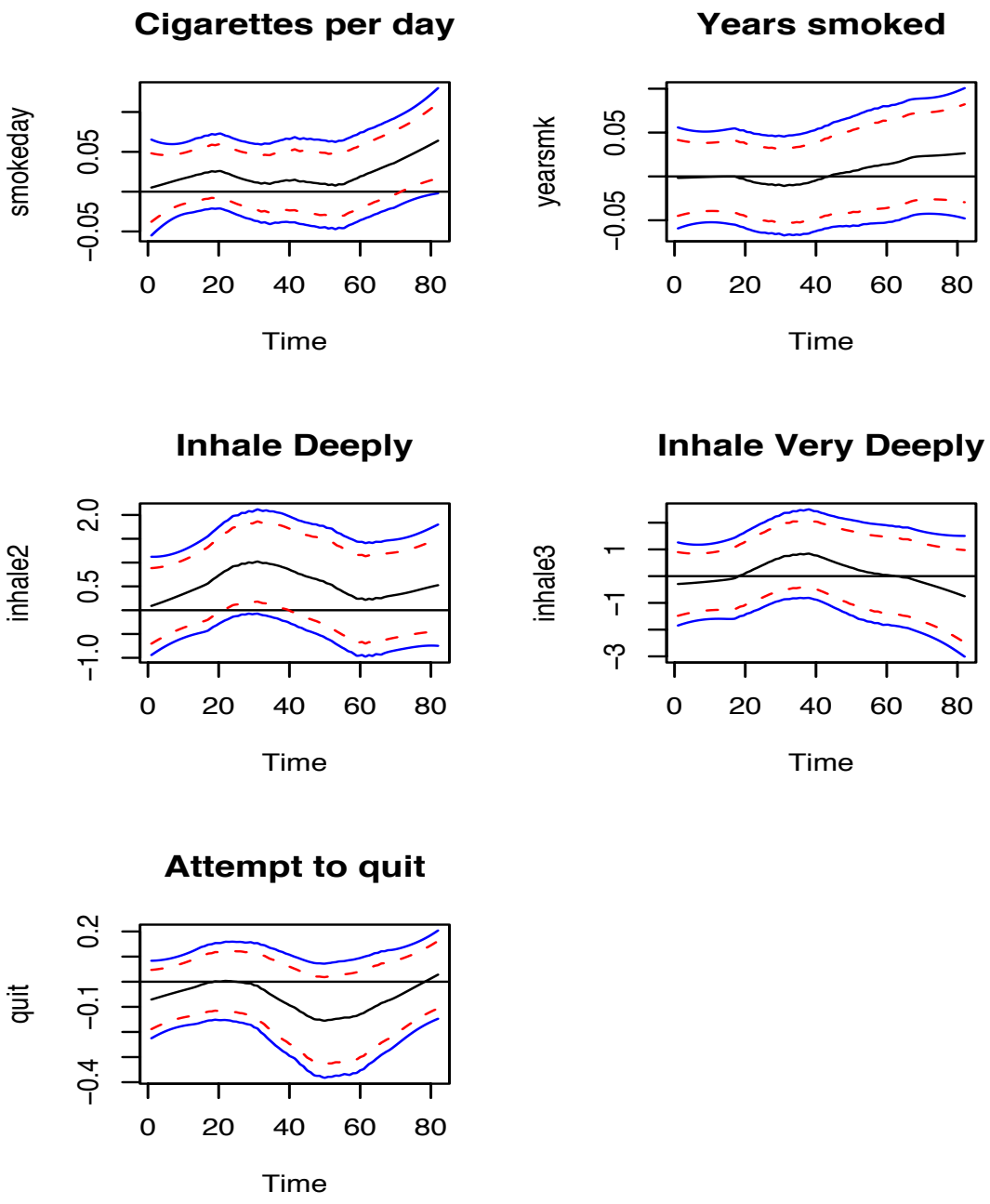



Figure 4: The Smoking Cessation data (part 3). Curve estimate (solid line in center), 95% bootstrap confidence band (solid lines) and 95% point-wise confidence intervals (dash lines).

