Published in final edited form as:
Stat Interface. 2018 ; 11(2): 203–221. doi:10.4310/SII.2018.v11.n2.a1.

Time-varying copula models for longitudinal data

Esra Kürüm,
Department of Statistics, University of California Riverside, Riverside, CA 92521, USA

John Hughes[†],
Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

Runze Li, and
Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA

Saul Shiffman
Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA

Abstract

We propose a copula-based joint modeling framework for mixed longitudinal responses. Our approach permits all model parameters to vary with time, and thus will enable researchers to reveal dynamic response–predictor relationships and response–response associations. We call the new class of models TIMECOP because we model dependence using a time-varying copula. We develop a one-step estimation procedure for the TIMECOP parameter vector, and also describe how to estimate standard errors. We investigate the finite sample performance of our procedure via three simulation studies, one of which shows that our procedure performs well under ignorable missingness. We also illustrate the applicability of our approach by analyzing binary and continuous responses from the Women's Interagency HIV Study and a smoking cessation program.

Keywords and phrases

Bimodal kernel; HIV; Joint model; Local regression; Varying coefficient model

AMS 2000 subject classifications

Primary 62G08; secondary 62H20

1. INTRODUCTION

Analyses of multivariate longitudinal outcomes are now common, but current models for such data cannot reveal the nature of time-varying dependence among the coordinates of a *d*-

Correspondence to: Esra Kürüm.

[†] Drs Kürüm and Hughes contributed equally to this work.

dimensional response unless $d = 2$ and the two processes in question are Bernoulli and Gaussian [43]. What is needed is a modeling approach that is flexible enough to permit the estimation of time-varying parameters for a response of higher dimension, having coordinate processes of practically any type. In this paper we develop a modeling framework that addresses these concerns.

A number of joint models for longitudinal binary and continuous responses are well established [see, for example, 7, 10, 26, 61, 57, 16, 29, 47, 43]. The main challenge in developing such models is that there is no natural multivariate distribution for responses of mixed type. One way to overcome this problem is to introduce a latent variable underlying the binary response, and assume that the continuous response and the latent variable are jointly Gaussian. The resulting joint distribution can then be factored, leading to one of two formulations: (1) a marginal distribution for the continuous response along with a conditional distribution for the binary response given the continuous response, or (2) a marginal distribution for the binary response along with a conditional distribution for the continuous response given the binary response. Some approaches of this sort are not limited to binary and continuous outcomes, but can also handle other types of outcomes and $d > 2$ [see, for example, 61, 16].

A second solution to the above mentioned problem is the joint mixed-effects model [28, 29]. In this approach a random effect is assumed for each outcome, and the outcomes are associated via a joint distribution for the random effects. As pointed out by Verbeke et al. [73], fitting these models becomes computationally burdensome as the number of outcomes increases, and maximum likelihood estimation is possible only when the dimension is low or strong assumptions are made. An example of the latter can be found in Roy and Lin [58], where corresponding random effects for various outcomes are assumed to be perfectly correlated. Another potential drawback of the mixed-effects approach is confounding [34], which may inflate the variance of fixed-effects estimators, preventing the discovery of important response–predictor relationships.

A third solution is to join a set of marginal distributions using a copula. (See Nelsen [52] for an introduction to copulas, and de Leon and Chough [13], Heinen and Rengifo [33], Madsen and Fang [49], Masarotto and Varin [51], Smith et al. [67], Song et al. [71], Wu and de Leon [74] for information on copula-based regression models.) This is the approach we adopt for the remainder of this article.

In a longitudinal study the relationship between a response and predictors, or the association between a pair of responses, may change over time. The inability of ordinary models to capture these dynamic patterns led Kürüm et al. [43] to develop time-varying models [3] for longitudinal binary and continuous responses. Kürüm et al. [43] adopted the latent variable approach and first type of factorization described above. This implies a two-step estimation procedure. In the second stage of the procedure, the association between the responses takes the form of a time-varying regression coefficient. The advantage of this method is that it allows all parameters, including association parameters, to be time varying. But the method of Kürüm et al. [43] relies on the assumption that the latent variable and the continuous response follow a joint normal distribution. Moreover, their method is limited to estimating the time-varying association between only two longitudinal outcomes.

In this paper we develop a flexible and intuitive framework for modeling multivariate longitudinal data. Although we focus on revealing time-varying dependence relationships, our framework can easily accommodate all manner of time-varying parameters for the coordinate processes: regression coefficients, variances, etc. To achieve this goal, we exploit the modularity of copula-based modeling, which allows us to model the marginal distributions and dependence structure separately before joining them by way of the probability integral transform [66]. In our view, this is perhaps the simplest and most natural way to construct a multivariate distribution for discrete or mixed responses, and is even more intuitive when the responses are continuous.

Our main contribution is twofold. First, our approach brings time-varying parameters to multivariate longitudinal modeling. This will allow researchers to uncover complex dynamic patterns of dependence and response–predictor relationships. Second, our approach brings arbitrary response type and dimension greater than two to time-varying joint modeling, i.e., our approach, unlike the approach of Kürüm et al. [43], is not limited to a binary–continuous response. Moreover, our model requires neither latent variables nor factorization, and so does not rely on the assumptions required by Kürüm et al. [43] to estimate time-varying dependence parameters.

Our motivating data were collected as part of the National Institutes of Health-funded Women's Interagency HIV Study (WIHS). The WIHS was spurred by alarming trends:

- Between 1990 and 1994, the rate of increase in AIDS cases reported for women (89%) was three times that for men (29%).
- The 1994 rate of AIDS cases among African-American women was twice that for Hispanic women and 17 times that for white women.
- By 1995, HIV infection had become the third leading cause of death among U.S. women between the ages of 25 and 44, and the leading cause of death among African-American women in this age group.

Although the Multicenter AIDS Cohort study (MACS) [39], a 10 year-long study of 5,000 homosexual/bisexual men, 90% of whom were white, contributed much to our understanding of HIV progression, the progression of HIV to AIDS, and survival after AIDS, the increasing rate of AIDS among women necessitated a similar longitudinal study for women and communities of under-represented race. For this reason, the WIHS was established to study the impact of HIV infection in U.S. women. One of many objectives of the program is to investigate nutritional, socioeconomic, and behavioral risk factors that may be related to the rate of disease progression [2].

Researchers have taken a special interest in the smoking behavior of HIV patients because it is known that smoking affects the immune system [25, 31, 36, 27]. But the findings are inconsistent. For instance, Nieman et al. [54] found that smokers progressed to AIDS more rapidly than nonsmokers, whereas the analyses performed by Galai et al. [27] and Burns et al. [5] on two different data sets found no difference between smokers and nonsmokers in the risk of developing AIDS. We will investigate this matter further by applying our model to two WIHS variables: CD4 cell percentage (a measure of HIV progression) and smoking status.

Our analysis will differ from the above mentioned analyses in two important ways. First, our analysis will not employ survival methods, for our data are not censored, and we do not wish to define a patient lifetime that ends when the patient has progressed to AIDS. We wish to explore the dynamics of the relationship between CD4 cell percentage and smoking status throughout the study. Second, some of the above mentioned analyses excluded patients who altered their smoking behavior during the study. It is not unusual for behaviors to change during a longitudinal study, and ignoring these changes may lead to biased results.

We will treat CD4 percentage and smoking status as response variables so that we can estimate the time-varying *partial* association [8] between them, i.e., we aim to reveal the association between CD4 percentage and smoking status conditional on predictors of interest, one of which is shared by the outcomes. We could regress one of CD4 percentage and smoking status on the other, but it is not obvious how to assign the role of response or predictor to either variable. (In HIV studies it is customary to regress CD4 cell percentage on smoking status, assuming that smoking leads to a change in CD4 cell percentage. But Mamary et al. [50] showed that the percentage of smokers among HIV patients is higher than the percentage of smokers in the general adult population. In addition, HIV patients tend to have a pessimistic view of their survival, which might result in lack of motivation to quit smoking or even an increase in cigarette consumption. These results suggest that being an HIV patient could be predictive of smoking, which implies that one should regress smoking status on CD4 cell percentage.) And a regression model would not allow us to reveal the nature of this association having controlled for the predictors, such as depression, nor would it allow us to reveal response–predictor relationships for both variables. We could of course achieve the latter goal by fitting two univariate regression models, but this would provide no information about the association, which is of interest to us. Moreover, joint modeling can lead to considerably more precise estimators [28, 70]. This gain in efficiency grows as the strength of dependence increases, especially for smaller samples.

We will also apply TIMECOP to binary and continuous responses from a smoking cessation study. Several studies have focused on understanding the motivation for smoking so that more successful smoking cessation programs can be designed. The intuitive link between urge to smoke and smoking, and the importance of urges in some theories of smoking, makes urge to smoke interesting to prevention scientists [64]. Besides urge to smoke (our continuous response), the data set contains a number of additional outcomes, including alcohol consumption, coffee consumption, presence of others smoking, and food consumption (all binary). We used the latter four variables to create our binary response, since it has in fact been observed that these stimuli increase the odds of smoking [17, 63, 37, 64].

The relationship between the above mentioned stimuli and smoking (and therefore, perhaps, urge to smoke) might vary over the course of the study, particularly before and after a subject quits smoking. However, previous studies ignored the possible changes in this relationship over time. Our main goal is to study the time-varying partial association between these factors and urge to smoke using our joint modeling approach. Exploring the dynamics of this association would help prevention scientists to achieve their goal of designing smoking cessation programs with high success rates.

The remainder of the paper is organized as follows. Section 2 describes our time-varying copula model for longitudinal mixed outcomes; our estimation procedure, which uses local regression techniques; and bandwidth selection. Section 3 assesses the finite sample behavior of our approach via simulation, and shows that our procedure can handle ignorable missingness. Sections 4 and 5 apply our method to a subset of the WIHS data and to smoking cessation data, respectively. And Section 6 contains concluding remarks.

2. TIME-VARYING COPULA MODELS

In longitudinal and ecological studies, response–predictor associations may change with time, temperature, or geographical location. Varying coefficient models, which allow regression parameters to change with some underlying covariate(s), were developed to address the inability of ordinary regression models to capture these dynamic relationships.

