A Semiparametric Change-Point Regression Model for Longitudinal Observations

Haipeng XING and Zhiliang YING

ABSTRACT: Many longitudinal studies involve relating an outcome process to a set of possibly time-varying covariates, giving rise to the usual regression models for longitudinal data. When the purpose of the study is to investigate the covariate effects when experimental environment undergoes abrupt changes or to locate the periods with different levels of covariate effects, a simple and easy-to-interpret approach is to introduce change-points in regression coefficients. In this connection, we propose a semiparametric change-point regression model, in which the error process (stochastic component) is nonparametric and the baseline mean function (functional part) is completely unspecified, the observation times are allowed to be subject-specific, and the number, locations and magnitudes of change-points are unknown and need to be estimated. We further develop an estimation procedure which combines the recent advance in semiparametric analysis based on counting process argument and multiple change-points inference, and discuss its large sample properties, including consistency and asymptotic normality, under suitable regularity conditions. Simulation results show that the proposed methods work well under a variety of scenarios. An application to a real data set is also given.

KEY WORDS: Change-points; Counting process; Time-varying coefficient.

Haipeng Xing is Associate Professor, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794 (email: *xing@ams.sunysb.edu*) and Zhiliang Ying is Professor, Department of Statistics, Columbia University, New York, NY 10027 (email: *zying@stat.columbia.edu*). This research was supported by the National Science Foundation under grant DMS-1206321. The authors thank an associate editor and two referees for their careful reading of the paper and valuable comments.

1. INTRODUCTION

Longitudinal studies have been widely used in the research of biological, health and social sciences. In a longitudinal study, repeated measurements of the response and predictor (or treatment) variables from different subjects are collected at irregular and possibly subject-specific time points, and hence inherit features from both multivariate and time series analysis; see Nesselroade and Baltes (1979), Diggle, Heagerty, Liang and Zeger (2002), Wooldridge (2002), Frees (2004) and references therein. To investigate the dynamic relationship between repeated measurements in an outcome process and covariates, various inference procedures have been developed in the form of regression analysis over the last several decades. This includes parametric methods for linear and nonlinear regressions (Laird, Donnelly, and Ware 1992; Diggle, Liang, and Zeger 1994; Davidian and Giltinan 1995), nonparametric regressions with kernel, spline and other smoothing techniques (Rice and Silverman 1991; Altman and Casella 1995; Silverman 1996; Lin and Carroll 2000, 2001), smoothing methods for nonparametric regressions with smooth time-varying coefficients (Hoover, Rice, Wu, and Yang 1998; Wu, Chiang, and Hoover 1998; Huang, Wu and Zhou 2002, 2004; Xue and Zhu 2007), functional linear models (Fan and Zhang 2000; James, Wang and Zhu 2009), and inference methods on semiparametric regressions (Moyeed and Diggle 1994; Zeger and Diggle 1994; Lin and Ying 2001; Lin and Carroll 2006; Li 2011). In particular, semiparametric regression models follow the idea of the Cox (1972) proportional hazards model and keep the regression form of covariate effects while leave the baseline (or intercept) unspecified, hence remove limitations of parametric and nonparametric regressions in practical analysis of longitudinal observations (Wu, Chiang and Hoover 1998).

In this paper, we consider a semiparametric regression analysis for longitudinal studies in which the baseline (or intercept) effect is unspecified and the covariate effects are piecewise constant with unknown multiple change-points. Such specification is motivated by the problem of studying the covariate effects when the experimental environment undergoes sharp changes and locating the periods with different levels of covariate effects, which has been largely ignored in the statistical literature. For example, in credit risk analysis, the authority and banks need to understand how individual firms responses to macroeconomic and firm-specific covariates when the economy experiences structural breaks (Xing, Sun, and Chen 2012); in clinical trials, doctors need to know the effect level of a treatment and the duration of each effect level, so that patients can be assigned with different dose levels over different periods (Desmond et al. 2002). The proposed specification is also closely related to the average regression effect problems (Xu and O'Quigley 2000, Schemper, Wakounig, and Heinze 2009) and intervention analysis problems in biological, health, and social sciences, in which one treats the starting and ending times of interventions as change-points (Asgharian and Wolfson 2001, Verbeke and Molenberghs 2000, Rosenfield et al. 2010, Wu, Tian, and Jiang 2011).

Since our model consists of both semiparametric and multiple change-points elements, existing inference procedures in longitudinal analysis do not apply here. For this reason, we also develop an estimation procedure for the change-point effects of covariates that has attractive statistical and computational properties. As the proposed model synthesizes semi-parametric and multiple change-points features, a natural strategy of inference procedure is to modify and synthesize two basic approaches that deal with "degenerate" cases of the model, using the recent advances of semiparametric regression analysis and multiple change-points analysis in time series. In particular, the first degenerate case we consider here involves semiparametric regressions without change-points and a related inference procedure based on counting process argument, developed in Lin and Ying (2001). The second degenerate case involves recent advances in multiple change-points analysis, which has been studied intensively in statistics, econometrics and engineering over the last several decades under both frequentist and Bayesian perspectives. The frequentist methods include the dynamic programming algorithm based on maximized likelihood (Bai 1997a,b, Bai and Perron 1998, 2003, Qu and Perron 2007), the binary segmentation procedure (Vostrikova 1981, Olshen et al. 2004), the model selection procedure based on information criteria (Yao 1988, Siegmund 2004, Davis, Lee, and Rodriguez-Yam 2006, Zhang and Siegmund 2006), and the penalized likelihood methods (Birgé and Massart 2001, Broman and Speed 2002, Lavielle 2005). The Bayesian approaches usually model the change-point as a process and solve the inference problem through the Markov chain Monte Carlo (MCMC) (Albert and Chib 1993, Chib 1998, Liu and Lawrence 1999, Wang and Zivot 2000, Chib, Nardari, and Shephard 2002) or the reversible jump MCMC (Green, 1995). Although these methods may best fit many other applications, they are difficult to be extended here due to the intrinsic counting process features of the proposed model. Therefore, we consider for the second degenerate case the exact approach of estimating model parameters and multiple change-points (Yao 1984, Lai, Liu, and Xing 2005, Lai, Xing, and Zhang 2008, Lai and Xing 2011). Combining such approach with counting process based estimating equations for semiparametric regression, we hence obtain an efficient estimation procedure. Since our model assumes prior distributions for the post-change value of regression coefficients and change-point process, the proposed estimation procedure has a Bayesian interpretation in the sense that the conditional distributions of regression coefficients at particular times, given all observations, can be approximated as the mixtures of posterior distributions. However, the large sample properties established in Section 2.6 have frequentist justification and do not require either regression coefficients or the change-point process has the prior distributional assumption. We also propose a bounded complexity of mixtures approximation and a hyperparameter estimation procedure. We further show that the estimated change-points and regression coefficients are consistent, and establish an oracle property in the sense that the estimators converge to the same limiting normal distribution as those would be used if the change-points were known.

The rest of the paper is organized as follows. Section 2 provides details of the proposed model and related estimating equations. It also develops inference procedure and hyperparameter estimation, and establishes asymptotic properties for the estimated coefficients and change-points. Extensive simulation results are given in Section 3 to assess the performance of the proposed method. Section 3 also includes a longitudinal analysis for the AIDS Clinical Trial Group (ACTG) Protocol 016 study. Section 4 gives a conclusive discussion and points out the related future work.

2. MODEL SPECIFICATION AND INFERENCE PROCEDURE

2.1 Model Specification

Suppose there are n subjects over time period $[0, \overline{C}]$ in the study, in which $\overline{C} = \max\{C_1, \ldots, C_n\}$ and C_i is the follow-up or censoring time of the ith subject $(i = 1, \ldots, n)$. Following Lin and Ying (2001), let $Y_i(t)$ and $\mathbf{X}_i(t)$ be the real- and $\mathbb{R}^d$-valued response and covariate vectors of the ith subject at time t, respectively, for $i = 1, \ldots, n$. We assume that the covariate history $\{\mathbf{X}_i(t) : 0 \leq t \leq C_i\}$ $(i = 1, \ldots, n)$ can be observed for each subject. For the ith subject, let $t_{i,j}$ be the observation time of the jth measurement, $j = 1, \ldots, K_i$, where K_i is the total number of observations on the ith subject. We also consider the observation times as realizations from an arbitrary counting process that is censored at the end of follow-up. In particular, let $N_i^*(t)$ is a discrete- or continuous-time counting process of the observation times of the ith subject, then $N_i(t) \equiv N_i^*(t \wedge C_i)$ is the number of observations taken in the ith subject by time t. Since the process $Y_i(t)$ is observed only at the jump points of $N_i(t)$, we also have $N_i(t) = \sum_{k=1}^{K_i} I(t_{i,k} \leq t)$, where $I(\cdot)$ is the indicator function. We further consider the marginal model for the effect of $\mathbf{X}_i(t)$ on $Y_i(t)$,

$$E\{Y_i(t)|\mathbf{X}_i(t)\} = \alpha(t) + \boldsymbol{\beta}(t)'\mathbf{X}_i(t), \text{ or } Y_i(t_{i,j}) = \alpha(t_{i,j}) + \boldsymbol{\beta}(t_{i,j})'\mathbf{X}_i(t_{i,j}) + \epsilon_i(t_{i,j}), \qquad (1)$$

in which $\epsilon_1(t), \ldots, \epsilon_n(t)$ are zero-mean stochastic processes independent of $\mathbf{X}(t)$, the intercept $\alpha(t)$ is an arbitrary function of t, and the d-dimensional regression coefficient $\boldsymbol{\beta}(t)$ are piecewise constant with unknown number and locations of change-points.

To develop an algorithm which incorporates the feature that both the number and locations of change-points in $\boldsymbol{\beta}(t)$ and the pre- and post-change values of $\boldsymbol{\beta}(t)$ are unknown, we shall, for the time being, assume that the number of change-points in $\boldsymbol{\beta}(t)$ follows a Poisson process $\{J(t); t \geq 0\}$ with rate λ and are independent of $\epsilon(t)$ and $\mathbf{X}(t)$. Then the duration between two adjacent change-points in $\boldsymbol{\beta}(t)$ follows an exponential distribution with mean $1/\lambda$, and $\boldsymbol{\beta}(t)$ between two adjacent change-points are constant. If a change-point occurs at time t, the post-change value of $\boldsymbol{\beta}(t)$ are assumed to be independent of its pre-change value. Such assumption can be characterized as $\boldsymbol{\beta}(t) = \boldsymbol{\omega}_{J(t)}$, where $\boldsymbol{\omega}_0, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots$ are independent and identically distributed (i.i.d.) normal random vectors with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Note that such assumption for $\boldsymbol{\beta}(t)$ allows us to consider the case of unknown number of change-points, analogous to nonparametric Bayes method using Gamma process prior. We also note that model (1) specifies the marginal mean of the process $Y_i(\cdot)$ while leaving its distributional form and dependence structure completely unspecified, and this semiparametric nature is similar to that of the Cox (1972) proportional hazards model and of the regression model discussed in Lin and Ying (2001).

We shall allow the observation time to depend on covariates. Specifically, we assume a marginal relationship between covariates and the intensity $N_i(t)$ of the observation times of the ith subject prior to the censoring time C_i,

$$E[dN_i(t)|C_i, \mathbf{X}_i(t)] = \xi_i(t) e^{\boldsymbol{\gamma}' \mathbf{Z}_i(t)} d\mu(t), \qquad i = 1, \cdots, n, \tag{2}$$

where $\mathbf{Z}_i(t)$ is the part of $\mathbf{X}_i(t)$ that affects the potential observation times, $\xi_i(t) = I(C_i > t)$, $\mu(t)$ is an arbitrary nondecreasing function, and $\boldsymbol{\gamma}$ is a vector of unknown regression parameters.