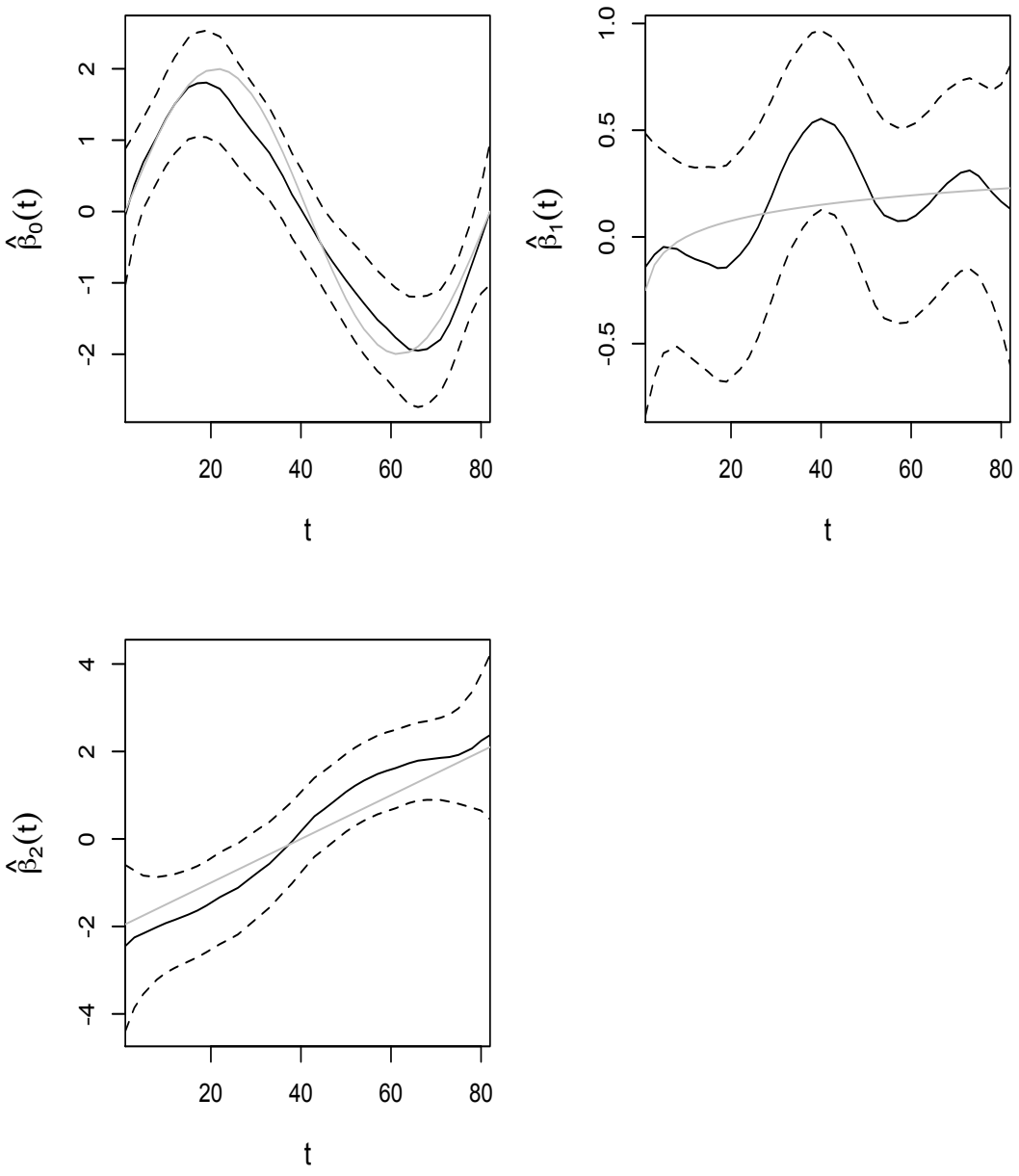



Figure 5: Simulation Model 1: The true varying coefficient functions (solid gray), their estimates (solid black) based on the proposed TS method and 95% bootstrap confidence bands (black dashed) from a the run with median IMSE among 500 Monte Carlo runs.