Varying coefficient models were introduced by Cleveland et al. [9] and popularized by Hastie and Tibshirani [32]. A linear varying coefficient model takes the form

$$Y = \boldsymbol{x}'\boldsymbol{\beta}(U) + \varepsilon, \quad (1)$$

where Y is the response variable, $\boldsymbol{x} = (x_1, \ldots, x_p)'$ is a vector of predictors, U is a scalar covariate, $\boldsymbol{\beta}(U) = (\beta_1(U), \ldots, \beta_p(U))'$ are unknown coefficient functions, and $\boldsymbol{\varepsilon}$ is an error such that $\mathbb{E}(\boldsymbol{\varepsilon} \,/\boldsymbol{x}, U) = 0$. Since varying coefficient models are local linear models [22], kernel smoothing is a natural approach to estimation for model (1), and so we use kernel smoothing for the class of models developed below: time-varying copula models for longitudinal data, TIMECOP for short. The TIMECOP framework permits not only regression coefficients but all parameters, including dependence parameters, to be time varying.

Suppose we have m independent subjects. For subject i we observe the d-variate process $\boldsymbol{Y}_i(t) = (Y_{i1}(t), \ldots, Y_{id}(t))'$ at random times $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})'$. That is, we observe $\boldsymbol{Y}_{ij} \equiv \boldsymbol{Y}_i(t_{ij}) = (Y_{i1}(t_{ij}), \ldots, Y_{id}(t_{ij}))'$ $(j = 1, \ldots, n_i)$. The number of observations and the observation times may vary from subject to subject.

We assume that coordinate k of the response has marginal distribution function F_{ik} and density/mass function f_{ik}, both of which may depend on time-varying parameters $\boldsymbol{\theta}_k(t)$, some of which may be regression coefficients $\boldsymbol{\beta}_k(t)$. For example, in the oft-cited bivariate continuous–binary case we might adopt a Gaussian linear model for one coordinate:

$$Y_{i1}(t) \sim \mathcal{N}\{\boldsymbol{x}'_{i1}(t)\boldsymbol{\beta}_1(t), \sigma^2(t)\}$$

and a logistic model for the other:

$$Y_{i2}(t) \sim \mathcal{B}([1 + \exp\{-\boldsymbol{x}'_{i2}(t)\boldsymbol{\beta}_2(t)\}]^{-1}),$$

where $\mathcal{B}(p)$ denotes a Bernoulli random variable with mean p, and $\boldsymbol{x}_{i1}(t)$ and $\boldsymbol{x}_{i2}(t)$ are vectors of predictors for subject i, measured at time t. Although it is often convenient to work within the familiar generalized linear model (GLM) framework, in which case our local model (Section 2.2) is reminiscent of the vector GLM [71], there are of course many other options for the marginal specifications: extreme value distributions, beta regression models, zero-inflated models, skew-normal models, heavy-tailed distributions, etc.

We model dependence using a time-varying d-copula $C_{\boldsymbol{\gamma}(t)}\{u_1(t), \ldots, u_d(t)\}$, where $\boldsymbol{\gamma}(t)$ are copula parameters [52]. A convenient choice is the Gaussian copula [38, 69]

$$\Phi_{\mathbf{R}(t)}[\Phi^{-1}\{u_1(t)\}, \ldots, \Phi^{-1}\{u_d(t)\}],$$

where $\Phi_{\mathbf{R}(t)}$ is the cdf of a d-variate multinormal random variable with mean vector $\mathbf{0}$ and correlation matrix $\mathbf{R}(t)$, and Φ^{-1} is the univariate standard normal quantile function. Other attractive choices are the t copula [15], which can accommodate tail dependence, or the skew t copula [68], which can accommodate tail dependence and asymmetric dependence.

We use a one-step estimation procedure based on optimization of an approximation to the local kernel-weighted log likelihood of $\boldsymbol{\theta}(t) = (\boldsymbol{\theta}'_1(t), \ldots, \boldsymbol{\theta}'_d(t), \boldsymbol{\gamma}'(t))'$. The approximation is based on the distributional transform (DT) (explained below) and was first proposed by Kazianka and Pilz [40] for fitting Gaussian copula geostatistical models.

2.1 Likelihood inference for Gaussian copula models

In this subsection we revert temporarily to an ordinary likelihood setting. This will ease notation and allow us to motivate our likelihood approximation as simply and clearly as possible. In Section 2.2 we will reintroduce time varyingness and describe our approach to local likelihood inference for TIMECOP.

Likelihood inference is fairly straightforward for Gaussian copula models with continuous marginals. To see this, first note that the density for the Gaussian d-copula is

$$\begin{aligned} c_{\mathbf{R}}(\boldsymbol{u}) &= \frac{\phi_{\mathbf{R}}\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\}}{\prod_{i=1}^{d} \phi\{\Phi^{-1}(u_i)\}} \\ &\propto \mid \mathbf{R} \mid^{-1/2} \exp\left\{-\frac{1}{2}z'(\mathbf{R}^{-1} - \mathbf{I})z\right\}, \end{aligned}$$

where $\boldsymbol{z} = (z_1, \dots, z_d)' = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))'$ and $\mathbf{I}$ is the $d \times d$ identity matrix. If the desired marginal distributions $F_1, \dots, F_d$ are continuous also, the likelihood of the parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$ has the form

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto c_{\mathbf{R}}\{F_1(y_1), \dots, F_d(y_d)\} \prod_{i=1}^{d} f_i(y_i),$$

where f_i is the density function corresponding to F_i. This implies the log likelihood

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{y}) = -\frac{1}{2}\log \mid \mathbf{R} \mid - \frac{1}{2}z'(\mathbf{R}^{-1} - \mathbf{I})z + \sum_{i=1}^{d} \log f_i(y_i), \quad (2)$$

where $z_i = \Phi^{-1}\{F_i(y_i)\}$. This log likelihood can be optimized to arrive at the maximum likelihood estimate of $\boldsymbol{\theta}$.

When some of the marginal distributions are discrete, the likelihood does not have the simple form given above because $z_i = \Phi^{-1}\{F_i(y_i)\}$ is not standard normal (since $F_i(y_i)$ is not standard uniform if F_i has jumps). In this case the true likelihood is more complicated [44, 71] and becomes unwieldy as the number of discrete coordinates increases. An appealing alternative to the true likelihood is an approximation based on the distributional transform.

It is well known that if $Y \sim F$ is continuous, $F(Y)$ has a standard uniform distribution. But if Y is discrete, $F(Y)$ tends to be stochastically larger, and $F(Y^-) = \lim_{x \nearrow Y} F(x)$ tends to be stochastically smaller, than a standard uniform random variable. This can be remedied by stochastically "smoothing" F at its jumps. This technique goes at least as far back as Ferguson [24], who used it in connection with hypothesis tests. More recently, the distributional transform has been applied to stochastic ordering [59], conditional value at risk [4], and the extension of limit theorems for the empirical copula process to general distributions [60].

Let $W \sim \mathscr{U}(0, 1)$, and suppose that $Y \sim F$ and is independent of W. Then the distributional transform

$$G(W, Y) = WF(Y^-) + (1 - W)F(Y)$$

follows a standard uniform distribution, and $F^{-1}\{G(W, Y)\}$ follows the same distribution as Y.

Kazianka and Pilz [40] suggested approximating $G(W, Y)$ by replacing it with its expectation with respect to W:

$$\begin{aligned} G(W, Y) &\approx \mathbb{E}_w G(W, Y) \\ &= \mathbb{E}_w\{WF(Y^{-}) + (1 - W)F(Y)\} \\ &= \mathbb{E}_w WF(Y^{-}) + \mathbb{E}_w(1 - W)F(Y) \\ &= F(Y^{-})\mathbb{E}_w W + F(Y)\mathbb{E}_w(1 - W) \\ &= \frac{F(Y^{-}) + F(Y)}{2}. \end{aligned}$$

To construct the approximate log likelihood, then, we replace $F_i(y_i)$ in (2) with

$$\frac{F_i(y_i^{-}) + F_i(y_i)}{2}$$

for each discrete coordinate of the response. Note that this becomes

$$\frac{F_i(y_i - 1) + F_i(y_i)}{2}$$

if the distribution has integer support.

This approximation, although crude, performs well as long as the discrete distribution in question has a sufficiently large variance, in which case we suggest using the approximation when the true likelihood is too cumbersome to obtain. We employed the approximation in the trivariate simulation study. We used the true likelihood in the bivariate simulation study and data applications, for those scenarios involved binary outcomes, for which the DT-based approximation tends to perform poorly.

2.2 Local likelihood inference for timecop

Now we return to TIMECOP, for which we recommend local likelihood inference. That is, we estimate $\boldsymbol{\theta}(t)$ at time t_0 by maximizing the local kernel-weighted log likelihood

$$\ell\{\boldsymbol{\theta}(t_0) \mid \mathbf{T}\} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \ell\{\boldsymbol{\theta}(t_0) \mid \mathbf{y}_{ij}\} K\{(t_0 - t_{ij})/h\}/h,$$

where $\mathbf{T} = (\boldsymbol{t}_1 \cdots \boldsymbol{t}_m)$, $\ell\{\boldsymbol{\theta}(t_0) / \boldsymbol{y}_{ij}\}$ is the log likelihood of $\boldsymbol{\theta}(t_0)$ given the outcomes for subject i at time point t_{ij}, K is a kernel, and h is a bandwidth. In this section, we describe how to use the approximate log likelihood described in the previous section. Specifically, if we partition the response vector so that the first d_1 coordinates are continuous and the remaining coordinates are discrete, we have

$$\ell\{\boldsymbol{\theta}(t_0) \mid \boldsymbol{y}_{ij}\} = -\frac{1}{2}\log \ |\mathbf{R}| \ - \frac{1}{2}(z_{ij}', z_{ij}^{*\,\prime})\{\mathbf{R}^{-1} - \mathbf{I}\}(z_{ij}', z_{ij}^{*\,\prime})' + \sum_{k=1}^{d} \log f_{ik}\{y_{ik}(t_{ij})\}, \quad (3)$$

where

$$z_{ij} = (\Phi^{-1}[F_{i1}\{y_{i1}(t_{ij})\}], \ldots, \Phi^{-1}[F_{id_1}\{y_{id_1}(t_{ij})\}])',$$
$$z_{ij}^{*} = (\Phi^{-1}\{u_{i(d_1+1)}(t_{ij})\}, \ldots, \Phi^{-1}\{u_{id}(t_{ij})\})'.$$

The distributional transform approximation enters through computation of the $u_{ik}(t_{ij})$ ($k = d_1 + 1, \ldots, d$):

$$u_{ik}(t_{ij}) = \frac{F_{ik}\{y_{ik}^{-}(t_{ij})\} + F_{ik}\{y_{ik}(t_{ij})\}}{2}.$$

We obtain $\hat{\boldsymbol{\theta}}(t_0)$ using the quasi-Newton method of Byrd et al. [6] so that estimated dependence and scale parameters can be appropriately constrained.