Model (2) has been proposed for recurrent event times in the survival analysis literature (Pepe and Cai, 1993; Lawless and Nadeau, 1995; Lin et al., 2000). In practice, $\mu(t)$ can be estimated by $\widehat{\mu}(t) = \sum_{i=1}^{n} \int_{t^*}^{t} \frac{dN_i(s)}{\sum_{j=1}^{n} \xi_j(s) e^{\boldsymbol{\gamma}'\mathbf{Z}_j(s)}}$, and $\boldsymbol{\gamma}$ can be estimated by $\widehat{\boldsymbol{\gamma}}$, which is the solution to $\sum_{i=1}^{n} \int_0^{\infty} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t;\boldsymbol{\gamma})\} dN_i(t) = \mathbf{0}$, where $\bar{\mathbf{Z}}(t;\boldsymbol{\gamma}) = \{\sum_{i=1}^{n} \xi_i(t) e^{\boldsymbol{\gamma}'\mathbf{Z}_i(t)} \mathbf{Z}_i(t)\} / \{\sum_{i=1}^{n} \xi_i(t) e^{\boldsymbol{\gamma}'\mathbf{Z}_i(t)}\}$; see Lawless and Nadeau (1995) and Lin et al. (2000). In our subsequent discussion, we shall omit $\boldsymbol{\gamma}$ in $\boldsymbol{\beta}$ for notational simplicity where there is no ambiguity. The censoring time C_i is allowed to depend on the vector of covariates $\mathbf{X}_i(t)$ in an arbitrary manner. We assume the usual noninformative censoring in the sense that $E[Y_i(t)|\mathbf{X}_i(t), C_i \geq t] = E[Y_i(t)|\mathbf{X}_i(t)]$.

2.2 Estimating Equations with a Normal Prior

We consider here an estimation procedure for the degenerate case that $\boldsymbol{\beta}(t)$ is a constant random variable with $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ prior given observations in interval (t_*, t^*), that is, $\boldsymbol{\beta}(t) = \boldsymbol{\theta}$ for $t \in (t_*, t^*)$, and $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In such case, model (1) degenerates to

$$Y_i(t_{i,j}) = \alpha(t_{i,j}) + \boldsymbol{\theta}'\mathbf{X}_i(t_{i,j}) + \epsilon_i(t_{i,j}), \qquad t \in (t_*, t^*), \tag{3}$$

in which $i \in \{j; C_j > t_*\}$. To make inference on $\boldsymbol{\theta}$, we consider estimating equations that are similar to the ones in Lin and Ying (2001). Specifically, let $\mathcal{A}_{t_*,t^*}(t;\boldsymbol{\theta}) = \int_{t_*}^{t} \alpha(s) d\mu(s)$, and

$$M_{t_*,t^*;i}(t;\mathcal{A},\boldsymbol{\theta}) = \int_{t_*}^{t} \left[\{Y_i(s) - \boldsymbol{\theta}'\mathbf{X}_i(s)\} dN_i(s) - \xi_i(s) e^{\boldsymbol{\gamma}'\mathbf{Z}_i(s)} d\mathcal{A}_{t_*,t^*}(s;\boldsymbol{\theta}) \right]. \tag{4}$$

Note that $\mathcal{A}_{t_*,t^*}(t;\boldsymbol{\theta}) = 0$ and $M_{t_*,t^*;i}(t;\mathcal{A},\boldsymbol{\theta}) = 0$, if the ith subject is not observed in (t_*, t^*) or $i \notin \{j; C_j > t_*\}$. Since substituting $E[dN_i^*(t)|\mathbf{X}_i(t)]$ and $Y_i(t_{i,j})$ into the expectation of $M_{t_*,t^*;i}(t;\mathcal{A},\boldsymbol{\theta})$ results in the expected $M_{t_*,t^*;i}(t;\mathcal{A},\boldsymbol{\theta})$ being 0, it follows that $M_{t_*,t^*;i}(t;\mathcal{A},\boldsymbol{\theta})$ are zero-mean stochastic processes. Therefore, it is reasonable to consider the estimating equation $\sum_{i=1}^n M_{t_*,t^*;i}(t;\mathcal{A},\boldsymbol{\theta}) = 0$ for $\mathcal{A}_{t_*,t^*}(t;\boldsymbol{\theta})$, which yields the estimate

$$\widehat{\mathcal{A}}_{t_*,t^*}(t;\boldsymbol{\theta}) = \sum_{i=1}^n \int_{t_*}^t \frac{\{Y_i(s) - \boldsymbol{\theta}'\mathbf{X}_i(s)\}dN_i(s)}{\sum_{j=1}^n \xi_j(s)e^{\boldsymbol{\gamma}'\mathbf{Z}_j(s)}}. \tag{5}$$

As $\boldsymbol{\theta}$ follows the prior distribution $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, a natural estimating equation for $\boldsymbol{\theta}$ could be $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta}) + \sum_{i=1}^n \int_{t_*}^{t^*} W(t)\mathbf{X}_i(t)dM_{t_*,t^*;i}(t;\mathcal{A},\boldsymbol{\theta}) = \mathbf{0}$, where $W(t)$ is a possibly data-dependent weight function. Replacing $\mathcal{A}_{t_*,t^*}(t;\boldsymbol{\theta})$ in this equation by $\widehat{\mathcal{A}}_{t_*,t^*}(t;\boldsymbol{\theta})$ in (5) yields

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta}) + \sum_{i=1}^n \int_{t_*}^{t^*} W(t)\{\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\}\big[Y_i(t) - \boldsymbol{\theta}'\mathbf{X}_i(t)\big]dN_i(t) = \mathbf{0}, \tag{6}$$

where $\bar{\mathbf{X}}_{t_*,t^*}(t) = \{\sum_{i=1}^n \xi_i(t)e^{\boldsymbol{\gamma}'\mathbf{Z}_i(t)}\mathbf{X}_i(t)\}/\{\sum_{i=1}^n \xi_i(t)e^{\boldsymbol{\gamma}'\mathbf{Z}_i(t)}\}$. Making use of (2) and the definition of $\bar{\mathbf{X}}_{t_*,t^*}(t)$, direct calculation shows that, for any function $g(\cdot)$,

$$E\Big[\sum_{i=1}^n \int_{t_*}^{t^*} \{\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\}g(t)dN_i(t)\Big|\{\mathbf{X}_i, C_i; i=1,\ldots,n\}\Big] = \mathbf{0}.$$

We then modify (6) and obtain a class of estimating functions for $\boldsymbol{\theta}$, $\mathbf{U}_g(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta}) + \sum_{i=1}^n \int_{t_*}^{t^*} W(t)\{\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\}\{Y_i(t) - \boldsymbol{\theta}'\mathbf{X}_i(t) - g(t)\}dN_i(t)$. Note that the slope of $\mathbf{U}_g(\boldsymbol{\theta})$ is invariant over g. Thus it is desirable to choose a g to minimize the variance of $\mathbf{U}_g(\boldsymbol{\theta})$. We further notice that

$$\begin{aligned}\mathbf{U}_g(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta}) &+ \sum_{i=1}^n \int_{t_*}^{t^*} W(t)\{\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\}\epsilon_i(t)dN_i(t)\\ &+ \sum_{i=1}^n \int_{t_*}^{t^*} W(t)\{\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\}\{\alpha(t) - g(t)\}\{dN_i(t) - \xi_i(t)d\mu(t)\}.\end{aligned}$$

Since the three terms on the right side of the above equation are uncorrelated, an optimal choice of $g(\cdot)$ should be $\alpha(t)$. As $\alpha(t)$ is unspecified, we may estimate $\alpha(t)$ by $\bar{Y}^*_{t_*,t^*}(t) - \boldsymbol{\theta}'\bar{\mathbf{X}}_{t_*,t^*}(t)$, where $\bar{Y}^*_{t_*,t^*}(t) = \big\{\sum_{i=1}^n \xi_i(t)e^{\boldsymbol{\gamma}'\mathbf{Z}_i(t)}Y_i(t)\big\}\big/\big\{\sum_{i=1}^n \xi_i(t)e^{\boldsymbol{\gamma}'\mathbf{Z}_i(t)}\big\}$, for $t \in (t_*, t^*)$.

In practice, $Y_i(t)$ in the above estimate may be replaced by the measurement of Y_i at the time point nearest to t. We then set $g(t) = \bar{Y}^*_{t_*,t^*}(t) - \boldsymbol{\theta}'\bar{\mathbf{X}}_{t_*,t^*}(t)$ and obtain the estimating equation

$$\mathbf{U}_{t_*,t^*}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta})+\sum_{i=1}^{n}\int_{t_*}^{t^*} W(t)\{\mathbf{X}_i(t)-\bar{\mathbf{X}}_{t_*,t^*}(t)\}\left[Y_i(t)-\bar{Y}^*_{t_*,t^*}(t)-\boldsymbol{\theta}'\{\mathbf{X}_i(t)-\bar{\mathbf{X}}_{t_*,t^*}(t)\}\right]dN_i(t). \tag{7}$$

Solving $\mathbf{U}_{t_*,t^*}(\boldsymbol{\theta}) = \mathbf{0}$ yields

$$\widehat{\boldsymbol{\theta}}_{(t_*,t^*)} = \boldsymbol{\Sigma}_{(t_*,t^*)}\boldsymbol{\mu}_{(t_*,t^*)}, \tag{8}$$

where $\boldsymbol{\mu}_{t_*,t^*} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \sum_{i=1}^{n}\int_{t_*}^{t^*} W(t)\{\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\}\{Y_i(t) - \bar{Y}^*_{t_*,t^*}(t)\}dN_i(t)$, $\boldsymbol{\Sigma}_{t_*,t^*} = \left[\boldsymbol{\Sigma}^{-1} + \sum_{i=1}^{n}\int_{t_*}^{t^*} W(t)\{\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\}^{\otimes 2}dN_i(t)\right]^{-1}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$. Note that equation (8) depends on hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which, if unspecified, can be estimated by the approach discussed in Section 2.5. The derived estimator $\widehat{\boldsymbol{\theta}}_{(t_*,t^*)}$ takes a form similar to the Bayes estimator under normality assumption. However, it is not exactly a Bayes estimator since there is no additional assumption on ϵ_i and no specification on $\alpha(t)$. Given $\widehat{\boldsymbol{\theta}}_{(t_*,t^*)}$, $\mathcal{A}_{t_*,t^*}(t;\boldsymbol{\theta})$ is naturally estimated by $\widehat{\mathcal{A}}_{t_*,t^*}(t;\widehat{\boldsymbol{\theta}}_{(t_*,t^*)})$.

Remark 1. Note that no matter what weight function $W(\cdot)$ is chosen, the proposed estimator $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normal. However, the variance of $\widehat{\boldsymbol{\theta}}$ depends on $W(t)$. If $\epsilon_i(t)$ are independent over time and the variances of the ϵ_i's do not depend on i, then an optimal choice of $W(t)$ is $1/\sigma^2_\epsilon(t)$, in which $\sigma^2_\epsilon(t)$ can be obtained by aggregating the observations around t. Substituting $\boldsymbol{\gamma}$ in (11) by $\widehat{\boldsymbol{\gamma}}$, we arrive at estimator $\widehat{\boldsymbol{\theta}}_{(t_*,t^*)}(\widehat{\boldsymbol{\gamma}})$. We argue in Appendix A.1 that, for any fixed $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, $n^{1/2}(\widehat{\boldsymbol{\theta}}_{(t_*,t^*)}(\widehat{\boldsymbol{\gamma}}) - \boldsymbol{\theta})$ converges in distribution to a d-variate zero-mean normal random vector, whose covariance doesn't depend on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as the effect of prior diminishes when $n \to \infty$.