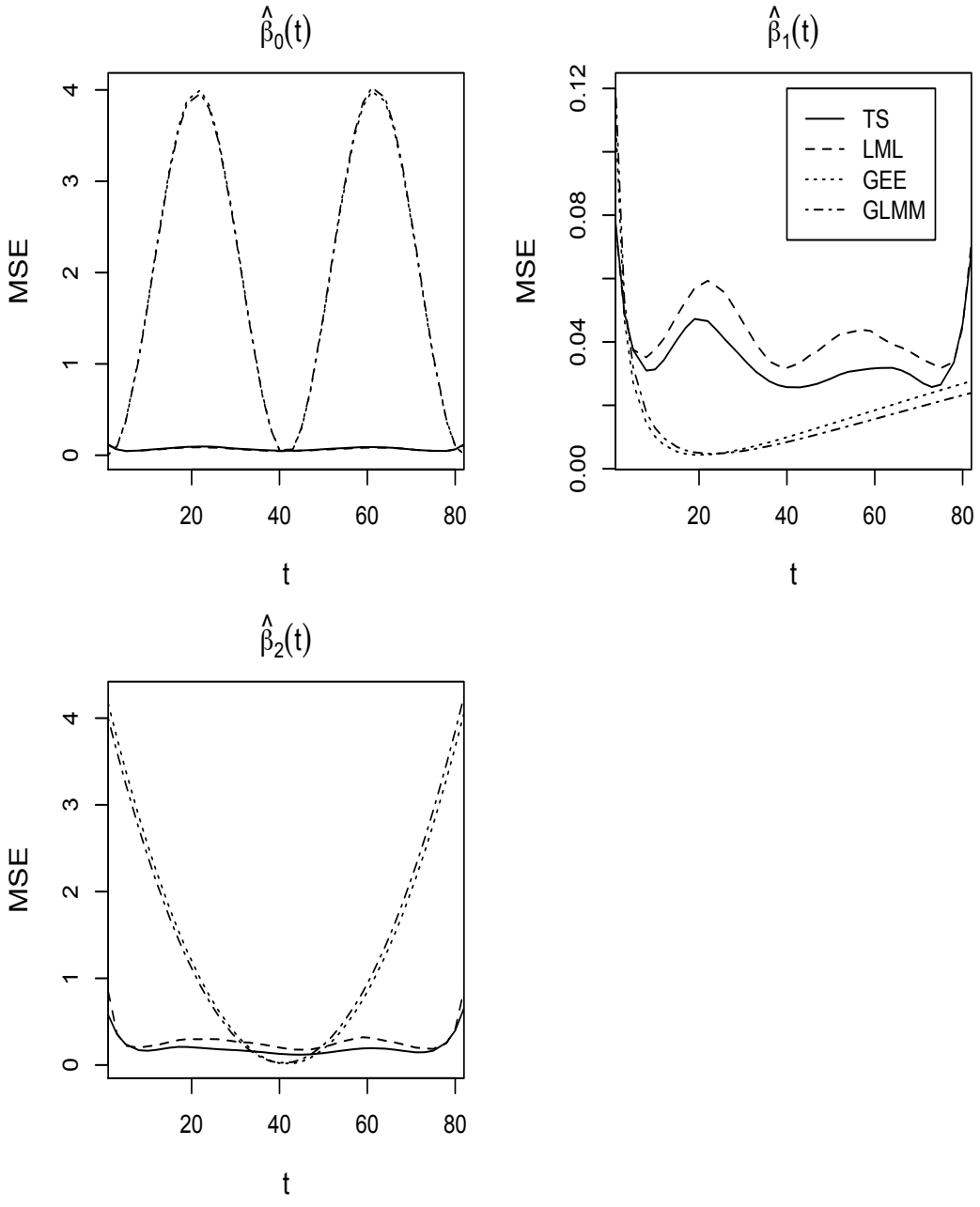



Figure 6: Simulation Model 1: Mean estimated mean square error (MSE) over time.