We used local constant estimation in our simulation studies, i.e., we assumed that $\boldsymbol{\theta}(t_0)$ is constant on a neighborhood of t_0. It is straightforward to use a higher-order polynomial approximation to $\boldsymbol{\theta}(t_0)$, but even a linear approximation—in which one assumes that

$$\boldsymbol{\theta}(t_0) = \boldsymbol{\theta}(t) + \dot{\boldsymbol{\theta}}(t)(t_0 - t)$$

for t on a neighborhood of t_0—increases the computational burden quite a bit while reducing bias only slightly.

In the final step of our procedure we estimate the variance of $\hat{\boldsymbol{\theta}}(t_0)$. Here we use results obtained by Fan et al. [18]. We begin with the approximate conditional variance (conditional on $\mathbf{T}$), which has a sandwich form:

$$\begin{aligned} \textstyle\sum(t_0) &= \mathbb{V}\{\hat{\boldsymbol{\theta}}(t_0) \mid \mathbf{T}\} \qquad (4) \\ &\approx \kappa(t_0)\mathscr{H}^{-1}(t_0)\mathscr{J}(t_0)\mathscr{H}^{-1}(t_0) \\ &= \kappa(t_0)[\ddot{\ell}\{\boldsymbol{\theta}(t_0) \mid \mathbf{T}\}]^{-1}\mathbb{V}[\dot{\ell}\{\boldsymbol{\theta}(t_0) \mid \mathbf{T}\}][\ddot{\ell}\{\boldsymbol{\theta}(t_0) \mid \mathbf{T}\}]^{-1}, \end{aligned}$$

where $\kappa(t_0) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} K^2\{(t_0 - t_{ij})/h\}/h^2$ and $\mathcal{H}(t_0)$ is the Hessian matrix, which can be estimated by $\ddot{\ell}\{\hat{\boldsymbol{\theta}}(t_0)\ /\mathbf{T}\}$ as a side effect of optimization. The variance of the score, $\mathcal{J}(t_0)$, can be estimated by

$$\frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \nabla \nabla' \ell\{\hat{\boldsymbol{\theta}}(t_0) \mid \boldsymbol{y}_{ij}\} K\{(t_0 - t_{ij})/h\}/h}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} K\,\{(t_0 - t_{ij})/h\}/h},$$

where ∇ denotes the gradient.

Invoking asymptotic normality [30, 14] and using $\hat{\Sigma}(t_0)$, we construct a pointwise $(1-\alpha)100\%$ confidence interval for the νth element of $\boldsymbol{\theta}(t_0)$ as

$$\hat{\boldsymbol{\theta}}_{\nu}(t_0) \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\Sigma}_{\nu}(t_0)}, \quad (5)$$

where $\hat{\Sigma}_{\nu}(t_0)$ is the νth diagonal element of $\hat{\Sigma}(t_0)$.

Note that our procedure does not account for intrasubject dependence, but theory suggests that intra-subject dependence can safely be ignored (for estimation of $\boldsymbol{\theta}(t)$) when the number of subjects is sufficiently large. As shown in Lin and Carroll [45], the method of kernel generalized estimation equations (kernel GEE) yields a root-n consistent estimator regardless of the working correlation structure. Furthermore, kernel GEE with working independence correlation matrix yields the most efficient estimator for the nonparametric regression function in a longitudinal setting. The procedure proposed here shares the spirit of kernel GEE, and so we suspect that our estimator is root-n consistent and is likely the most efficient. Theoretical justification is beyond the of scope of this paper and must be left to a future investigation.

As for inference, Fan et al. [18] noted that the intervals given by (5) might be too narrow for some datasets (due to intra-subject dependence), in which case one can get better coverage by using the wider intervals obtained from an undersmoothed fit. Undersmoothing is effective here because any two kernel-weighted intra-subject observations are nearly uncorrelated when h is sufficiently small and the serial dependence is short- or medium-range [19]. See Section 3 for details regarding appropriate undersmoothing for TIMECOP.

Practical application of our approach depends on selection of a suitable bandwidth. For this we recommend a form of cross-validation proposed by Fan and Zhang [22]. We leave out a single subject at a time rather than a single observation, since the latter approach is inappropriate when there is intra-subject dependence [35]. After removing the ith subject, we estimate $\boldsymbol{\theta}(\cdot)$ based on the remaining subjects. After doing this for each of the m subjects, we combine the results to form the cross-validation score

$$\mathrm{CV}(h) = -\sum_{i=1}^{m}\sum_{j=1}^{n_i} \ell\{\hat{\boldsymbol{\theta}}^{\backslash i}(t_{ij}) \mid \boldsymbol{y}_{ij}, h\},$$

where $\hat{\boldsymbol{\theta}}^{\backslash i}(t_{ij})$ is the leave-i-out estimate for time t_{ij}. We compute the cross-validation score for a range of bandwidths and select the bandwidth that minimizes the score.

We recommend that a bimodal kernel [12] be used because doing so leads to a more accurate estimate in the presence of intra-subject dependence. Informally, a bimodal kernel removes serial dependence by down weighting observations that are very close to t_0. This prevents undersmoothing by preventing the estimation procedure from “mistaking” local similarity for structure that should be fitted. We use the member of the so called ε-optimal class of bimodal kernels recommended by De Brabanter et al. [12]. Specifically, we use

$$K_{\varepsilon}(u) = \frac{4}{4 - 3\varepsilon - \varepsilon^2}\begin{cases} \frac{3}{4}(1-u^2)\mathbf{1}\{|u| \le 1\} & \text{if } |u| \ge \varepsilon \\ \frac{3}{4}\frac{1-\varepsilon^2}{\varepsilon}|u| & \text{if } |u| < \varepsilon \end{cases}$$

with $\varepsilon = 0.1$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

3. SIMULATED APPLICATION

We investigated the finite sample performance of our estimator using a simulation study designed to mimic the WIHS data that we analyze in Section 4. The response in our study was binary–continuous. Specifically, we let $Y_{i1}(t)$ be a Gaussian process with mean $\boldsymbol{x}'_{i1}(t)\boldsymbol{\beta}_1(t)$ and variance $\sigma^2(t)$, and $Y_{i2}(t)$ be a Bernoulli process with mean

$$[1 + \exp\{-\boldsymbol{x}'_{i2}(t)\boldsymbol{\beta}_2(t)\}]^{-1},$$

as described above in Section 2. For the time-varying coefficients $\boldsymbol{\beta}_1(t)$ and $\boldsymbol{\beta}_2(t)$, we used

$$\begin{aligned} \boldsymbol{\beta}_1(t) &= (\beta_{10}(t), \beta_{11}(t), \beta_{12}(t), \beta_{13}(t))' \\ &= 0.2(\cos 2\pi t, \sin 2\pi t, -\sin 2\pi t, 1 + \sin 2\pi t)' \\ \boldsymbol{\beta}_2(t) &= (\beta_{20}(t), \beta_{21}(t))' = 0.2(\sin 2\pi t, 1 + \cos 2\pi t)', \end{aligned}$$

with $\beta_{k0}(t)$ an intercept and $\beta_{k1}(t)$, $\beta_{12}(t)$, and $\beta_{13}(t)$ slopes. We simulated the predictors, independently, from the standard normal distribution. $Y_{i1}(t)$’s time-varying standard deviation was $\sigma(t) = 0.8+0.2 \sin 2\pi t$. And the cross-sectional correlation function was $\rho(t) = 0.2 + 0.15 \sin 2\pi t$.

Although it does not enter into our estimation procedure, we simulated both processes with CAR(1) dependence:

$$
\begin{aligned}
\rho_1(s,t) &= 2^{-7\,|\,s-t\,|}\\
\rho_2(s,t) &= 5^{-7\,|\,s-t\,|},
\end{aligned}
$$

for any two times s and t in the unit interval. These functions correspond to consequential but short-range dependence. Specifically, if we define the effective range to be the distance $/s-t/$ at which the correlation between two observations has dropped to 0.05, these functions have an effective range of approximately 0.2.

We simulated a single dataset as follows. For subject i,

1. let $n_i = 10$;
2. simulate n_i measurement times $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})'$ from the standard uniform distribution;
3. for $j = 1, \ldots, n_i$, construct the 2×2 correlation matrix with off-diagonal entries $\rho(t_{ij})$, and use the Cholesky root to impose the correlation on $(W_{i1}(t_{ij}), W_{i2}(t_{ij}))' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
4. construct the $n_i \times n_i$ correlation matrices $\mathbf{R}_1(\boldsymbol{t}_i)$ and $\mathbf{R}_2(\boldsymbol{t}_i)$ according to the CAR(1) specification given above, and use the corresponding Cholesky roots to impose the correlation structures on $(W_{i1}(t_{i1}), \ldots, W_{i1}(t_{in_i}))'$ and $(W_{i2}(t_{i1}), \ldots, W_{i2}(t_{in_i}))'$;
5. form $\boldsymbol{Z} = (W_{i1}(t_{i1}), W_{i2}(t_{i1}), \ldots, W_{i1}(t_{in_i}), W_{i2}(t_{i1n_i}))'$;
6. apply the probability integral transform (PIT) to each element of $\boldsymbol{Z}$ to arrive at

   $$
   \begin{aligned}
   \boldsymbol{U} &= (U_{i1}(t_{i1}), U_{i2}(t_{i1}), \ldots, U_{i1}(t_{in_i}), U_{i2}(t_{in_i}))'\\
   &= (\Phi^{-1}\{Z_{i1}(t_{i1})\}, \Phi^{-1}\{Z_{i2}(t_{i1})\}), \ldots, \Phi^{-1}\{Z_{i1}(t_{in_i})\}, \Phi^{-1}\{Z_{i2}(t_{in_i})\}',
   \end{aligned}
   $$

   the elements of which are uniformly distributed; and
7. apply the inverse PIT to $\boldsymbol{U}$ to produce the response

   $$
   (Y_{i1}(t_{ij}), Y_{i2}(t_{ij}))' = (F_{i1}^{-1}\{U_{i1}(t_{ij})\}, F_{i2}^{-1}\{U_{i2}(t_{ij})\})',
   $$

   where F_{i1}^{-1} and F_{i2}^{-1} are the inverse cdfs corresponding to the desired Gaussian and Bernoulli marginal distributions described above and $j = 1, \ldots, n_i$.