Remark 2. In the procedure above, the unspecified function $\alpha(t)$ is estimated by $\widehat{\alpha}_{t_*,t^*}(t) = \bar{Y}_{t_*,t^*}(t) - \boldsymbol{\theta}'\bar{\mathbf{X}}_{t_*,t^*}(t)$. Plugging it into (3) yields a reduced regression form

$Y_i(t) - \bar{Y}_{t_*,t^*}(t) = \boldsymbol{\theta}\big[\mathbf{X}_i(t) - \bar{\mathbf{X}}_{t_*,t^*}(t)\big] + \epsilon_i(t)$. As the inference procedure in Section 2.3 involves the likelihood of model (3), we may approximate it in the following way. Assume $W(t) = 1/\sigma_\epsilon^2(t)$ and $\epsilon_i(t)$ are independent zero mean normal random variables. Since the posterior distribution of $\boldsymbol{\theta}$ in the reduced regression is $N(\widehat{\boldsymbol{\theta}}_{(t_*,t^*)}, \boldsymbol{\Sigma}_{t_*,t^*})$, the likelihood of the reduced model can be obtained by integrating conditional likelihood of $\{Y_i(t) - \bar{Y}_{t_*,t^*}(t); t \in (t_*, t^*), 1 \le i \le n\}$ over the posterior normal distribution of $\boldsymbol{\theta}_{t_*,t^*}$, and we denote it by

$$\psi_{t_*,t^*} = \exp\left\{\frac{1}{2}\widehat{\boldsymbol{\theta}}_{(t_*,t^*)}\boldsymbol{\Sigma}^{-1}_{(t_*,t^*)}\widehat{\boldsymbol{\theta}}_{(t_*,t^*)} - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{\mu}\right\}|\boldsymbol{\Sigma}_{t_*,t^*}|^{1/2}/|\boldsymbol{\Sigma}|^{1/2}. \tag{9}$$

2.3 An Estimation Procedure for $\boldsymbol{\beta}(t)$

We now consider an estimation procedure of $\boldsymbol{\beta}(t)$ for a grid of the sample period: $0 < s_1 < \cdots < s_L = \overline{C}$ (for convenience, we assume $\overline{C} = 1$), where $s_l = l/L$ and $L \sim O(n^\xi)$ for some $\xi \in (0,1)$. We assume that change-points only happen at times $s_1, \ldots, s_L$. We define the variables $J_1 = 1$ and $J_l = J(t_l-) - J(t_{l-1}-)$ for $l = 2, \ldots, L$ to indicate if $\boldsymbol{\beta}(t)$ are same at the periods (s_{l-2}, s_{l-1}) and (s_{l-1}, s_l), then $\{J_l\}$ is a sequence of independent Bernoulli random variables with success probability $p_l = 1 - \exp(-\lambda/L)$. Let $\mathcal{Y}_{(s_m,s_k)} = \{\mathbf{X}_i(t), Y_i(t); i = 1, \ldots, n, t \in (s_m, s_k)\}$ and $\boldsymbol{\beta}_{(s_m,s_k)}$ be the regression coefficient in model (1) for $t \in (s_m, s_k)$ when s_m and s_k are two adjacent change-points. To estimate $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,s_L)}$, we note that, for any estimating function $U(\cdot|\mathcal{Y}_{(0,s_L)})$, we have

$$U(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_L)}) = \sum_{1 \le m \le l \le k \le L} \pi_{mlk} U(\boldsymbol{\beta}_{(s_{m-1},s_k)}|\mathcal{Y}_{(s_{m-1},s_k)}), \tag{10}$$

in which π_{mlk} is the probability that two most recent change-times around s_l are s_m and s_k ($s_m \le s_l < s_k$). Due to the mixture feature of (10), we should compute the mixture probabilities $\{\pi_{mlk}\}$ first.

Let $R_l = \max\{s_{m-1}|J_m = 1, m \le l\}$, i.e., R_l represents the time of the most recent change-point up to time t_{l-1}, and $\eta_{m,l} = P(R_l = s_{m-1}|\mathcal{Y}_{(0,s_l)})$. Then the posterior distribution of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,s_l)}$ is expressed as

$$f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_l)}) = \sum_{m=1}^{l} \eta_{m,l} f(\boldsymbol{\beta}_{(s_{m-1},s_l)}|\mathcal{Y}_{(s_{m-1},s_l)}). \tag{11}$$

$f(\boldsymbol{\beta}_{(s_{m-1},s_l)}|\mathcal{Y}_{(s_{m-1},s_l)})$ is the posterior distribution of $\boldsymbol{\beta}(s_l)$ given $R_l = s_{m-1}$ and $\mathcal{Y}_{(s_{m-1},s_l)}$, and the mixture probabilities are expressed as $\eta_{m,l} = \eta^*_{m,l}/\sum_{m=1}^{l}\eta^*_{m,l}$, and

$$\eta^*_{m,l} = \begin{cases} p_l \psi_{s_l,s_l} & m = l, \\ (1-p_l)\eta_{m,l-1}\psi_{s_m,s_l}/\psi_{s_m,s_{l-1}} & m < l; \end{cases} \tag{12}$$

see proofs in Appendix A.2. Note that ψ_{s_m,s_l} represents the likelihood of $\mathcal{Y}_{(s_{m-1},s_l)}$ given $R_l = s_{m-1}$, which can be approximated by (9) for $(t_*, t^*) = (s_{m-1}, s_l)$. To construct an estimating function of $\boldsymbol{\beta}(s_l)$ using $\mathcal{Y}_{(0,s_L)}$ for $1 \le l < L$, we denote $\widetilde{R}_{l+1} = \min\{s_k|J_k = 1, k > l\}$ and $\widetilde{\eta}_{k,l+1} = P(\widetilde{R}_{l+1} = s_k|\mathcal{Y}_{s_{l+1},s_L})$. We then show in Appendix A.2 that the posterior distribution of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(s_l,s_L)}$ is

$$f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(s_l,s_L)}) = p_l f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_0) + (1-p_l)\sum_{k=l+1}^{L} \widetilde{\eta}_{k,l+1} f(\boldsymbol{\beta}_{(s_l,s_k)})|\mathcal{Y}_{(s_l,s_k)}), \tag{13}$$

in which $f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_0)$ represents the density of $\boldsymbol{\beta}(s_l)$ without any observations, and the mixture probabilities $\widetilde{\eta}_{k,l+1} = \widetilde{\eta}^*_{k,l+1}/\sum_{k=l+1}^{L}\widetilde{\eta}^*_{k,l+1}$ and

$$\widetilde{\eta}^*_{k,l+1} = \begin{cases} p_{l+1}\psi_{s_{l+1},s_{l+1}} & k = l+1, \\ (1-p_{l+1})\eta_{l+2,k}\psi_{s_{l+1},s_k}/\psi_{s_{l+2},s_k} & k > l+1. \end{cases} \tag{14}$$

Following the probability argument in Appendix A.2, one can use the Bayes theorem to combine functions (11) and (13) to obtain the posterior of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,s_L)}$

$$f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_L)}) = \sum_{1\le m\le l\le k\le L} \pi_{mlk} f(\boldsymbol{\beta}_{(s_{m-1},s_k)}|\mathcal{Y}_{(s_{m-1},s_k)}), \tag{15}$$

in which $\pi_{mlk} = \pi^*_{mlk} \big/ \sum_{1\le i\le l\le j\le L} \pi^*_{ilj}$ and

$$\pi^*_{mlk} = \begin{cases} p_l \eta_{m,l} & m \le l = k, \\ (1-p_l)\eta_{m,l}\widetilde{\eta}_{k,l+1}\psi_{s_m,s_k} \big/ \big(\psi_{s_m,s_l}\psi_{s_{l+1},s_k}\big) & m \le l < k. \end{cases} \tag{16}$$

As the above procedure provides explicit formulas to compute the mixture weights $\{\pi_{mlk}\}$, we use (10) to construct the estimation procedure as follows. First, we use expressions (12), (14), and (16) to compute the mixture probabilities $\{\pi_{mlk}\}$, then we use observations $\mathcal{Y}_{(s_{m-1},s_k)}$ to compute the estimate of $\boldsymbol{\beta}_{(s_{m-1},s_k)}$ by applying the estimating procedure and formula (8) in Section 2.2. Finally, in the spirit of (10), we construct the following estimate of $\boldsymbol{\beta}(s_l)$ given observations $\mathcal{Y}_{(0,s_L)}$,

$$\widehat{\boldsymbol{\beta}}(s_l) = \sum_{1\le m\le l\le k\le L} \pi_{mlk}\widehat{\boldsymbol{\beta}}_{(s_{m-1},s_k)}. \tag{17}$$

2.4 An Estimation Procedure with Bounded Complexity Mixture

In practical analysis, one usually uses a fine grid to obtain a better estimate on change-points. This indicates that the number of mixture in (12) increases with l, resulting in unbounded computational complexity and memory requirements in estimating $\boldsymbol{\beta}(s_l)$ as l changes from 1 to L. To address the issue of computational efficiency, we follow Lai and Xing (2011) and consider a *bounded complexity mixture* (BCMIX) approximation procedure with much lower computational complexity. The idea of BCMIX approximation is to keep only a fixed number B of weights at every s_l, in particular, the most recent b ($1 \le b \le B$)

weights $\eta_{m,l}$ (with $l-b<m\leq l$) and the largest $B-b$ of the remaining weights.

Denote $\mathcal{K}_{l-1}$ be the set of indices j for which $\eta_{j,l-1}$ in (11) is kept at time s_{l-1}; thus $\mathcal{K}_{l-1}\in\{l-1,\ldots,l-b\}$. At time s_l, define $\eta^*_{m,l}$ as in (12) for $l\in\{l\}\cup\mathcal{K}_{l-1}$, and let j_l be the index not belonging to $\{l,\ldots,l-b+1\}$ such that $\eta^*_{j_l,l}=\min\{\eta^*_{j,l}:j\in\mathcal{K}_{l-1}\text{ and }j\leq l-b\}$, choosing j_l to be the minimizer farthest from l if the above set has two or more minimizers. Define $\mathcal{K}_l=\{l\}\cup(\mathcal{K}_{l-1}-\{j_l\})$ and let $\eta_{m,l}=\left(\eta^*_{m,l}\big/\sum_{i\in\mathcal{K}_l}\eta^*_{i,l}\right)$ for $m\in\mathcal{K}_l$. We then get a BCMIX approximation to (11). Similarly, let $\widetilde{\mathcal{K}}_{l+1}$ denote the set of indices for which $\widetilde{\eta}^*_{k,l+1}$ as in (14) for $k\in\{l\}\cup\widetilde{\mathcal{K}}_{l+1}$, and let j_l be the index not belonging to $\{l,\ldots,l+b-1\}$ such that $\widetilde{\eta}^*_{j_l,l}=\min\{\widetilde{\eta}^*_{k,l}:k\in\widetilde{\mathcal{K}}_{l+1}\text{ and }k\geq l+b\}$, choosing j_l to be the minimizer farthest from l if the above set has two or more minimizers. Define $\widehat{\mathcal{K}}_l=\{l\}\cup(\widetilde{\mathcal{K}}_{l+1}-\{j_l\})$ and let $\widetilde{\eta}_{k,l}=\left(\widetilde{\eta}^*_{k,l}\Big/\sum_{i\in\widetilde{\mathcal{K}}_m}\widetilde{\eta}^*_{i,l}\right)$ for $k\in\widetilde{\mathcal{K}}_m$, we obtain a BCMIX approximation to (13).

The BCMIX approximation to (15) can be obtained by combining the above BCMIX approximations to (11) and (13) via the Bayes theorem and the probability argument in Section 2.3,

$$f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_L)})\sim\sum_{m\in\mathcal{K}_l,k\in\widetilde{\mathcal{K}}_{l+1}}\pi_{mlk}f(\boldsymbol{\beta}_{(s_{m-1},s_k)}|\mathcal{Y}_{(s_{m-1},s_k)}),\tag{18}$$

in which $\pi_{mlk}=\pi^*_{mlk}\big/\sum_{i\in\mathcal{K}_l,j\in\widetilde{\mathcal{K}}_{l+1}}\pi^*_{ilj}$ and π^*_{mlk} are given by (16) for $m\in\mathcal{K}_l$ and $k\in\widetilde{\mathcal{K}}_{l+1}$. Therefore, the BCMIX estimate of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(s_{m-1},s_k)}$, which approximates the estimate (17), is expressed as

$$\widehat{\boldsymbol{\beta}}(s_l)\approx\sum_{m\in\mathcal{K}_l,k\in\widetilde{\mathcal{K}}_{l+1}}\pi_{mlk}\widehat{\boldsymbol{\beta}}_{(s_{m-1},s_k)}.\tag{19}$$

2.5 Estimation of Hyperparameters

The estimation procedures in Sections 2.2-2.4 contain hyperparameters $\Phi=\{\lambda,\boldsymbol{\mu},\boldsymbol{\Sigma}\}$. In practical analysis, Φ is unknown and need to be estimated from data. From the definition (12) of $\eta^*_{m,l}$, it follows that the conditional likelihood of $\mathcal{Y}_{(s_{l-1},s_l)}$ given $\mathcal{Y}_{(0,s_{l-1})}$ is $f(\mathcal{Y}_{(s_{l-1},s_l)}|\mathcal{Y}_{(0,s_{l-1})})=\sum_{m=1}^{l}\eta^*_{m,l}$, in which $\eta^*_{m,l}$ are functions of the hyperparameter Φ. Given

Φ and the observed data $\mathcal{Y}_{(0,s_L)}$, the log-likelihood function is expressed as

$$l(\Phi) = \sum_{l=1}^{L} \log f(\mathcal{Y}_{(s_{l-1},s_l)}|\mathcal{Y}_{(0,s_{l-1})}) = \sum_{l=1}^{L} \log\Big\{\sum_{m=1}^{l} \eta^*_{m,l}\Big\}. \quad (20)$$

Given the semiparametric feature in our model, we cannot compute this likelihood exactly. Therefore, we compute $\eta^*_{m,l}$ via approximation of ψ_{s_m,s_l} as stated in Remark 2 and the comment right after (12).