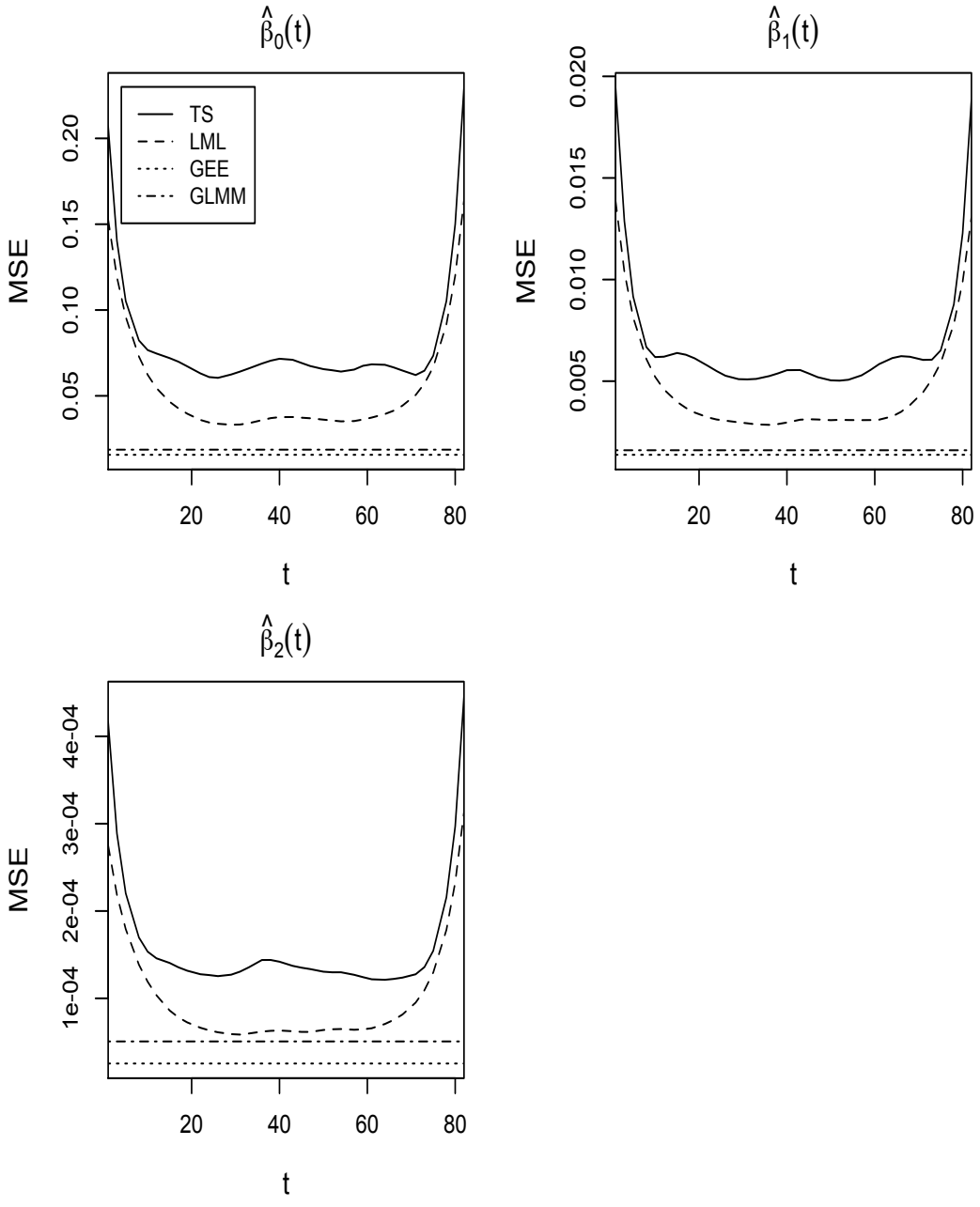



Figure 7: Simulation Model 2: Mean estimated mean square error (MSE) over time.