Note that the simulated outcomes need not have precisely the same dependence structure as the underlying copula realization [41, 48]. This is because, for non-Gaussian outcomes, the margins impose bounds (the so called Fréchet–Hoeffding bounds) on the achievable correlation. For example [56], the maximum correlation for two binary random variables with expectations p_1 and p_2 is

$$\min\left\{\sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}}, \sqrt{\frac{p_2(1-p_1)}{p_1(1-p_2)}}\right\}.$$

We used a pilot study to choose several bandwidths. For each bandwidth we simulated and fitted 500 datasets, each having $m = 300$ subjects, and estimated $\boldsymbol{\theta}(t) = (\boldsymbol{\beta}_1'(t), \sigma(t), \boldsymbol{\beta}_2'(t), \rho(t))'$ at 200 grid points equally spaced over the unit interval. Figures 1 and 2 show the results for $h_0 = 0.1$, the bandwidth that minimized the cross-validation score described in Section 2.

Our procedure performed well overall for this scenario with respect to bias, as the biases are generally small. The standard errors for the selected bandwidth of $h_0 = 0.1$ were very accurate for the slope functions $\beta_{11}(t)$, $\beta_{12}(t)$, $\beta_{13}(t)$, and $\beta_{21}(t)$, and so the coverage rates for those functions were very close to the desired 95%. For the other parameters, especially $\sigma(t)$ and $\rho(t)$, the procedure tended to yield optimistic confidence intervals. We remedied this by using a smaller bandwidth for variance estimation. Specifically, we used h_1 in $O(n^{-1/4})$ since the asymptotically optimal bandwidth is in $O(n^{-1/5})$. For our simulation scenario, this leads to $h_1 = 0.067$. We see from the plots in Figures 1 and 2 that this bandwidth yielded accurate confidence intervals.

It is of interest to observe the performance of our methodology in a missing completely at random (MCAR) scenario [46] since missingness of this type is common in longitudinal studies. We designed our MCAR study as follows: for any subject and any time point, if the value of the second predictor for the continuous response is greater than some cutoff value, delete that observation. Our goal was to create approximately 15% missingness in each simulated data set. Since the second predictor is standard normal, a cutoff value of 1.03 allowed us to achieve our target rate. Figure 3 shows selected results under this missingness scenario, for the same bandwidth that was used in the first study. The plots show that our procedure performed comparably for the two studies, which suggests that our approach can handle ignorable missingness.

Our approach performs well with considerably fewer subjects if the binary data are replaced by, say, count data. To demonstrate this we present selected results from a second simulation study. For the second study we simulated a trivariate process for 100 subjects. The first coordinate was Gaussian and identical to the process used in the first study. The second coordinate was Poisson with mean $\exp\{\boldsymbol{x}_{i2}'(t)\boldsymbol{\beta}_2(t)\}$ and the same serial dependence as the Bernoulli process from the first study. And the third coordinate was Beta$\{\alpha(t), 2\}$ with $\alpha(t) = 5 + 0.2 \sin 2\pi t$ and serial dependence $\rho_3(s, t) = 10^{-6/s-t/}$. The three dependence functions for the joint process at time t were

$$\begin{aligned} \rho_{12}(t) &= 0.2 + 0.15 \sin 2\pi t \quad \text{Gaussian}-\text{Poisson} \\ \rho_{13}(t) &= 0.2 + 0.15 \cos 2\pi t \quad \text{Gaussian}-\text{Beta} \\ \rho_{23}(t) &= 0.3 + 0.15 \sin 2\pi t \quad \text{Poisson}-\text{Beta}. \end{aligned}$$

The results are shown in Figure 4. Note that we once again undersmoothed to obtain accurate confidence intervals for the non-slope parameter functions.

Additional simulation studies suggest that TIMECOP performs well for several, or even many, outcomes in realistic scenarios similar to those considered above, i.e., for at least 100 subjects and at least ten observations per subject, say.

4. APPLICATION TO HIV DATA

In this section we apply our proposed methodology to data from the Women's Interagency HIV Study (WIHS). These data contain information on 372 women recruited between 1994 and 1995 from HIV primary care clinics, research programs, community outreach sites, women's support groups, drug rehabilitation programs, and HIV testing sites in Chicago, Los Angeles, New York City, San Francisco, and Washington, DC. Participants were evaluated at WIHS sites every six months with an extensive interview that included physical and oral examinations, blood and gynecological specimen collection, and collection of information regarding participants' daily activities (such as their sexual behaviors and tobacco use). Our analysis is restricted to 292 participants who were HIV positive. Among these subjects, 26% self-identified as Latina or Hispanic, 45% of the women were of African-American non-Hispanic origin, and 12% were of white non-Hispanic origin. Sixty-six percent of the participants were smokers, and, while taking part in the study, 8.3% of the smokers quit smoking while 8.1% of the non-smokers started smoking. Although our data set contains follow-up information on women aged 25–55 until 2006, many participants failed to attend some of their scheduled visits, which led to unequal numbers of measurements and different measurement times. The number of observations for each subject varies from one to eight.

It is known that cigarette smoking has effects on the immune system [25, 27, 31, 36], but it is not yet clear whether any of these effects influence the progression of HIV. Burns et al. [5] analyzed data on a cohort of 3,221 HIV-seropositive men and women enrolled in the Terry Beirn Community Programs for Clinical Research on AIDS. They used proportional hazards regression analysis to assess the differences between never, former, and current cigarette smokers in terms of clinical outcomes, and found no association between cigarette smoking and the overall risk of disease progression or death. Similarly, Galai et al. [27] used Kaplan-Meier analysis and multivariate Cox regression models to investigate the effect of cigarette smoking on the development of AIDS in the Multicenter AIDS Cohort Study of homosexual men. Their analysis revealed that smoking was not significantly associated with progression to AIDS. However, Nieman et al. [54] found that in a case series of 84 individuals, smokers progressed to AIDS more rapidly than nonsmokers. Their analysis used life tables and compared the median time to develop AIDS for smokers and nonsmokers.

Given these inconsistent findings, our primary interest was in investigating the association between HIV progression (as measured by CD4 cell percentage) and smoking status among women with HIV enrolled in the WIHS, while revealing response–predictor relationships for both responses. Based on the HIV literature [75, 55] and exploratory analyses, we chose a number of predictors. For the continuous response we used baseline CD4 cell percentage (measured at the first visit), number of sexual partners, hematocrit value (the volume percentage of red cells in the blood), mean corpuscular volume (a measure of average red blood cell size), platelet count, and Center for Epidemiologic Studies Depression (CESD) scale score. For the binary response we used the CESD scale score and race. All predictors save race are continuous, and we centered them. The race variable originally had five levels: African American, white, Asian/Pacific Islander, native American/Alaskan native, and other. However, our subset of the data had just two participants in each of the Asian/Pacific Islander and native American/Alaskan native categories. Hence, we recategorized race into three levels: African American, white, and other.

To minimize modeling bias we assumed a maximally flexible model, which is to say that we permitted all parameters, including the dependence parameter $\rho(t)$, to be time varying. If the confidence bands for some parameters suggest that those parameters may be constant with respect to time, our methodology can be used to fit a semivarying model [76].

We assume that CD4 cell percentage, $Y_{i1}(t)$, is a Gaussian process with mean $\boldsymbol{x}'_{i1}(t)\boldsymbol{\beta}_1(t)$ and variance σ^2(t), where $\boldsymbol{\beta}_1(t) = (\beta_{10}(t), \beta_{11}(t), \beta_{12}(t), \beta_{13}(t), \beta_{14}(t), \beta_{15}(t), \beta_{16}(t))'$ and $\boldsymbol{x}'_{i1}(t) = (1, x_{i11}(t), x_{i12}(t), x_{i13}(t), x_{i14}(t), x_{i15}(t), x_{i16}(t))'$ with for subject i

$x_{i11}(t)$: the baseline CD4 (BaseCD4) cell percentage at the first visit,

$x_{i12}(t)$: the number of sexual partners (PART) at time t,

$x_{i13}(t)$: the hematocrit (HCV) value at time t,

$x_{i14}(t)$: the mean corpuscular volume (MCV) at time t,

$x_{i15}(t)$: the platelet count (PLAT) at time t,

$x_{i16}(t)$: the CESD scale score at time t.

We assume that smoking status, $Y_{i2}(t)$, is a Bernoulli process with mean

$$[1 + \exp\{-\boldsymbol{x}'_{i2}(t)\boldsymbol{\beta}_2(t)\}]^{-1},$$

where $\boldsymbol{\beta}_2(t) = (\beta_{20}(t), \beta_{21}(t), \beta_{22}(t))'$ and $\boldsymbol{x}'_{i2}(t) = (1, x_{i21}(t), x_{i22}(t), x_{i23}(t))'$ with

$x_{i21}(t)$: the CESD scale score of subject i at time t,

$x_{i22}(t)$: the first dummy variable for race (RACE 1)

($x_{i22}(t) = 1$ if subject i is African American),

$x_{i23}(t)$: the second dummy variable for race (RACE 2)

($x_{i23}(t) = 1$ if subject i is white).

We used the $K_{0.1}$ bimodal kernel and chose a bandwidth of $h = 14$ using the cross-validation procedure described in Section 2. Note that the time covariate in this study is the age of the participant. The estimated time-varying regression coefficient functions and variance for the continuous response (CD4 cell percentage) are shown in Figure 5.