Another issue is on the choice of B and b in the BCMIX algorithm. As the asymptotic properties in Section 2.6 holds for any fixed pair (b, B) $(b < B)$, we could choose a small number of (b, B)'s in the algorithm. Specifically, we have compared the performance of $(b, B) = (10, 20), (10, 15)$ and $(5, 10)$ via some simulation studies, the estimated values don't show much difference. Hence we recommend using $(b, B) = (10, 20)$ in practical applications.

2.6 Asymptotic Properties of the Estimates

In this section, we establish the consistency for the proposed estimates of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,s_L)}$, and provide a segmentation procedure for estimating change-points of $\boldsymbol{\beta}(t)$. We then argue that the segmentation procedure further yields estimates of $\boldsymbol{\beta}(t)$ that are asymptotically normal with the same limiting distribution as those given in Section 2.1 when change-points are assumed to be known. The following theorems, whose proof is given in Appendix A.3, assume the following regularity conditions:

(C1) The true change-points occur at $0 < \tau_1 < \cdots < \tau_M < \overline{C}$, where M and τ_j are unspecified.

(C2) For any $(t_*, t^*) \subset (\tau_{k-1}, \tau_k)$, $k = 1, \ldots, M+1$, the limiting value of $n^{-1}\boldsymbol{\Sigma}^{-1}_{t_*,t^*}$ is bounded away from 0 and infinity.

Remark 1. Condition (C1) does not impose any assumption on the structures of change-points except saying that there are finite but unknown number of change-points. Condition

(C2) is also mild. It basically requires that the covariates must not lie in a lower dimensional hyperplane. In particular, if $\mathbf{X}$ is time independent with a nonsingular covariance matrix and $d\mu/dt > 0$ almost everywhere, then (C2) is satisfied. In general, it is satisfied if the covariance matrix (as a function of t) of $\mathbf{X}(t)$ is nonsingular and $d\mu/dt > 0$ for almost all t.

Theorem 1. *Let $\mathcal{O}_n = \cup_{k=1}^{M+1}(\tau_{k-1} + B/L, \tau_k - B/L)$, where $\tau_0 = 0$ and $\tau_{M+1} = \overline{C}$. Then for $L \sim O(n^\xi)$ $(\xi \in (0,1))$,*

$$\sup_{t_m \in \mathcal{O}_n} ||\widehat{\boldsymbol{\beta}}(t_m) - \boldsymbol{\beta}(t_m)|| \xrightarrow{P} 0 \text{ as } n \to \infty.$$

In the theorem, the consistency holds for any $\xi \in (0,1)$, and the BCMIX parameter B is a fixed number and doesn't depend on n. Theorem 1 essentially tells us that, if there are finite unknown change-points, as long as the partition is not too fine relative to n, the estimated regression coefficients could converge uniformly to the true values in the sample period after we remove small intervals around change-points to avoid boundary effects. A practical concern is how to choose L or the number of grids. Our empirical experience shows that, if one wants to get more accurate estimates of change-points, a larger L which is close to n can be used. For example, in our simulation studies of Section 4, we use $\xi = 0.99$ so that each interval (s_{l-1}, s_l) is approximately 1 percent of the total length $\overline{C}$ of the sample period.

Note that the above consistency still holds when we replace the hyperparameter by their estimates given in Section 2.5. We then consider how to determine the number of change-points in $(0, s_L)$. Denote $\Delta_l = ||\widehat{\boldsymbol{\beta}}(s_{l+w}) - \widehat{\boldsymbol{\beta}}(s_{l-w})||^2$, where $w > 0$ is a fixed number that is comparable to B, for instance, $w = B$ or $2B$. We then estimate the M change-points in $\boldsymbol{\beta}(t)$ sequentially by making use of $\{\Delta_l : w < l < L - w\}$. Let $\widehat{\nu}_1$ be the maximizer of Δ_l over $w < l < L - w$. After $\widehat{\nu}_1, \ldots, \widehat{\nu}_{k-1}$ have been defined, we can define $\widehat{\nu}_k$ as the maximizer of Δ_l over l that lies outside the w-neighborhoods of $\widehat{\nu}_i$ for $1 \le i \le k-1$, i.e.,

$$\Delta_{\widehat{\nu}_k} = \max\{\Delta_l : w < l < L - w, \min_{1 \le i \le k-1} |l - \widehat{\nu}_i| \ge w\}. \tag{21}$$

Note that the jump-times $\tau_1, \ldots, \tau_M$ are ordered, whereas the estimates $\widehat{\nu}_k$ of the times of the change-points in (21) are unordered and do not depend on M. Under the model of M change-points, we can take $\widehat{\nu}_1, \ldots, \widehat{\nu}_M$ and order them as $\nu_{1,n} < \cdots < \nu_{M,n}$ to provides estimates of the M change times. We can then estimate the constant regression coefficient over $(\nu_{k-1,n}, \nu_{k,n})$, $1 \le k \le M$, where $\nu_{0,n} = 0$ and $\nu_{M+1,n} = s_L$. This yields the following pseudo log-likelihood profile with the estimated change times $\nu_{1,n}, \ldots, \nu_{M,n}$: $\ell_n(M) = \sum_{k=1}^{M} \log \psi_{\nu_{k,n}, \nu_{k+1,n}}$. With an upper bound A on the number M of change-points, we propose to estimate M by $\widehat{M}_n = \text{argmax}_{1 \le k \le M}\{\ell_n(k) - \frac{1}{2} d(k+1)\log n\}$, in which d is the dimension of the regression coefficient $\boldsymbol{\beta}(t)$.

Theorem 2. *Under the assumption of Theorem 1,* $\widehat{M}_n \xrightarrow{P} M$ *and* $\nu_{k,n} \xrightarrow{P} \tau_k$, $k = 1, \ldots, M$.

Given the estimated change-points $v_{k,n}$ for τ_k, we can choose $c_n \sim o(n^{\xi})$. Then with probability approaches to 1, $\boldsymbol{\beta}(t) = \boldsymbol{\beta}_k$ for $t \in (v_{k,n} + c_n B/L, v_{k+1,n} - c_n B/L)$, in which $\boldsymbol{\beta}_k$ is the constant value of regression parameters in $(\nu_{k-1,n}, \nu_{k,n})$. We next estimate $\boldsymbol{\beta}_k$ by $\widetilde{\boldsymbol{\beta}}_{v_{k,n}+c_n B/L, v_{k+1,n}-c_n B/L}$, which is defined by (11). Such estimates have the following oracle property.

Theorem 3. $\sqrt{n}(\widetilde{\boldsymbol{\beta}}_{v_{k-1,n}+c_n B/L, v_{k,n}-c_n B/L} - \boldsymbol{\beta}_k)$ *converges in distribution to a d-variate zero-mean normal random vector.*

Note that since the covariance function of the error process $\epsilon(t)$ is completely unspecified, we don't have analytical form for the limiting covariance; though in such case, consistent estimators of the limiting covariance can be constructed. This is analogous to the GEE method in longitudinal data analysis.

Remark 2. The proposed estimation procedure doesn't require prior knowledge of change-point locations. Theorems 1-3 essentially addresses the following concerns. For any finite number of unknown change-points, the time-varying regression coefficients and locations of multiple change-points can be consistently estimated (Theorems 1 & 2). Furthermore, if one is allowed to do one more step, he could use the segmented data to estimate the constant regression coefficients within certain time intervals, and such estimation can be no worse than the case of change-points being known (Theorem 3).

3. NUMERICAL STUDIES

3.1 Performance of the Proposed Estimation Procedure

We conduct extensive simulation studies to evaluate the performance of the proposed estimation procedure in practical settings. We assume the responses are generated from the following semiparametric random-effects model

$$Y(t) = \alpha(t) + \beta_{01}(t)X_1 + \beta_{02}(t)X_2 + \epsilon(t), \tag{22}$$

where $\alpha(t) = a_0 + a_1\sqrt{t}$ or $a_0 + a_1 \sin t$, X_1 is Bernoulli with 0.5 success probability, X_2 is independent standard normal, $\epsilon(t)$ is normal with mean ϕ and variance σ_ϵ^2 for all t. We consider two scenarios here, one is $\phi = 0$ and the other is that, $\phi \sim \text{Normal}(0, \sigma_\phi^2)$. The counting process $N^*(t)$ for the observation times was set to be a random-effect Poisson process with intensity rate η, where η is an independent gamma random variable with mean 1 and variance σ_η^2. For nonzero σ_η^2, the observation times within the same subject are positively correlated. The censoring time is an independent uniform $(0, \widetilde{\tau})$ variable, where $\widetilde{\tau}$ was chosen to yield a desired number of uncensored observations. We assume $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t))'$ changes over time in two scenarios, $\overline{C} = \max\{C_1, \dots, C_n\}$ and C_i is the simulated follow-up or censoring time of the ith subject $(i = 1, \dots, n)$.

Scenario 1. $\beta_1(t) = 1_{t\in(0,0.5\overline{C})}$ and $\beta_2(t) = 1_{t\in(0.5\overline{C},\overline{C})}$;

Scenario 2. $\beta_1(t) = 1_{t\in(0,\overline{C}/3)} + 0.51_{t\in(2\overline{C}/3,\overline{C})}$ and $\beta_2(t) = 1_{t\in(\overline{C}/3,2\overline{C}/3)} - 0.51_{t\in(2\overline{C}/3,\overline{C})}$,

We considered the situations of independent observation times, and evaluated the accuracies of the estimated $\boldsymbol{\beta}(t)$ and change times. The weight function $W(\cdot)$ was set to 1. The accuracies of the estimator $\beta_j(t)$ were assessed by comparing the L_1 errors between the true $\boldsymbol{\beta}(t)$ and $\widehat{\boldsymbol{\beta}}(t)$, i.e.,

$$||\beta_j(t) - \widehat{\beta}_j(t)||_1 = \sum_{i=1}^{\bar{n}} ||\beta_j(t_i) - \widehat{\beta}_j(t_i)||_1 \cdot (t_i - t_{i-1})/\overline{C}, \qquad j = 1, 2,$$

where $t_0 = 0$. The hyperparameter in all simulations are estimated by maximizing the approximated likelihood (20) over a coarse grid of hyperparameter space. We also choose $\xi = 0.99$ such that $\overline{C}/L \in (0.002, 0.005)$ in all simulations, and use formula (19) with $(b, B) = (10, 20)$ to compute the estimate of $\boldsymbol{\beta}(t)$. For comparison purpose, we also consider an oracle estimate which assumes the change-times are known. In such case, we have two or three segments in each scenarios, and the estimate of $\boldsymbol{\beta}(t)$ in each segment can be estimated by (8) in Section 2.2 (since there are more than enough data in each segment, we use noninformative prior $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}^{-1} = \mathbf{0}$ in the oracle estimate). Table 1 summarizes the estimation error in $\boldsymbol{\beta}(t)$ with $a_0 = 0$, $a_1 = 1$, $\sigma_\phi = 0.2$ and $\sigma_\eta = 0.5$ for the two estimates, and each cell in the table is based on 1,000 simulations. The choices of $\widetilde{\tau} = 4$ and 10 yielded on average 2 and 10 observations per subject. Table 1 shows that the performance of our proposed estimates is comparable to that of the oracle estimate.

We also estimate the change times and compare them with the actual change times. Tables 2 and 3 summarizes the performance of estimated change times in Scenarios 1 and 2, respectively. In particular, the columns of $\widehat{M}_n$ gives the percentages of the estimated number of change-points, using the segmentation procedure in Section 2.6. Note that the percentage of getting right estimates of the number of change-point is more than 90% in the table. Since the maximal censoring time $\overline{C}$ in each simulation may be different, we use the relative measure $10^3|\widehat{\tau}_i - \tau_i|/\overline{C}$ to show how close the estimated and the true change-points are. The columns $10^3|\widehat{\tau}_i - \tau_i|/\overline{C}$ provide the average of the relative difference between the estimated and actual change times with their standard errors (in parentheses) for the case $\widehat{M}_n = 1$ in Scenario 1 and $\widehat{M}_n = 2$ in Scenario 2, respectively. Note that most of relative difference is less than 1%.

3.2 Performance under Misspecified Assumptions

We also investigate the performance of the proposed estimation procedure when the assumptions of model (1) are not satisfied. In particular, we consider marginal models (22) with four different assumptions for $\boldsymbol{\beta}(t)$ but same specifications for X_1, X_2 and $\epsilon(t)$ as in Section 3.1,

Model (A1). $\beta_1(t) = 1_{t\in(0,0.5\overline{C})}$ and $\beta_2(t) = 1_{t\in(0.5\overline{C},\overline{C})}$;

Model (A2). $\beta_1(t) = (t+0.1)^{1/3}$ and $\beta_2(t) = \sin(t+0.2)$ for $t \in (0, \overline{C})$;

Model (A3). $\beta_1(t) = 1_{t\in(0,0.5\overline{C})}$ and $\beta_2(t) = 1$ for $t \in (0, \overline{C})$;

Model (A4). $\beta_1(t) = 1_{t\in(0,0.75\overline{C})}$ and $\beta_2(t) = 1_{t\in(0.6\overline{C},0.95\overline{C})} - 0.51_{t\in(0.95\overline{C},\overline{C})}$.

Among these four models, (A1) is in the class of the proposed model, (A2) implies a regression model with smooth time-varying coefficients, for which the inference procedures based on smoothing techniques have been developed by Hoover et al. (1998) and Wu et al. (1998), and (A3) and (A4) consider the case that components of $\boldsymbol{\beta}(t)$ do not experience change-points at the same time. For comparison purpose, we assume $a_0 = 0, a_1 = 1, \sigma_\phi = 0.2$, $\sigma_\eta = 0.5$, $\sigma_\epsilon = 0.5$, $\tau = 10$ and $n = 100$ in the study.

We then simulate observations from models (A1)-(A4) and use the proposed procedure in Section 2.4 to estimate the regression coefficients $\boldsymbol{\beta}(t)$. Since model (A2) can also be estimated via nonparametric smoothing, we use the kernel smoothing approach in Wu et al. (1998) to estimate regression coefficients in models (A1)-(A4). Figure 1 shows the estimation results for models (A1)-(A4) using our procedure and nonparametric smoothing approach. We note that our estimation procedure, as incorporating the possibility of multiple change-points, works well in the four models, while regular smoothing method, if ignoring the existence of change-points, doesn't provide a satisfactory performance. The results for models (A3) and (A4) indicates that, when change-points are nonsynchronous and two change-points are too close, neither our methods nor nonparametric smoothing approach could locate the change-points correctly.

3.3 A Real Example

We now apply the proposed method to the AIDS Clinical Trial Group (ACTG) Protocol 016 study, a placebo-controlled trial on the safety and efficacy of zidovudine (AZT) in the treatment of patients with mildly symptomatic HIV infection (Fischl et al. 1990; Lin and Ying, 2001). A total of 351 patients were assigned to placebo and 360 to AZT. The study extended over 120 weeks, with a median follow-up of 48 weeks. The immunologic response was measured by the CD4 counts. All patients were supposed to be evaluated every two weeks for the first 16 weeks, and every month thereafter. The actual observation times deviated substantially from the schedule. The number of observations per patient ranged from 1 to 18, with a median of 8. It has been suggested that it is biologically plausible that the regression effects of AZT and baseline CD4 count are not constant over the entire study period. In particular, after the immune system starts responding to antiviral agents, one would expect the effect on the immune system of HIV patients will keep at a fixed level and last for certain periods, and then diminish after a certain point.

To address this concern, we fit models (1) and (2) to CD4 counts data. As commonly been done, we first apply the logarithm transformation to the CD4 counts and use the transformed values as the outcome $Y(t)$. We also consider two covariates $\mathbf{X} = (X_1, X_2)'$, where X_1 is the treatment indicator taking the values 1 versus 0 for AZT versus placebo, and X_2 is the (logarithmic) baseline CD4 count. Both X_1 and X_2 are time independent.

As patients' visit times are scheduled regularly, the observation times do not depend on covariates and hence we have $\boldsymbol{\gamma} = (0, 0)'$ and $\mathbf{Z} = (X_1, X_2)$ in (2). We then use the procedures in Sections 2.4 and 2.5 to estimate the hyperparameters and regression coefficients with $W(\cdot) = 1$ and each subperiod representing a week. We also consider a benchmark model, in which $Y(t)$ and $\mathbf{X}(t)$ are characterized by model (2) but coefficients $\boldsymbol{\beta}(t)$ are smooth function of time. We estimate this smooth time-varying model by the kernel methods in Wu et al. (1998). Figure 2 shows the estimated treatment indicator X_1 and baseline CD4 count X_2 effects (solid curve) with 95% confidence bands (dotted curves) from our model and the estimates from the smooth time-varying model (dashed curve). We find that, in the about first 70 weeks, the regression coefficients are almost constant (that is, the treatment effect is around 0.1, and the baseline effect is around 0.72). Indeed, we than use the segmentation procedure in Section 2.6 to get the similar segment and estimated regression coefficient for the segment. We note that, between the 90th and 100th weeks, the values of $\boldsymbol{\beta}(t)$ changes significantly, and for the period after the 100th week, it is difficult to give conclusive discussion as the sample size becomes much smaller. We also observe that the smoothing estimates are also within the confidence band given by the proposed model, indicating both methods can catch the trend of covariate effects, while it is easier to use the proposed approach to identify periods of covariates effect with constant level.

4. CONCLUSION

Longitudinal data analysis is an important statistical tool in the research of biological, health and social sciences. Existing regression analysis for longitudinal data assumes constant or smooth time-varying regression parameters over the study period. When the experimental environment undergoes abrupt changes or the purpose of the study is to locate periods in which the covariate effects are at different constant levels, a semiparametric change-point model is more appropriate.

For this purpose, the proposed model and developed estimation procedures provide a convenient tool for statistical inference. In particular, the semiparametric specification makes the proposed model sufficiently flexible for practical analysis, and multiple change-points setting allows one to perform intervention analysis and locate periods with different constant covariate effects. The developed estimation procedure is statistically and computationally attractive, as it combines the recent advance in semiparametric analysis based on counting process argument and multiple change-points inference. Furthermore, the developed procedure yields consistent and efficient estimates of covariates effect and change-points, which are also shown via simulation and empirical studies in the paper. The developed approach can be extended to the case of piecewise linear or quadratic regression coefficients and other statistical applications such as generalized linear models or survival analysis, for which we will discuss related issues in further studies.

ACKNOWLEDGEMENT

We thank gratefully the associate editor and two anonymous referees for their constructive comment on how to improve this paper.

APPENDIX

A.1 Asymptotic properties of $\widehat{\boldsymbol{\theta}}_{t_*,t^*}(\boldsymbol{\gamma})$ and $\widehat{\boldsymbol{\theta}}_{t_*,t^*}(\widehat{\boldsymbol{\gamma}})$ in Section 2.2

Note that the segmented estimating equations and procedures in Section 2.2 are essentially same as the ones in Lin and Ying (2001, Section 2) except the sample period in Section 2.2 is (t_*, t^*) instead of $(0, \infty)$ in Lin and Ying (2001). As Lin and Ying (2001, Appendix A.2) have provided detailed proof for the asymptotic normalities of $\widehat{\boldsymbol{\theta}}_{0,\infty}(\boldsymbol{\gamma})$ and $\widehat{\boldsymbol{\theta}}_{0,\infty}(\widehat{\boldsymbol{\gamma}})$, one simply replace the sample period $(0, \infty)$ in Lin and Ying's (2001, Appendix A.2) argument by (t_*, t^*) and is able to obtain the asymptotic normalities discussed in Remark 1.

A.2 Proof of posterior distributions in Section 2.3

We first derive (11) or the posterior distribution of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,s_l)}$. Let $f(\cdot|\mathcal{Y}_{(0,t_l]})$ denote the density function of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,s_l)}$. Note that conditional on $J_l = 1$ or 0, $f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_l)})$ is proportional to the sum of $p_l f(\boldsymbol{\beta}(s_l), \mathcal{Y}_{(s_{l-1},s_l)}|\mathcal{Y}_{(0,s_{l-1})}, J_l = 1)$ and $(1-p_l)f(\boldsymbol{\beta}(s_l), \mathcal{Y}_{(s_{l-1},s_l)}|\mathcal{Y}_{(0,s_{l-1}]}, J_l = 0)$. Note that $p_l f(\boldsymbol{\beta}(s_l), \mathcal{Y}_{(s_{l-1},s_l]}|\mathcal{Y}_{(0,s_{l-1}]}, J_l = 1) = \eta^*_{l,l} f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_l]}, J_l = 1) = \eta^*_{l,l} N(\widehat{\boldsymbol{\beta}}_{(s_{l-1},s_l)}, \boldsymbol{\Sigma}_{s_{l-1},s_l})$, where $\eta^*_{l,l} = pf(\mathcal{Y}_{(s_{l-1},s_l]}|\mathcal{Y}_{(0,s_{l-1}]}, J_l = 1) = p_l \psi_{s_l,s_l}$. Furthermore, some algebra shows that $(1-p_l)f(\boldsymbol{\beta}(s_l), \mathcal{Y}_{(s_{l-1},s_l)}|\mathcal{Y}_{(0,s_{l-1}]}, J_l = 0)$ is equivalent to $\sum_{m=1}^{l-1} \eta^*_{m,l-1} N(\widehat{\boldsymbol{\beta}}_{(s_{m-1},s_l)}, \boldsymbol{\Sigma}_{(s_{m-1},s_l)})$, where $\eta^*_{m,l-1} = (1-p_l)\eta_{m,l-1}\psi_{s_m,s_l}/\psi_{s_m,s_{l-1}}$. Hence combining above equations yields the posterior distribution (11) and (12) . (13) can be similarly proved by using the fact that $\{\boldsymbol{\beta}(s_l)\}$ can be considered as a reversible Markov chain with stationary distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

We next apply the Bayes theorem to obtain (15), i.e., the posterior distribution of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,s_L)}$. Let $f(\cdot|\mathcal{Y}_{(s_l,s_L)})$ and $f(\cdot|\mathcal{Y}_{(0,s_L)})$ denote the density functions of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(s_l,s_L)}$ and $\mathcal{Y}_{(0,s_L)}$, respectively, and let f denote the stationary density function of $\boldsymbol{\beta}(s_l)$ which is same as the prior Normal$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Making use of the assumption that outcomes are conditionally independent in the period (s_{l-1}, s_l) given covariates and $\boldsymbol{\beta}(s_l)$, we can show that $f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_L)}) \propto f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,s_l)})f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(s_l,s_L)})/f(\boldsymbol{\beta}(s_l))$, where $f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(0,t_l)})$ and $f(\boldsymbol{\beta}(s_l)|\mathcal{Y}_{(t_l,s_L)})$ are posterior distributions of $\boldsymbol{\beta}(s_l)$ given $\mathcal{Y}_{(0,t_l)}$ and $\mathcal{Y}_{(t_l,s_L)}$, respectively. Then we plug in (11), (13) and the Normal prior into the above equation, and use the fact $N(\widehat{\boldsymbol{\beta}}_{(s_{m-1},s_l)}, \boldsymbol{\Sigma}_{s_{m-1},s_l}) \cdot N(\widehat{\boldsymbol{\beta}}_{(s_l,s_k)}, \boldsymbol{\Sigma}_{s_l,s_k})/N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is equivalent to $\frac{\psi_{s_m,s_k}}{\psi_{s_m,s_l}\psi_{s_{l+1},s_k}} N(\widehat{\boldsymbol{\beta}}_{(s_{m-1},s_k)}, \boldsymbol{\Sigma}_{s_{m-1},s_k})$, we then obtain (15) and (16).