- From panel (a) we see that the intercept function is time varying and increases with age.
- The plot in panel (b) suggests that the effect for baseline CD4 is time varying and decreases with age. Moreover, the effect is always significant and positive for ages between 25 and 55.
- The confidence band in panel (c) suggests that the coefficient for number of sexual partners may be time invariant. And the effect is significantly different from zero only between ages 43 and 55.
- The plot in panel (d) suggests that the effect of hematocrit may be time invariant, but the effect is significant and positive for ages between 28 and 50.
- According to panel (e), the effect of mean corpuscular volume may be time invariant, but the effect is significant and positive after age 26.
- Panel (f) shows that the effect of platelet count is significant and positive after age 29. But the confidence band is too wide to support the conclusion that the effect is time varying.
- From panel (g) we see that the effect of CESD score is always significant and negative, i.e., depression is associated with a lower CD4 cell percentage. The effect may be constant with respect to time, however.
- The final panel (just barely) allows us to conclude that the variance of CD4 percentage is time varying and increases with age.

Figure 6 shows the estimated time-varying regression coefficient functions for the binary response (smoking status) along with the estimated time-varying association, $\hat{\rho}(t)$, between CD4 cell percentage and smoking status.

- Panel (a) shows that the intercept function is time varying and increases with age.
- Panel (b) suggests that the coefficient for CESD score is time varying. The coefficient is significant until age 50 and decreases with age. We see that the effect is positive, which implies an association between depression and smoking. The association is evidently weaker for older patients.
- In panel (c) the upper and lower three-curve groups are the estimates and confidence bands for the RACE 1 and RACE 2 variables, respectively. The confidence bands reveal that the coefficient for RACE 1 is time varying while the coefficient for RACE 2 may be time invariant. We see that the coefficient for RACE 1 is always significant and positive, and until age 45 is greater than the coefficient for RACE 2. That is, African Americans have higher odds of smoking than do patients of other races. After age 45 the confidence bands for RACE 1 and RACE 2 overlap, which indicates that the difference between the two groups becomes insignificant. We also observe that RACE 2 is not a significant predictor of smoking.

In Section 1 we mentioned that joint modeling can result in more precise estimation of marginal parameters. To demonstrate this for the WIHS data, we fitted a univariate time-varying model for each of the outcomes. In each univariate model, the other response was included as an additional predictor. Figure 7 compares standard errors for the joint model to those for the univariate models, for selected parameters. We see that joint modeling leads to considerable efficiency gains near the endpoints. The difference is negligible between ages 35 and 45, which is not surprising: approximately 55% of our data were observed in this time interval, and, as Gueorguieva and Agresti [29] argue, for larger sample sizes the efficiency gained through joint modeling is less pronounced.

Finally, let us interpret the last panel of Figure 6, which shows the estimated time-varying partial association. Judging from the confidence band in this plot, we do not have sufficient evidence to conclude that $\rho(t)$ is time varying. But we can conclude that the partial association is always significant and negative, i.e., for women enrolled in the WIHS, decreased CD4 cell percentage is partially associated with smoking. Although we studied women only, this finding provides evidence that smokers progress to AIDS more rapidly than nonsmokers. In addition to its negative association with CD4 percentage, smoking is known to decrease the adherence to highly active antiretroviral therapy [23]. Smoking also poses additional threats to HIV-positive patients, such as pulmonary-related complications (pneumonia, asthma, and chronic obstructive pulmonary disease) and increased incidence of opportunistic infections [1, 11, 42]. Therefore, the findings of our study and others suggest that smoking cessation counseling is a necessary component of any program that seeks to enhance quality of life and disease management for HIV patients. Niaura et al. [53] provided a review of existing cessation techniques for HIV patients, and also suggested ways to improve research studies so that more effective cessation treatments can be discovered.

Since some of the parameters may be constant with respect to time, a semivarying model is probably the most appropriate model for these data. Although our method can be adapted to the semivarying setting, it seems clear that fitting such a model to the WIHS data would not result in substantive changes to our conclusions.

5. APPLICATION TO SMOKING CESSATION DATA

In this section we apply TIMECOP to the smoking cessation data mentioned in the introduction.

According to a 2004 report by the U.S. Department of Health and Human Services [72], cigarette smoking is one of the leading preventable causes of several diseases, including coronary heart disease, acute myeloid leukemia, and bladder, esophageal, laryngeal, lung, oral, and throat cancers. Therefore, prevention scientists have designed studies to explore the motivation behind smoking and factors that might promote smoking. These studies revealed that alcohol, coffee, food, and presence of others smoking increase the odds of smoking [17, 63, 37, 64].

Urge to smoke is another variable that is often of interest in smoking cessation research because of (1) the intuitive link between urge to smoke and smoking, and (2) the importance of urges in some theories of smoking (for instance, many theories posit that the influence of emotional states on smoking is impacted through the urge to smoke). Shiffman et al. [64]

investigated the association between urge to smoke and smoking and concluded that there is a strong and positive association between these variables, especially for lower levels of urge to smoke.

A drawback of the previous analyses was that they did not explore the dynamics of the association between these factors and smoking. However, we suspect that the relationship between these stimuli and smoking (and therefore, perhaps, urge to smoke) might vary over the course of a cessation study, especially before and after a subject quits smoking. Thus we created a binary response by combining the factors that might lead to smoking (specifically, the binary outcome is zero iff none of the factors is present), and applied TIMECOP with the new binary outcome and urge to smoke as response variables.

The data were collected using hand-held palm-top computers that prompted each participant at random times. When prompted, the subjects recorded their answers to a series of questions about their current setting and activities as well as current mood and urge to smoke. The data collection process is described below.

First, the subjects were monitored for a two-week interval during which they engaged in their ordinary smoking behavior. They were asked to record all their smoking occasions during this period, and to respond to the random assessment prompts. Patients were instructed to quit smoking at the end of this two-week period. When a patient had abstained for 24 hours, the current day was recorded as that patient's quit day. After the subjects quit, they were required to continue responding to the random assessment prompts and to record any episodes of smoking (lapses) or strong temptations. Although all subjects were instructed to quit on a certain date, different subjects had different quit days, and the prompts were random. Thus the subjects have unequal numbers of measurements and different measurement times.

In our analysis we focus on the randomly scheduled assessment data collected two weeks before and after the quit day, so that we can study the differences between these periods. We analyzed the data for 206 smokers, each of which had from 46 to 222 observations.

Based on previous analyses of smoking cessation data [62, 65, 64, 43], we used the mood variables negative affect, arousal, and attention disturbance as predictors for urge to smoke. (see [64] for more information regarding these scores). We used the same set of predictors for the binary response. All of these predictors are continuous.

We assume that urge to smoke, $Y_{i1}(t)$, is a Gaussian process with mean $\boldsymbol{x}_{i1}'(t)\boldsymbol{\beta}_1(t)$ and variance $\sigma^2(t)$, where $\boldsymbol{\beta}_1(t) = (\beta_{10}(t), \beta_{11}(t), \beta_{12}(t), \beta_{13}(t))'$ and $\boldsymbol{x}_{i1}'(t) = (1, x_{i11}(t), x_{i12}(t), x_{i13}(t))'$ with for subject i

$x_{i11}(t)$: the centered score of negative affect at time t,

$x_{i12}(t)$: the centered score of arousal at time t,

$x_{i13}(t)$: the centered score of attention disturbance at time t.

We assume that smoking triggers, $Y_{i2}(t)$, is a Bernoulli process with mean

$$[1 + \exp\{-\boldsymbol{x}_{i1}'(t)\boldsymbol{\beta}_2(t)\}]^{-1},$$

where $\boldsymbol{\beta}_2(t) = (\beta_{20}(t), \beta_{21}(t), \beta_{22}(t), \beta_{23}(t))'$.

As in the first data application, we used the $K_{0.1}$ bimodal kernel in our estimation procedure and chose a bandwidth of $h = 7$ using the cross-validation procedure described in Section 2. The estimated time-varying regression coefficient functions and variance for the continuous response (urge to smoke) are shown in Figure 8.

- The plot in panel (a) suggests that the intercept function is time varying, and decreases after the quit day.
- The confidence band in panel (b) shows that the coefficient for negative affect is time varying. The effect is always significant and positive, that is, as negative affect increases, urge to smoke also increases. This effect increases until five days after the quit day and then starts to decrease.
- According to panel (c) we observe that the effect of attention disturbance may be time invariant, but it is significantly different from zero and always positive.
- Panel (d) allows us to conclude that the coefficient for attention disturbance is time varying. The effect becomes significant and negative just prior to quit day. Approximately at the same date that the coefficient for negative affect starts to decrease, the effect of attention disturbance starts to increase.
- The final panel shows that the variance of urge to smoke varies over time. The variance is almost constant prior to the quit day, and then it starts to decrease.

Figure 9 shows the estimated time-varying regression coefficient functions for the binary response (factors that may lead to smoking) along with the estimated time-varying association, $\hat{\rho}(t)$, between this binary response and urge to smoke.

- From panel (a) we see that the intercept function is time varying and significant.
- Panel (b) shows that the coefficient for negative affect score might be time invariant. The effect is significant and negative until quit day, after which the effect becomes insignificant.
- In panel (c) we observe that the effect of arousal may be time invariant but is significant and positive.
- The confidence band in panel (d) suggests that the effect of attention disturbance is significant and just barely time varying. The effect is positive and decreases until the quit day. Then the effect becomes nearly constant.

According to the confidence band in the last panel of Figure 9, which shows the estimated time-varying partial association, we do not have sufficient evidence to conclude that $\rho(t)$ is time varying. However, we observe that this partial association is always significant and positive, i.e., for smokers enrolled in this study, exhibiting at least one of the factors is associated with an increased urge to smoke. This result could help prevention scientists to design better cessation programs. For instance, before someone tries to quit smoking, it may be wise of him/her to decrease the urge to smoke by mitigating some or all of the factors.

As in the application to HIV data, these results suggest that some of the parameters may be time invariant, in which case a semivarying model would be more appropriate. But fitting a semivarying model to these data would not lead to substantive changes to our conclusions.

6. CONCLUSION

In this article we developed a new class of joint models for longitudinal responses, and an estimation procedure based on the local likelihood approach. This new class of models can accommodate (1) responses of mixed type, (2) time-varying association parameters, and (3) all manner of time-varying marginal parameters. We demonstrated the efficacy of our approach via three simulation studies, one of which showed that our approach performs well under ignorable missingness. Then we used our methodology to analyze a bivariate response taken from the Women's Interagency HIV Study. This analysis revealed a significant negative partial association between CD4 cell percentage and smoking status. We also applied our approach to smoking cessation data. Our analysis revealed a significant and positive partial association between urge to smoke and factors that may lead to smoking.