A.3 Asymptotic properties in Section 2.6

To prove Theorem 1, we first prove the following lemmas, which make the same assumptions as those in Theorem 1. The first three lemmas are used to analyze the BCMIX weights η_{ml}, $m \in \mathcal{K}_l$. A similar analysis yields corresponding results for the BCMIX weights $\eta_{k,l}$,

$k \in \widetilde{\mathcal{K}}_l$. Lemma 3 combines these results to provide the asymptotic behavior of the weights π_{mlk} in (18) which we use to prove Theorem 1.

The BCMIX weights η_{ml}, $m \in \mathcal{K}_l$, are difficult to analyze directly because they are defined recursively via (12) for which there is renormalization (from η^*_{ml} to η_{ml}) at every l. We shall approximate them by a more tractable version in Lemmas 2 and 3. To fix the idea, first assume that $B - b = 1$, which is tantamount to allowing one more change-point prior to t besides the most recent possibilities $s_l, \ldots, s_{l-b+1}$. We approximate the BCMIX by the AMOC ("at most one change") algorithm which assumes at most one change can occur before the "ancient" one of the $b+1$ components is removed from the mixture. First note that for $l-b+1 \le m \le l$ and $t \in (0, s_l]$, $P\{\boldsymbol{\beta}(t) \text{ is constant} | \mathcal{Y}_{(0,s_l)}\} \propto \psi_{1l} \prod_{j=1}^{l-1}(1-p_j)$, and $P\{\boldsymbol{\beta}(t) \text{ has only one jump at } s_m | \mathcal{Y}_{(0,s_l)}\} \propto \psi_{1,m-1}\psi_{ml}p_m \prod_{j \neq m}(1-p_j)$. Note that the sum of the above probabilities is $P\{\tau_1 \ge s_{l-b+1}, \tau_2 > s_m | \mathcal{Y}_{(0,s_l)}\}$, i.e., the conditional probability that there is at most one change up to the mth interval and that the change can only occur at intervals $s_{l-b+1}, \ldots, s_l$. Then we can estimate τ_1 by

$$t_1^{(n)} = \inf\left\{s_l : (1-p_{l-b+1})\psi_{1l} < p_{l-b+1}\psi_{1,l-b}\psi_{l-b+1,l}\right\}, \tag{23}$$

and then repeat the same procedure, with $\mathcal{Y}_{(0,s_l)}$ replaced by $\mathcal{Y}_{(t_1^{(n)},s_l)}$, to estimate τ_2 from $\mathcal{Y}_{(t_1^{(n)},s_l)}$. Proceeding inductively in this way yields the change-time estimates $t_1^{(n)} < t_2^{(n)} < \cdots$. Note that the AMOC filter weights for $s_l < t_1^{(n)}$ are

$$\eta^A_{1l} = (1-p_m)\psi_{1l}/P^A_l, \quad \eta^A_{jl} = p_m\psi_{1,m-1}\psi_{ml}/P^A_m \text{ for } l-b+1 \le m \le l, \tag{24}$$

where P^A_m is the normalizing constant to make the $b+1$ weights in $\{\eta^A_{jl}\}$ add up to 1. While keeping the most recent b indices as in BCMIX, (23) basically compares $P\{\boldsymbol{\beta}(t) \text{ is constant} | \mathcal{Y}_{(0,s_l)}\}$ and $P\{\boldsymbol{\beta}(t) \text{ has only one jump at interval } m-b | \mathcal{Y}_{(0,s_l)}\}$ and keeps the index with the larger posterior probability, analogous to BCMIX. The AMOC weights $\{\eta^A_{jl}\}$ for $t_k^{(n)} \le s_l < t_{k+1}^{(n)}$ are defined similarly, with $\mathcal{Y}_{(0,s_l)}$ replaced by $\mathcal{Y}_{(t_k^{(n)},s_l)}$. The following lemma gives the asymptotic properties of the AMOC weights.

Lemma 1. As $n \to \infty$, $P\{\max_{1\le j\le M} |t_j^{(n)} - \tau_j| \le b/L\} \longrightarrow 1$. *Moreover,* $\max_{s_l\le\tau_1} |\eta_{1,l}^{\mathrm{A}} - 1| \xrightarrow{P} 0$ *and* $\max_{t_j^{(n)}<s_l<\tau_{j+1}} |\eta_{t_j^{(n)}L,l}^{\mathrm{A}} - 1| \xrightarrow{P} 0$ *for* $1 \le j \le M-1$.

Proof. Note that some simple algebra shows that

$$\begin{aligned}\frac{\psi_{1,l-j}\psi_{l-j+1,l}p_{l-j+1}}{\psi_{1,l}(1-p_{l-j+1})} =&|\boldsymbol{\Sigma}|^{-1/2}\exp\Big\{-\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{\mu}\Big\}(e^{\lambda/L}-1)\Big|\boldsymbol{\Sigma}_{0,s_j}^{-1}\boldsymbol{\Sigma}_{0,s_l}\boldsymbol{\Sigma}_{s_{l-j+1},s_l}^{-1}\Big|^{-1/2}\\ &\cdot\exp\Big\{\frac{1}{2}(\boldsymbol{\beta}_{0,s_{l-j}}-\boldsymbol{\beta}_{s_{l-j+1},s_l})'\boldsymbol{\Sigma}_{0,s_j}^{-1}\boldsymbol{\Sigma}_{0,s_l}\boldsymbol{\Sigma}_{s_{l-j+1},s_l}^{-1}(\boldsymbol{\beta}_{0,s_{l-j}}-\boldsymbol{\beta}_{s_{l-j+1},s_l})\Big\}.\end{aligned} \tag{25}$$

First consider $s_l \le \tau_1 + 2b$. Note that for $n \to \infty$ such that $s_l - s_m > 0$ and $s_m > 0$, strong law of large number indicates that $n^{-1}\boldsymbol{\Sigma}_{0,s_j}^{-1}\boldsymbol{\Sigma}_{0,s_l}\boldsymbol{\Sigma}_{s_{l-j+1},s_l}^{-1}$ converges to a constant almost surely. We then have that $\max_{s_l\le\tau_1} |\eta_{1,l}^{\mathrm{A}} - 1| \xrightarrow{P} 0$ and $P\{t_1^{(n)} \le \tau_1\} \to 0$. For $s_l > \tau_1 + b$, we notice that the right hand side of (25) increases exponentially with rate n, hence $P\{t_1^{(n)} \le \tau_1 + b/L\} \to 1$ as $n \to \infty$. Replacing $\mathcal{Y}_{(0,s_l)}$ in the preceding argument by $\mathcal{Y}_{(\tau_1,s_l)}$ then proves the corresponding results for $t_2^{(n)}$ and $\max_{t_1^{(n)}<s_l<\tau_2} |\eta_{t_1^{(n)},l}^{A} - 1|$. Proceeding inductively in this way then completes the proof. $\square$

Note that the AMOC filter weights (24) can be represented recursively by using $J_*(l) = \max\{j \le l : I_j = 1\}$, as in the BCMIX approximation to (12) but with η_{jl}^A in place of η_{jl}. The analog of the set $\mathcal{K}_{l-1}$ for the AMOC filter is $\mathcal{K}_{l-1}^A = \{l-1, \ldots, l-(b\vee\lfloor t_{j*}^{(n)}L\rfloor), \lfloor t_{j*}^{(n)}L\rfloor\}$, where $\lfloor t_{j*}^{(n)}L\rfloor$ is the largest k that is $t_k^{(n)}L \le s_{m-1}$. Thus, the main difference between AMOC and the more flexible BCMIX is that AMOC allows one additional index $\lfloor t_{j*}^{(n)}L\rfloor$ to be included in $\mathcal{K}_{l-1}^A$ besides the most recent $l-1, \ldots, l-b$, while BCMIX allows $B-b$ more previous indices that need not be $\lfloor t_{j*}^{(n)}L\rfloor$, thereby removing the "at most one change" requirement. Whereas AMOC filter weights have the explicit formula (24) which plays an important role in the proof of Lemma 1, the recursive representation of BCMIX does not have a similar explicit formula. On the other hand, "at most one change" dominates "more than one change" in probability, and Lemma 1 and its proof can be used to prove the following lemma for BCMIX, in which we also weaken the assumption $B-b=1$ for AMOC to $1 \le B-b = O(1)$.

Lemma 2. As $n \to \infty$,

$$\max_{\lim_{n\to\infty} s_l < \tau_1} |\eta_{1l} - 1| \xrightarrow{P} 0. \tag{26}$$

Moreover, for $1 \le j \le M$,

$$\max_{\tau_j + b/L < s_l < \tau_{j+1}} \Big| \sum_{l \in \mathcal{K}_l, \tau_j \le s_k \le \tau_j + B/L} \eta_{kl} - 1 \Big| \xrightarrow{P} 0. \tag{27}$$

Proof. First note that (26) basically says that $\eta_{1,l-1}$ behaves asymptotically like $\eta^A_{1,l-1}$ so that even though BCMIX allows $B-b \ge 1$ (instead of $B = b+1$), the additional weights are negligible compared with $\eta_{1,m-1}$. Since the η_{kl} are defined recursively, (26) and (27) can be proved by induction on l. Concerning (27) with $j=1$, we can use an argument similar to that in Lemma 1 to show that the weight η_{1l} is eliminated by time $\tau_1 + \delta$, with probability approaching 1 as $n \to \infty$. For $s_l \ge \tau_1 + b/L$, the weight $\eta^A_{t_1^{(n)},l}$ in Lemma 1 is now replaced by the sum of weights η_{jl} in the set $\{j \in \mathcal{K}_l : \tau_1 \le s_j \le \tau_1 + B/L\}$. We can then modify the induction proof of Lemma 1 to prove (26) with $j=1$ for the range $\tau_1 + b/L \le s_l < \tau_2$, and then proceed to $j = 2, 3, \ldots, k$. $\square$

For the backward BCMIX filter, a time-reversal argument establishes the analogs of (26) and (27) for $\widetilde{\eta}_{kl}$. In particular, the analog of (26) is $\max_{s_l > \tau_k} |\widetilde{\eta}_{kl} - 1| \xrightarrow{P} 0$. Combining these results on the forward and backward BCMIX estimate weights via (15) yields the following

Lemma 3. For $0 \le h \le M$, we have $\max_{s_l < \tau_1} \Big| \sum_{k \in \widetilde{\mathcal{K}}_m, \tau_1 - B/L \le s_k \le \tau_1} \xi_{1lk} - 1 \Big| \xrightarrow{P} 0$, $\max_{s_l > \tau_M} \Big| \sum_{j \in \mathcal{K}_l, \tau_M \le s_j \le \tau_k + B/L} \xi_{jlL} - 1 \Big| \xrightarrow{P} 0$, and

$$\max_{\tau_h + b/L \le t \le \tau_{h+1} - b/L} \Big| \sum_{j \in \mathcal{K}_l, k \in \widetilde{\mathcal{K}}_l, \tau_h \le s_j \le \tau_h + B/L, \tau_{h+1} - B/L \le s_k \le \tau_{h+1}} \xi_{jlk} - 1 \Big| \xrightarrow{P} 0.$$

Proof of Theorem 1. From (24) and asymptotic results on the limit of $n^{-1}\boldsymbol{\Sigma}_{t_*,t^*}^{-1}$ and the estimating function (7), it follows that for $1 \le h \le M$, $\max_{\tau_h \le s_j \le \tau_h + B/L, \tau_{h+1} - B/L \le s_l \le \tau_{h+1}} ||\widetilde{\boldsymbol{\beta}}_{s_j,s_l} - \boldsymbol{\beta}_{h+1}|| \xrightarrow{P} 0$, where $\boldsymbol{\beta}_{h+1}$ is the true value of $\boldsymbol{\beta}(t)$ in the interval (τ_h, τ_{h+1}), and that $\max_{\tau_1 - B/L \le s_l \le \tau_1} ||\widetilde{\boldsymbol{\beta}}_{0,s_l} - \boldsymbol{\beta}_1|| \xrightarrow{P} 0$. Combining this with Lemma 3, (C1) and (19) then yields the desired conclusion. □

Proof of Theorem 2. Let $\delta_h^{(n)} = ||\boldsymbol{\beta}_h - \boldsymbol{\beta}_{h-1}||$ and order them as $\delta_{(1)}^{(n)} \ge \delta_{(2)}^{(n)} \ge \cdots \ge \delta_{(M)}^{(n)}$. This ordering induces a corresponding ordering $\tau_{(h)}$ of the τ_h; as observations are in continuous time, we assume that there are no ties. Recall that the $\widehat{\nu}_h$ are defined by (21) which involves $\Delta_m = ||\widehat{\boldsymbol{\beta}}(t_{m-w}) - \widehat{\boldsymbol{\beta}}(t_{m+w})||^2$. By the segmentation procedure in Section 2.3, $\Delta_m = ||\boldsymbol{\beta}(t_{m-w}) - \boldsymbol{\beta}(t_{m+w})||^2 + o_p(1)$, in which the $o_p(1)$ term is uniform in $t \in \{s_1, \ldots, s_L\}$ such that $\min_{1 \le k \le M} |s_m - \tau_k| \ge b/L$ and hence $\max_{1 \le h \le M} \left| \nu_h - \tau_{(h)} \right| \Big/ (b/L) \xrightarrow{P} 0$. Making use of this result, putting $\psi_{\lfloor \nu_{k,n} L \rfloor, \lfloor \nu_{k+1,n} L \rfloor}$, which is given explicitly by (9), into the right hand side of (20) that defines $\ell_n(M)$ and applying the law of large numbers to $\ell_n(h)$ with $h < M$, and the functional central limit theorem to $\ell_n(h)$ with $h \ge M$, we conclude that $\widehat{M}_n \xrightarrow{P} M$.

Proof of Theorem 3. The interval $(v_{k-1,n} + c_n B/L, v_{k,n} - c_n B/L)$ approaches to (τ_{k-1}, τ_k) with probability 1, then the estimate $\widetilde{\boldsymbol{\beta}}_{v_{k-1,n} + c_n B/L, v_{k,n} - c_n B/L}$ has the similar asymptotic properties as $\widehat{\boldsymbol{\beta}}_{(\tau_{k-1}, \tau_k)}$, which yields Theorem 3.

REFERENCES

Asgharian, M. and Wolfson, D. B. (2001), "Covariates in Multipath Change-Point Problems: Modelling and Consistency of the MLE," *The Canadian Journal of Statistics*, 29, 515-528.

Albert, J. H. and Chib, S. (1993), "Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts," *Journal of Business,* *Economics, and Statistics*, 11, 1-15.

Altman, N. S., and Casella, G. (1995), "Nonparametric Empirical Bayes Growth Curve Analysis," *Journal of the American Statistical Association*, 90, 508-515.

Bai, J. (1997a), "Estimation of a Changepoint in Multiple Regression Models," *Review of Economics and Statistics*, 79, 551-563.

—— (1997b), "Estimation Multiple Breaks One at a Time," *Econometric Theory*, 13, 315-352.

Bai, J. and Perron, P. (1998), "Testing for and Estimation of Multiple Structural Changes," *Econometrica*, 66, 817-858.

Bai, J. and Perron, P. (2003), "Computation and Analysis of Multiple Structural Change Models," *Journal of Applied Econometrics*, 18, 1-22.

Birgé, L. and Massart, P. (2001), "Gaussian Model Selection," *Journal of European Mathematical Society*, 3, 203-268.

Broman, K. W. and Speed, T. P. (2002), "A model Selection Approach for the Identification of Quantitative Trait Loci in Experiment Crosses," *Journal of Royal Statistical Society, Ser.* B., 64, 641-656.

Chib, S. (1998), "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics*, 86, 221-241.

Chib, S., Nardari, F. and Shephard, N. (2002), "Markov Chain Monte Carlo Methods for Stochastic Volatility Models," *Journal of Econometrics*, 108, 281-316.

Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187-220.

Davidian, M., and Giltinan, D. M. (1995), *Nonlinear Models for Repeated Measurement Data*, London: Chapman and Hall.

Davis, R., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006), "Structural Break Estimation for Nonstationary Time Series Models," *Journal of American Statistical Association*, 101, 223-239.

Desmond, R. A., Weiss, H. L., Arani, R. B., Soong, S., Wood, M. J., Fiddlan, P. A. Gnann, J. W. and Whitley, R. J. (2002). "Clinical Applications for Change-Point Analysis of Herpers Zoster Pain," *Journal of Pain and Symptom Management*, 23, 510-516.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*. 2nd ed. Oxford: Oxford University Press.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Fan, J., and Zhang, J. T. (2000), "Functional Linear Models for Longitudinal Data," *Journal of the Royal Statistical Society*, Ser. B, 62, 303-322.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711-732.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data," *Biometrika*, 85, 809–822.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements," *Biometrika*, 89, 111-128.

—— (2004), "Polynomial Spline Estimation and Inference for Varying Coefficient Models with Longitudinal Data," *Statistica Sinica*, 14, 763-788.

James, G., Wang, J. and Zhu, J. (2009), "Functional Linear Regression That's Interpretable," *The Annals of Statistics*, 37, 2083-2108.

Lai, T. L., Liu, H. and Xing, H. (2005), "Autoregressive Models with Piecewise Constant Volatility and Regression Parameters," *Statistica Sinica*, 15, 279-301.

Lai, T. L., and Xing, H. (2011), "A Simple Bayesian Approach to Multiple Change-Points," *Statistica Sinica*, in press.

Lai, T. L., Xing, H. and Zhang, N. (2008), "Stochastic Segmentation Models for Array-Based Comparative Genomic Hybridization Data Analysis," *Biostatistics*, 9, 290-307.

Laird, N. M., Donnelly, C., and Ware, J. H. (1992), "Longitudinal Studies With Continuous Responses," *Statistical Methods in Medical Research*, 1, 225-247.

Lavielle, M. (2005), "Using Penalized Contrasts for the Change-Point Problem," *Signal Processing*, 85, 1501-1510.

Lawless, J. F. and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158-168.

Li, Y. (2011), "Efficient Semiparametric Regression for Longitudinal Data with Nonparametric Covariance Estimation," *Biometrika*, 98, 355-370.

Lin, D. Y., and Ying, Z. (2001), "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data," *Journal of the American Statistical Association*, 96, 103-113.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000), "Semiparametric Regression for the Mean and Rate Functions of Recurrent Events," *Journal of the Royal Statistical Society*, Ser. B, 62, 711-730.

Lin, X., and Carroll, R. J. (2000), "Nonparametric function estimation for clustered data when the predictor is measured without/with error," *Journal of the American Statistical Association*, 95, 520-534.

—— (2001), "Semiparametric regression for clustered data using generalized estimating equations," *Journal of the American Statistical Association*, 96, 1045-1056.

—— (2006), "Semiparametric Estimation in General Repreated Measures Problems," *Journal of the Royal Statistical Society*, Ser. B, 68, 69-88.

Liu, J. S. and Lawrence, C. E. (1999), "Bayesian Inference on Biopolymer Models," *Bioinformatics*, 15, 38-52.

Moyeed, R. A. and Diggle, P. J. (1994), "Rates of Convergence in Semi-parametric Modeling of Longitudinal Data," *Australian Journal of Statistics*, 36, 75–93.

Nesselroade, J. R. and Baltes, P. B. (eds.) (1979), *Longitudinal Research in the Study of Behavior and Development*, New York: Academic Press.

Pepe, M. S. and Cai, J. (1993), "Some Graphical Displays and Marginal Regression Analyses for Recurrent Failure Times and Time Dependent Covariates," *Journal of the American Statistical Association*, 88, 811-820.

Qu, Z. and Perron, P. (2007), "Estimating and Testing Structural Changes in Multivariate Regressions," *Econometrica*, 75, 459-502.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004), "Circular Binary Segmentation for the Analysis of Array-based DNA Copy Number Data," *Biostatistics*, 5, 557-572.

Rice, J. A. and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society*, Ser. B, 53, 233-243.

Rosenfield, D., Zhou, E., Wihelm, F. H., Conrad, A., Roth, W. T. and Meuret, A. E. (2010). "Change Point Analysis for Longitudinal Physiological Data: Detection of Cardiorespiratory Changes Preceding Panic Attacks," *Biological Psychology*, 84, 112-120.

Schemper, M., Wahounig, S., and Heinze, G. (2009). "The estimation of average hazard ratios by weighted Cox regression," *Statistics in Medicine*, 28, 2473-2489.

Siegmund, D. (2004). "Model Selection in Irregular Problems: Applications to Mapping Quantitative Trait Loci," *Biometrika*, 91, 785-800.

Silverman, B. W. (1996),"Smoothed Functional Principal Components by Choice of Norm," *The Annals of Statistics*, 24, 1-24.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Vostrikova, L. J. (1981), "Detecting Disorder in Multidimensional Random Processes," *Soviet Mathematics Doklady*, 24, 55-59.

Wang, J. and Zivot, E. (2000), "A Bayesian Time Series Model of Multiple Structural Changes in Level, Trend, and Variance," *Journal of Business, Economics and Statistics*, 18, 374-386.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data.* Cambridge and London: MIT Press.

Wu, C. O., Chiang, C. T. and Hoover, D. R. (1998), "Asymptotic Confidence Regions for Kernel Smoothing of a Time-Varying Coefficient Model with Longitudinal Data, " *Journal of the American Statistical Association*, 88, 1388-1402.

Wu, C. O., Tian, X. and Jiang, W. (2011), "A Shared Parameter Model for the Estimation of Longitudinal Concomitant Intervention Effects," *Biostatistics*, doi:10.1093/biostatistics/kxq084

Xing, H., Sun, N. and Chen, Y. (2012), "Credit Rating Dynamics in the Presence of Unknown Structural Breaks," *Journal of Banking and Finance*, 36, 78-89.

Xu, R. and O'Quigley, J. (2000). "Estimating average regression effect under non-proportional hazards," *Biostatistics*, 1, 423-439.

Xue, L. and Zhu, L. (2007), "Empirical Likelihood for a Varying Coefficient Model with Longitudinal Data," *Journal of the American Statistical Association*, 102, 642-654.

Yao, Y. (1984), "Estimation of a Noisy Discrete-Time Step Functions: Bayes and Empirical Bayes Approach," *The Annals of Statistics*, 12, 1434-1447.

Yao, Y. (1988), "Estimating the Number of Change-Points via Schwarz Criterion," *Statistics and Probability Letters*, 6, 181-189.

Zeger, S. L. and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.

Zhang, N. and Siegmund, D. (2006), "A Modified Bayes Information Criterion with Applications to Comparative Genomic Hybridization Data," *Biometrics*, 63, 22-32.