Our proposed methodology assumes that all parameters have the same degree of smoothness. However, it may be of interest to develop an estimation procedure capable of handling multiple degrees of smoothness. Perhaps the two-step estimation procedure proposed by Fan and Zhang [21] can be adapted for this purpose.

In this paper we used confidence bands to assess features of parameter functions. It may be desirable to develop hypothesis testing procedures, in which case results presented in [20] may prove useful.

Acknowledgments

Hughes' research was supported by a grant from the Simons Foundation (#243657) and Li's research was supported by National Science Foundation DMS 1512422, National Institutes of Health (NIH), National Institute on Drug Abuse (NIDA) grants P50 DA039838 and P50 DA036107, NIH, National Library of Medicine (NLM), T32 LM012415 and National Nature Science Foundation of China (NNSFC) grants 11690015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA, the NLM, the NIH or the NNSFC.

Data in this manuscript were collected by the Women's Interagency HIV Study (WIHS) Collaborative Study Group with centers (Principal Investigators) at New York City/Bronx Consortium (Kathryn Anastos); Brooklyn, NY (Howard Minkoff); Washington, DC, Metropolitan Consortium (Mary Young); The Connie Wofsy Study Consortium of Northern California (Ruth Greenblatt); Los Angeles County/Southern California Consortium (Alexandra Levine); Chicago Consortium (Mardge Cohen); Data Coordinating Center (Stephen Gange). The WIHS is funded by the National Institute of Allergy and Infectious Diseases (UO1-AI-35004, UO1-AI-31834, UO1-AI-34994, UO1-AI-34989, UO1-AI-34993, and UO1-AI-42590) and by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (UO1-HD-32632). The study is co-funded by the National Cancer Institute, the National Institute on Drug Abuse, and the National Institute on Deafness and Other Communication Disorders. Funding is also provided by the National Center for Research Resources (UCSF-CTSI Grant Number UL1 RR024131).

References

1. Arcavi L, Benowitz NL. Cigarette smoking and infection. Archives of Internal Medicine. 2004; 164(20):2206–2216. [PubMed: 15534156]
2. Barkan SE, Melnick SL, Preston-Martin S, Weber K, Kalish LA, Miotti P, Young M, Greenblatt R, Sacks H, Feldman J. The women's interagency hiv study. Epidemiology. 1998; 9(2):117–125. [PubMed: 9504278]
3. Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves. Journal of American Statistical Association. 1998; 93(443):961–976.
4. Burgert C, Rüschendorf L. On the optimal risk allocation problem. Statistics & Decisions. 2006; 24(1):153–171.
5. Burns DN, Hillman D, Neaton JD, Sherer R, Mitchell T, Capps L, Vallier WG, Thurnherr MD, FMG. for the Terry Beirn Community Programs for Clinical Research on AIDS. Cigarette smoking, bacterial pneumonia, and other clinical outcomes in HIV-1 infection. Journal of Acquired Immune Deficiency Syndromes. 1996; 13(4):374–383.
6. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 1995; 16(5):1190–1208.
7. Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. Journal of the American Statistical Association. 1992; 87(419):651–658.
8. Christensen, R. Plane Answers to Complex Questions: The Theory of Linear Models. Springer; 2011.
9. Cleveland, WS., Grosse, E., Shyu, WM. Local Regression Models. Wadsworth & Brooks/Cole; Pacific Grove, CA: 1992.
10. Cox DR, Wermuth N. Response models for mixed binary and quantitative variables. Biometrika. 1992; 79(3):441–461.
11. Crothers K, Griffith TA, McGinnis KA, Rodriguez-Barradas MC, Leaf DA, Weissman S, Gibert CL, Butt AA, Justice AC. The impact of cigarette smoking on mortality, quality of life, and comorbid illness among HIV-positive veterans. Journal of General Internal Medicine. 2005; 20(12):1142–1145. [PubMed: 16423106]
12. De Brabanter K, De Brabanter J, Suykens J, Moor BD. Kernel regression in the presence of correlated errors. Journal of Machine Learning Research. 2011; 12:1955–1976.
13. de Leon, AR., Chough, KC. Analysis of mixed data: Methods & applications. Chapman and Hall/CRC; 2013.
14. de Melo EFL, Mendes BVM. Local estimation of copula based value-at-risk. Brazilian Review of Finance. 2009; 7(1):19–50.
15. Demarta S, McNeil AJ. The t copula and related copulas. International Statistical Review. 2005; 73(1):111–129.
16. Dunson DB. Bayesian latent variable models for clustered mixed outcomes. Journal of the Royal Statistical Society Series B. 2000; 62(2):355–366.
17. Emurian HH, Nellis MJ, Brady JV, Ray RL. Event time-series relationship between cigarette smoking and coffee drinking. Addictive Behaviors. 1982; 7(4):441–444. [PubMed: 7183199]
18. Fan J, Farmen M, Gijbels I. Local maximum likelihood estimation and inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1998; 60(3):591–608.
19. Fan, J., Gijbels, I. Local Polynomial Modeling and Its Applications. Boca Raton: Chapman and Hall/CRC; 1996.
20. Fan J, Zhang C, Zhang J. Generalized likelihood ratio statistics and Wilks phenomenon. Annals of Statistics. 2001; 29(1):153–193.
21. Fan J, Zhang W. Statistical estimation in varying coefficient models. The Annals of Statistics. 1999; 27(5):1491–1518.
22. Fan J, Zhang W. Statistical methods with varying coefficient models. Statistics and Its Interface. 2008; 1:179–195. [PubMed: 18978950]
23. Feldman JG, Minkoff H, Schneider MF, Gange SJ, Cohen M, Watts DH, Gandhi M, Mocharnuk RS, Anastos K. Association of cigarette smoking with HIV prognosis among women in the

HAART era: A report from the women's interagency HIV study. American Journal of Public Health. 2006; 96(6):1060–1065. [PubMed: 16670229]

24. Ferguson, TS. Mathematical Statistics: A Decision Theoretic Approach. New York: Academic Press; 1967.
25. Ferson M, Edwards A, Lind A, Milton GW, Hersey P. Low natural killer-cell activity and immunoglobulin levels associated with smoking in human subjects. International Journal of Cancer. 1979; 23(5):603–609. [PubMed: 457307]
26. Fitzmaurice GM, Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. Journal of the American Statistical Association. 1995; 90(431):845–852.
27. Galai N, Park LP, Wesch J, Visscher B, Riddler S, Margolick JB. Effect of smoking on the clinical progression of HIV-1 infection. Journal of Acquired Immune Deficiency Syndromes. 1997; 14(5): 451–8.
28. Gueorguieva R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. Statistical Modelling: An International Journal. 2001; 1(3): 177–193.
29. Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. Journal of the American Statistical Association. 2001; 96(455):1102–1112.
30. Hall P, Tajvidi N. Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. Statistical Science. 2000; 15(2):153–167.
31. Halonen M, Barbee RA, Lebowitz MD, Burrows B. An epidemiologic study of the interrelationships of total serum immunoglobulin E, allergy skin-test reactivity, and eosinophilia. Journal of Allergy and Clinical Immunology. 1982; 69(2):221–228. [PubMed: 7056953]
32. Hastie T, Tibshirani R. Varying-coefficient models. Journal of the Royal Statistical Society Series B. 1993; 55(4):757–796.
33. Heinen A, Rengifo E. Multivariate reduced rank regression in non-gaussian contexts, using copulas. Computational Statistics & Data Analysis. 2008; 52(6):2931–2944.
34. Hodges J, Reich B. Adding spatially-correlated errors can mess up the fixed effect you love. The American Statistician. 2010; 64(4):325–334.
35. Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998; 85(4):809–822.
36. Hughes DA, Haslam PL, Townsend PJ, Turner-Warwick M. Numerical and functional alterations in circulatory lymphocytes in cigarette smokers. Clinical & Experimental Immunology. 1985; 61(2): 459–466. [PubMed: 2931227]
37. Hymowitz N, Cummings KM, Hyland A, Lynn WR, Pechacek TF, Hartwell TD. Predictors of smoking cessation in a cohort of adult smokers followed for five years. Tobacco Control. 1997; 6:S57–S62.
38. Joe, H. Multivariate Models and Dependence Concepts. Chapman and Hall; 1997.
39. Kaslow RA, Blackwelder WC, Ostrow DG, et al. No evidence for a role of alcohol or other psychoactive drugs in accelerating immunodeficiency in hiv-1— positive individuals: A report from the multicenter aids cohort study. JAMA: The Journal of the American Medical Association. 1989; 261(23):3424–3429. [PubMed: 2524608]
40. Kazianka H, Pilz J. Copula-based geostatistical modeling of continuous and discrete data including covariates. Stochastic Environmental Research and Risk Assessment. 2010; 24(5):661–673.
41. Klaassen CA, Wellner JA, et al. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. Bernoulli. 1997; 3(1):55–77.
42. Kohli R, Lo Y, Homel P, Flanigan TP, Gardner LI, Howard AA, Rompalo AM, Moskaleva G, Schuman P, Schoenbaum EE. Bacterial pneumonia, HIV therapy, and disease progression among HIV-infected women in the HIV epidemiologic research (HER) study. Clinical Infectious Diseases. 2006; 43(1):90–98. [PubMed: 16758423]
43. Kürüm E, Li R, Shiffman S, Yao W. Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data. Statistica Sinica. 2016; 26(3):979–1000. [PubMed: 27667908]
44. Li M, Boehnke M, Abecasis GR, Song PXK. Quantitative trait linkage analysis using Gaussian copulas. Genetics. 2006; 173(4):2317–2327. [PubMed: 16751671]

45. Lin X, Carroll R. Semiparametric regression for clustered data using generalized estimating equations. Journal of the American Statistical Association. 2001; 96(455):1045–1056.

46. Little, RJA., Rubin, DB. Statistical Analysis with Missing Data. New York: John Wiley; 1987.

47. Liu X, Daniels M, Marcus B. Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. Journal of the American Statistical Association. 2009; 104(486):429–438. [PubMed: 20161053]

48. Madsen L, Birkes D. Simulating dependent discrete data. Journal of Statistical Computation and Simulation. 2013; 83(4):677–691.

49. Madsen L, Fang Y. Joint regression analysis for discrete longitudinal data. Biometrics. 2011; 67(3): 1171–1175. [PubMed: 21039391]

50. Mamary EM, Bahrs D, Martinez S. Cigarette smoking and the desire to quit among individuals living with HIV. AIDS Patient Care and STDs. 2002; 16(1):39–42. [PubMed: 11839217]

51. Masarotto G, Varin C. Gaussian copula marginal regression. Electronic Journal of Statistics. 2012; 6:1517–1549.

52. Nelsen, RB. An Introduction to Copulas. New York: Springer; 2006.

53. Niaura R, Chander G, Hutton H, Stanton C. Interventions to address chronic disease and HIV: Strategies to promote smoking cessation among HIV-infected individuals. Current HIV/AIDS Reports. 2012; 9(4):375–384. [PubMed: 22972495]

54. Nieman RB, Fleming J, Coker RJ, Harris JR, Mitchell DM. The effect of cigarette smoking on the development of AIDS in HIV-1-seropositive individuals. AIDS. 1993; 7(5):705–710. [PubMed: 8318178]

55. Obirikorang C, Yeboah FA. Blood haemoglobin measurement as a predictive indicator for the progression of HIV/AIDS in resource-limited setting. Journal of Biomedical Science. 2009; 16(1): 102. [PubMed: 19922646]

56. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics. 1988:1033–1048. [PubMed: 3233244]

57. Regan MM, Catalano PJ. Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. Biometrics. 1999; 55(3):760–768. [PubMed: 11315004]

58. Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. Biometrics. 2000; 56(4):1047–1054. [PubMed: 11129460]

59. Rüschendorf L. Stochastically ordered distributions and monotonicity of the OC-function of sequential probability ratio tests. Statistics. 1981; 12(3):327–338.

60. Rüschendorf L. On the distributional transform, Sklar's theorem, and the empirical copula process. Journal of statistical planning and inference. 2009; 139(11):3921–3927.

61. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. Journal of the Royal Statistical Society: Series B. 1997; 59(3):667–678.

62. Shiffman S. Relapse following smoking cessation: A situational analysis. Journal of Consulting and Clinical Psychology. 1982; 50(1):71–86. [PubMed: 7056922]

63. Shiffman S, Balabanis M, Fertig J, Allen J. Associations between alcohol and tobacco. Alcohol and Tobacco: From Basic Science to Clinical Practice NIAAA Research Monograph. 1995; 30:17–36.

64. Shiffman S, Gwaltney CJ, Balabanis MH, Liu KS, Paty JA, Kassel JD, Hickcox M, Gnys M. Immediate antecedents of cigarette smoking: An analysis from ecological momentary assessment. Journal of Abnormal Psychology. 2002; 111(4):531–545. [PubMed: 12428767]

65. Shiffman S, Hickcox M, Paty JA, Gnys M, Kassel JD, Richards TJ. Progression from a smoking lapse to relapse: Prediction from abstinence violation effects, nicotine dependence, and lapse characteristics. Journal of Consulting and Clinical Psychology. 1996; 64(5):993–1002. [PubMed: 8916628]

66. Sklar A. Fonctions de répartition à n dimensions et leurs marges. Publications de L'Institut de Statistiques de l'Universite de Paris. 1959; 8:229–231.

67. Smith M, Min A, Almeida C, Czado C. Modeling longitudinal data using a pair-copula decomposition of serial dependence. Journal of the American Statistical Association. 2010; 105(492):1467–1479.

68. Smith MS, Gan Q, Kohn RJ. Modelling dependence using skew t copulas: Bayesian inference and applications. Journal of Applied Econometrics. 2012; 27(3):500–522.

69. Song PXK. Multivariate dispersion models generated from Gaussian copula. Scandinavian Journal of Statistics. 2000; 27(2):305–320.

70. Song, PX-K. Springer Series in Statistics. Vol. Chapter 6. New York: Springer; 2007. Correlated Data Analysis: Modeling, Analytics, and Applications; p. 121-155.

71. Song PXK, Li M, Yuan Y. Joint regression analysis of correlated data using Gaussian copulas. Biometrics. 2009; 65(1):60–68. [PubMed: 18510653]

72. U.S. Department of Health and Human Services. The health consequences of smoking: a report of the surgeon general. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2004. p. 62

73. Verbeke, G., Molenberghs, G., Rizopoulos, D. Random effects models for longitudinal data. In: van Montfort, K.Oud, J., Satorra, A., editors. Longitudinal Research with Latent Variables. Springer-Verlag; Berlin: 2010. p. 37-96.

74. Wu B, de Leon AR. Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology. Journal of Agricultural, Biological, and Environmental Statistics. 2014; 19(1):39–56.

75. Zeger SL, Diggle P. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics. 1994; 50(3):689–699. [PubMed: 7981395]

76. Zhang W, Lee SY, Song X. Local polynomial fitting in semivarying coefficient model. Journal of Multivariate Analysis. 2002; 82(1):166–188.

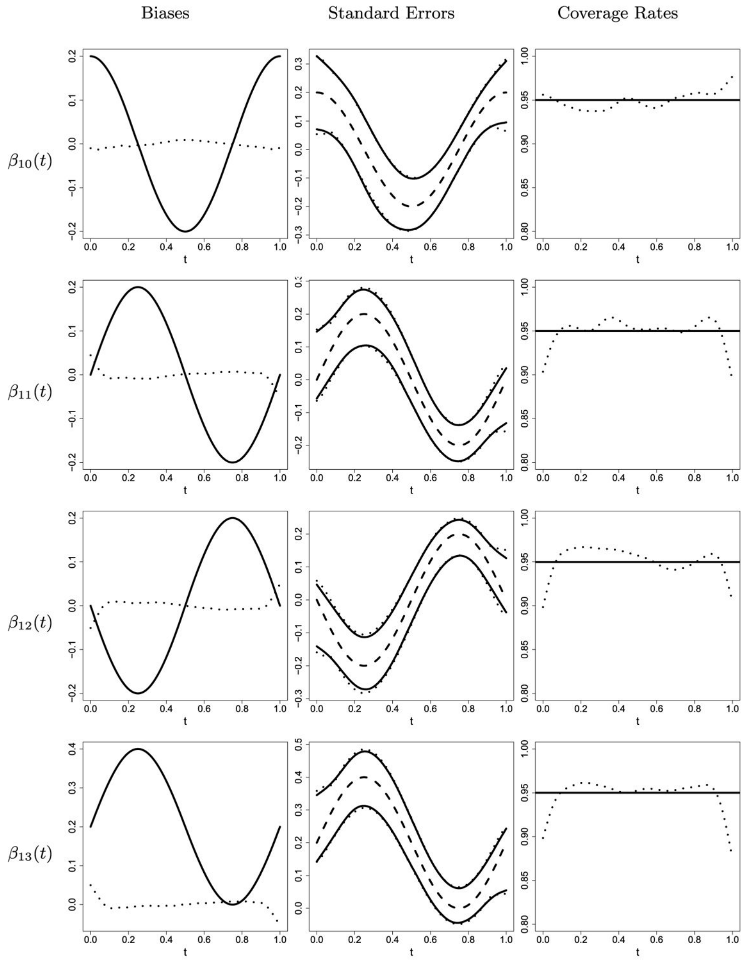

Figure 1.

Part 1 of the results from our simulation study. Each row shows three plots for a given time-varying parameter. The first plot shows the true function (solid) and the empirical bias of our estimator (dotted). The second plot shows the true function (dashed), the empirical pointwise 95% confidence band (solid), and the mean theoretical pointwise 95% confidence band (dotted). The third plot shows the desired coverage rate (solid) and the empirical pointwise coverage rates (dotted).

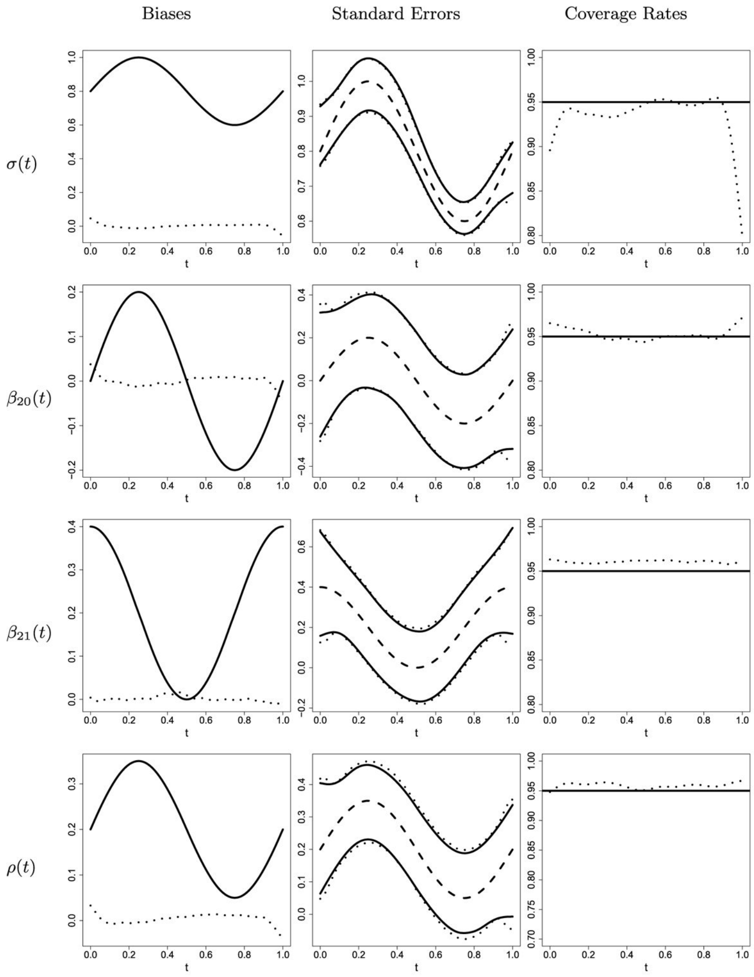

Figure 2.