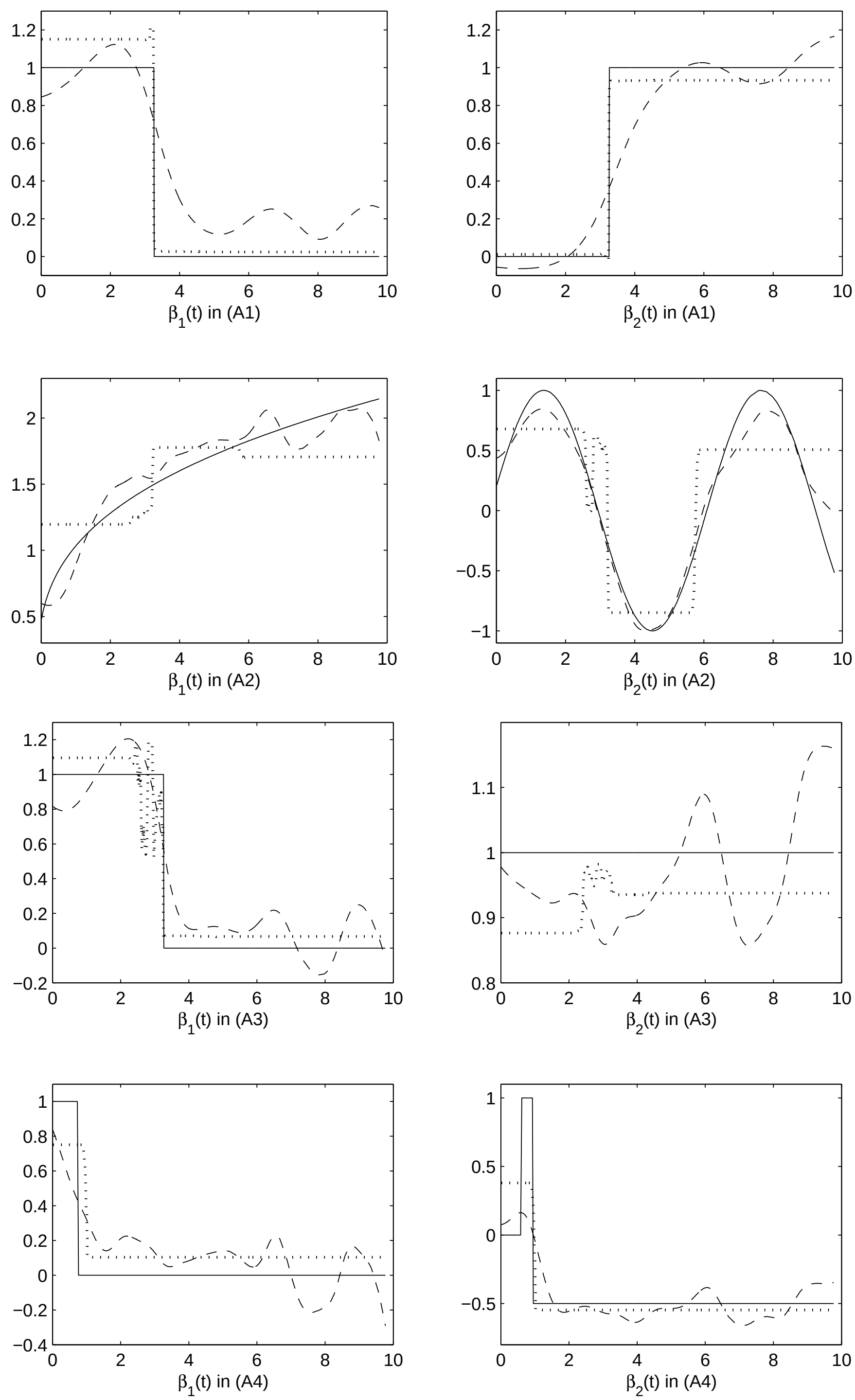

Figure 1: Regression coefficients $\beta_1(t)$ (left) and $\beta_2(t)$ (right) in models (A1)-(A4) (True value: Solid; Our estimate: Dotted; Smoothing estimate: Dashed).

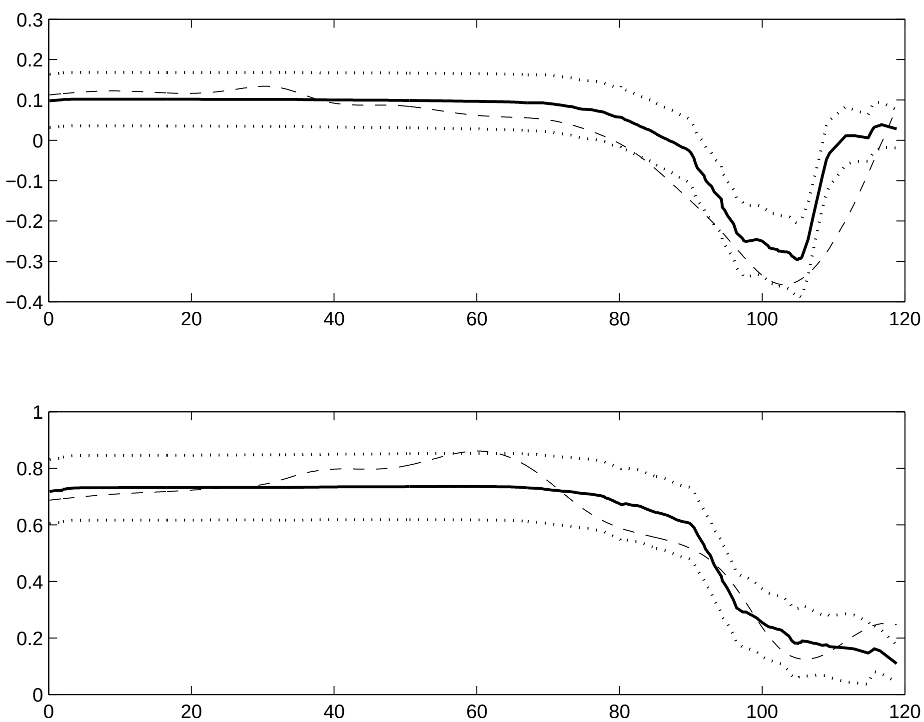

Figure 2: The estimated treatment indicator X_1 (top) and baseline CD4 count X_2 (bottom) effects with 95% confidence bands (Estimates: Solid; Confidence bands of the estimates: Dotted; Estimates from a smooth time-varying model: Dashed).

Table 1: Summary Statistics for Scenarios 1 and 2

n	$\alpha(t)$	σ_ϵ	$\widetilde{\tau}$	Oracle estimate		Our estimate	
				Bias1	Bias2	Bias1	Bias2
				Scenario 1			
100	$\sin t$	0.5	4	0.129 (0.002)	0.070 (0.001)	0.117 (0.002)	0.069 (0.001)
		0.5	10	0.095 (0.002)	0.048 (0.001)	0.105 (0.002)	0.055 (0.001)
		1	4	0.176 (0.003)	0.095 (0.002)	0.193 (0.004)	0.126 (0.003)
		1	10	0.131 (0.002)	0.067 (0.001)	0.141 (0.002)	0.080 (0.001)
	$\sqrt{t}$	0.5	4	0.112 (0.002)	0.059 (0.001)	0.120 (0.002)	0.071 (0.001)
		0.5	10	0.085 (0.002)	0.043 (0.001)	0.088 (0.002)	0.047 (0.001)
		1	4	0.178 (0.003)	0.096 (0.002)	0.196 (0.003)	0.128 (0.003)
		1	10	0.124 (0.002)	0.063 (0.002)	0.130 (0.002)	0.073 (0.001)
200	$\sin t$	0.5	4	0.081 (0.001)	0.040 (0.001)	0.086 (0.001)	0.045 (0.001)
		0.5	10	0.067 (0.001)	0.033 (0.001)	0.073 (0.001)	0.037 (0.001)
		1	4	0.131 (0.002)	0.065 (0.001)	0.138 (0.002)	0.075 (0.001)
		1	10	0.092 (0.002)	0.045 (0.001)	0.099 (0.002)	0.053 (0.001)
	$\sqrt{t}$	0.5	4	0.080 (0.001)	0.040 (0.001)	0.085 (0.002)	0.046 (0.001)
		0.5	10	0.060 (0.001)	0.030 (0.001)	0.061 (0.001)	0.032 (0.001)
		1	4	0.130 (0.002)	0.064 (0.001)	0.136 (0.002)	0.075 (0.001)
		1	10	0.087 (0.002)	0.043 (0.001)	0.089 (0.002)	0.047 (0.001)
				Scenario 2			
100	$\sin t$	0.5	4	0.149 (0.002)	0.098 (0.001)	0.138 (0.002)	0.096 (0.002)
		0.5	10	0.104 (0.002)	0.052 (0.001)	0.108 (0.002)	0.062 (0.001)
		1	4	0.215 (0.003)	0.117 (0.002)	0.234 (0.003)	0.267 (0.005)
		1	10	0.152 (0.002)	0.077 (0.001)	0.158 (0.002)	0.103 (0.001)
	$\sqrt{t}$	0.5	4	0.136 (0.002)	0.073 (0.001)	0.145 (0.002)	0.100 (0.002)
		0.5	10	0.099 (0.002)	0.050 (0.001)	0.101 (0.002)	0.059 (0.001)
		1	4	0.219 (0.003)	0.119 (0.002)	0.239 (0.003)	0.269 (0.005)
		1	10	0.148 (0.002)	0.075 (0.001)	0.154 (0.002)	0.098 (0.002)
200	$\sin t$	0.5	4	0.096 (0.001)	0.047 (0.001)	0.100 (0.001)	0.059 (0.001)
		0.5	10	0.073 (0.001)	0.037 (0.001)	0.075 (0.001)	0.042 (0.001)
		1	4	0.159 (0.002)	0.078 (0.001)	0.164 (0.002)	0.108 (0.002)
		1	10	0.106 (0.002)	0.053 (0.001)	0.110 (0.002)	0.065 (0.001)
	$\sqrt{t}$	0.5	4	0.097 (0.001)	0.048 (0.001)	0.102 (0.001)	0.060 (0.001)
		0.5	10	0.069 (0.001)	0.035 (0.001)	0.070 (0.001)	0.038 (0.001)
		1	4	0.159 (0.002)	0.078 (0.001)	0.165 (0.002)	0.109 (0.002)
		1	10	0.101 (0.002)	0.052 (0.001)	0.105 (0.002)	0.062 (0.001)

NOTE: Bias1 and Bias2 represent the $L1$ estimation bias for β_1 and β_2, the standard errors are shown in parentheses.

Table 2: Performance of $|\widehat{\tau}_1 - \tau_1|/\overline{C}$ and $\widehat{M}_n$ in Scenario 1

			$n = 100$			$n = 200$		
$\alpha(t)$	σ_ϵ	$\widetilde{\tau}$	$10^3\lvert\widehat{\tau}_1 - \tau_1\rvert/\overline{C}$	$\widehat{M}_n = 1$	$\widehat{M}_n = 2$	$10^3\lvert\widehat{\tau}_1 - \tau_1\rvert/\overline{C}$	$\widehat{M}_n = 1$	$\widehat{M}_n = 2$
$\sin t$	0.5	4	9.32 (0.44)	96.8	3.2	6.39 (0.46)	96.8	3.2
	0.5	10	6.22 (0.45)	83.5	16.5	3.91 (0.40)	85.0	15.0
	1	4	19.5 (0.95)	96.4	3.6	10.9 (0.59)	98.2	1.8
	1	10	11.2 (0.67)	91.8	8.2	5.57 (0.34)	92.0	8.0
$\sqrt{t}$	0.5	4	11.1 (0.66)	93.9	6.1	7.03(0.57)	90.3	9.7
	0.5	10	8.17 (0.86)	89.2	10.8	5.89 (0.79)	86.1	13.9
	1	4	20.6 (0.99)	96.2	3.8	10.6 (0.54)	97.7	2.3
	1	10	9.20 (0.54)	90.3	9.7	5.37 (0.51)	89.6	10.4

Table 3: Performance of $|\widehat{\tau}_k - \tau_k|/\overline{C}$ and $\widehat{M}_n$ in Scenario 2

n	$\alpha(t)$	σ_ϵ	$\widetilde{\tau}$	$10^3\lvert\widehat{\tau}_1 - \tau_1\rvert/\overline{C}$	$10^3\lvert\widehat{\tau}_2 - \tau_2\rvert/\overline{C}$	$\widehat{M}_n = 1$	$\widehat{M}_n = 2$	$\widehat{M}_n = 3$
100	$\sin t$	0.5	4	8.80 (0.39)	7.12 (0.31)	1.3	98.7	0.0
		0.5	10	3.74 (0.18)	3.24 (0.14)	0.1	99.1	0.8
		1	4	15.6 (0.76)	15.4 (0.74)	7.4	91.1	1.5
		1	10	7.54 (0.38)	6.10 (0.40)	1.7	97.8	5
	$\sqrt{t}$	0.5	4	9.69 (0.44)	7.60 (0.34)	2.0	97.8	0.2
		0.5	10	3.43 (0.15)	2.86 (0.12)	0.5	99.5	0.0
		1	4	17.2 (0.87)	15.8 (0.78)	7.7	91.1	1.2
		1	10	7.34 (0.33)	6.11 (0.39)	1.4	98.2	0.4
200	$\sin t$	0.5	4	4.29 (0.21)	2.97 (0.13)	0.3	99.7	0.0
		0.5	10	2.21 (0.34)	2.02 (0.22)	0.3	99.4	0.3
		1	4	8.12 (0.42)	5.97 (0.33)	1.7	97.9	0.4
		1	10	4.50 (0.33)	4.54 (0.66)	2.0	97.7	3
	$\sqrt{t}$	0.5	4	4.65 (0.21)	3.18 (0.14)	0.3	99.7	0.0
		0.5	10	1.69 (0.10)	1.36 (0.07)	0.2	99.8	0.0
		1	4	9.02 (0.58)	6.37 (0.32)	1.3	98.7	0.0
		1	10	3.61 (0.23)	3.74 (0.54)	1.7	98.0	0.3