Part 2 of the results from our simulation study. Each row shows three plots for a given time-varying parameter. The first plot shows the true function (solid) and the empirical bias of our estimator (dotted). The second plot shows the true function (dashed), the empirical pointwise 95% confidence band (solid), and the mean theoretical pointwise 95% confidence band (dotted). The third plot shows the desired coverage rate (solid) and the empirical pointwise coverage rates (dotted).

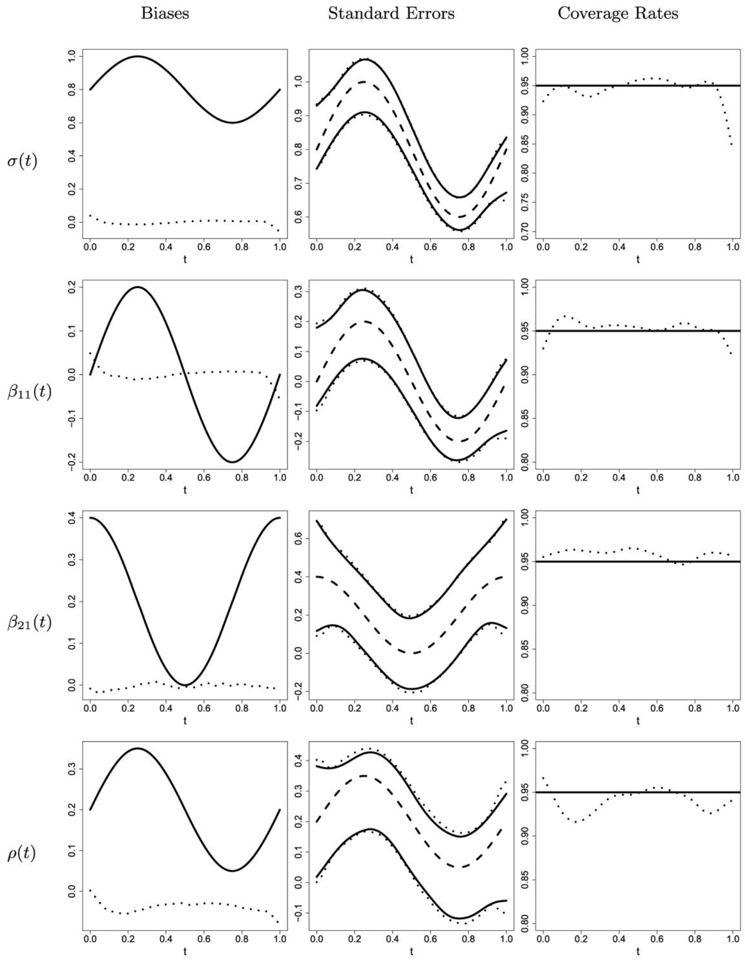

Figure 3.

Selected results from our simulation study with MCAR. Each row shows three plots for a given time-varying parameter. The first plot shows the true function (solid) and the empirical bias of our estimator (dotted). The second plot shows the true function (dashed), the empirical pointwise 95% confidence band (solid), and the mean theoretical pointwise 95% confidence band (dotted). The third plot shows the desired coverage rate (solid) and the empirical pointwise coverage rates (dotted).

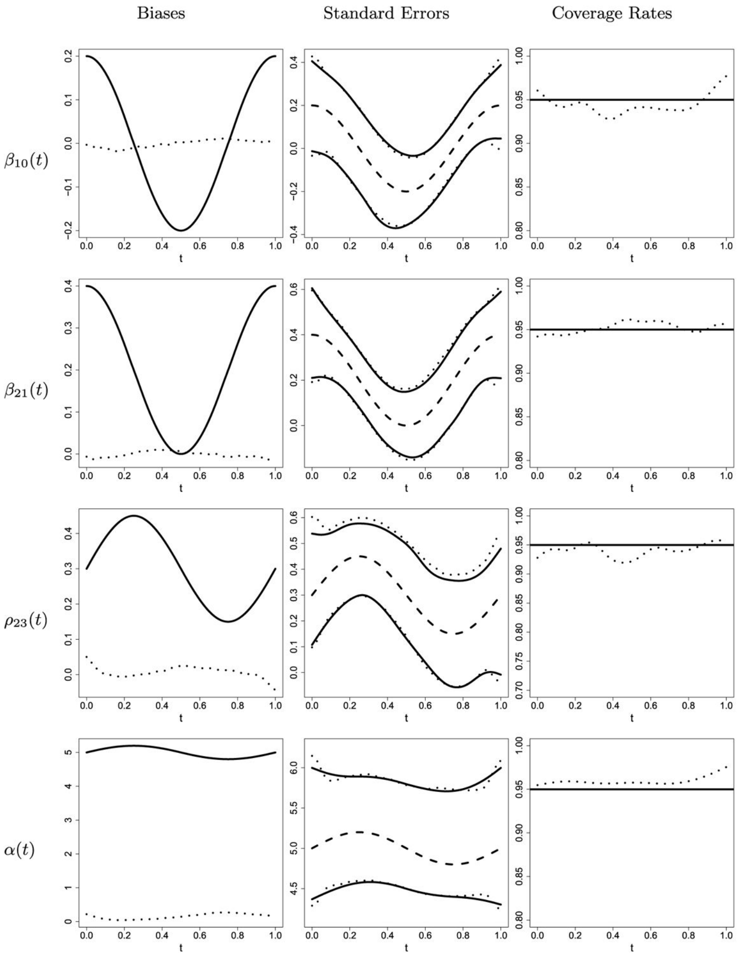

Figure 4.

Selected results from our trivariate simulation study. Each row shows three plots for a given time-varying parameter. The first plot shows the true function (solid) and the empirical bias of our estimator (dotted). The second plot shows the true function (dashed), the empirical pointwise 95% confidence band (solid), and the mean theoretical pointwise 95% confidence band (dotted). The third plot shows the desired coverage rate (solid) and the empirical pointwise coverage rates (dotted).

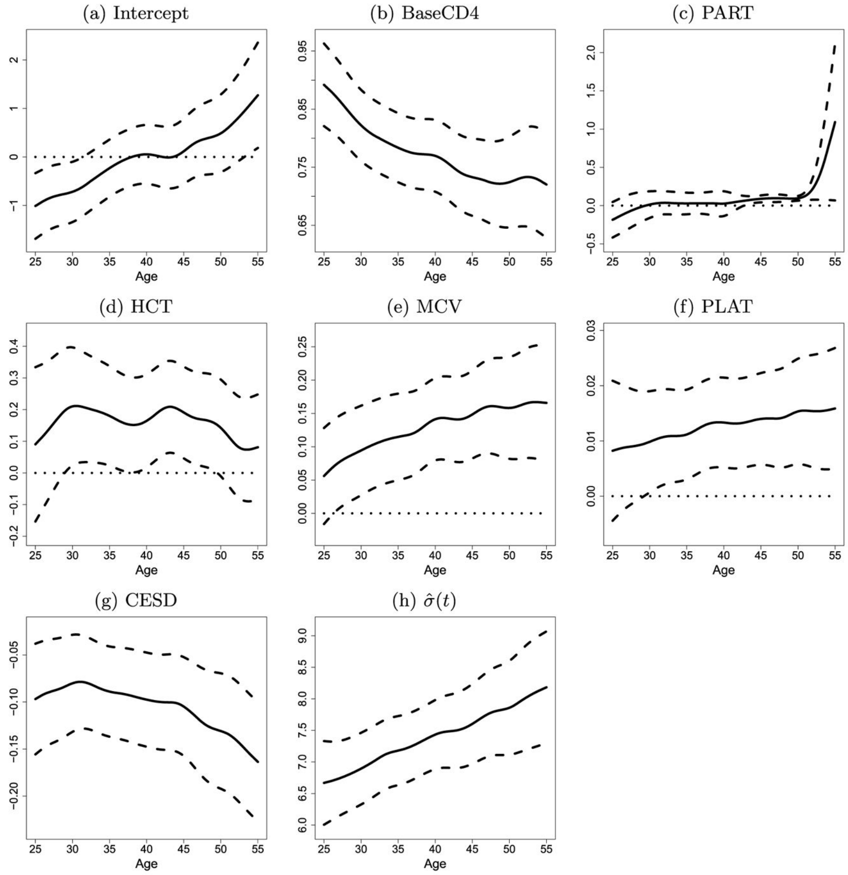

Figure 5.

The results of our data analysis for the continuous response, CD4 cell percentage. For each panel, the solid curve shows the estimate, the dashed curves show the estimated 95% pointwise confidence band, and the dotted line marks zero.

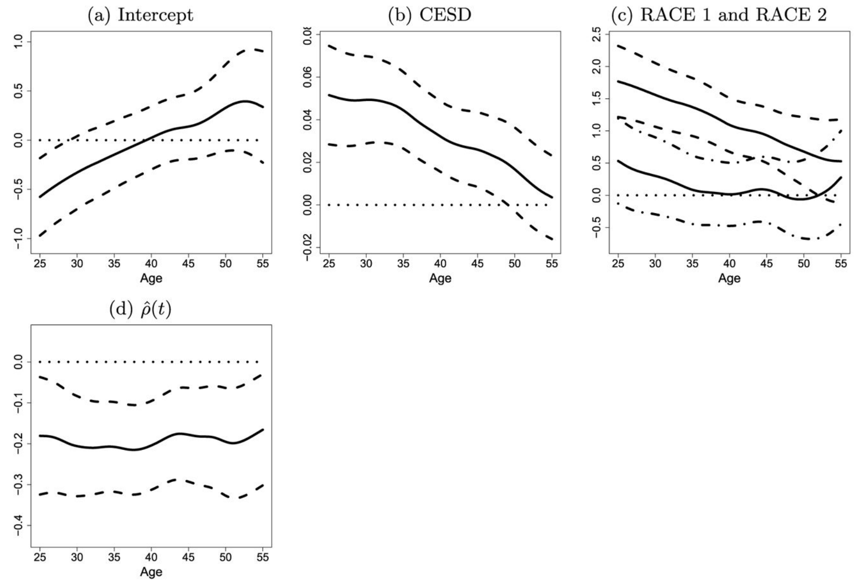

Figure 6.

The results of our data analysis for the binary response (smoking status), and the estimated time-varying association between smoking status and CD4 cell percentage. For each panel, the solid curve shows the estimate, the dashed (or dashed-dotted) curves show the estimated 95% pointwise confidence band, and the dotted line marks zero.

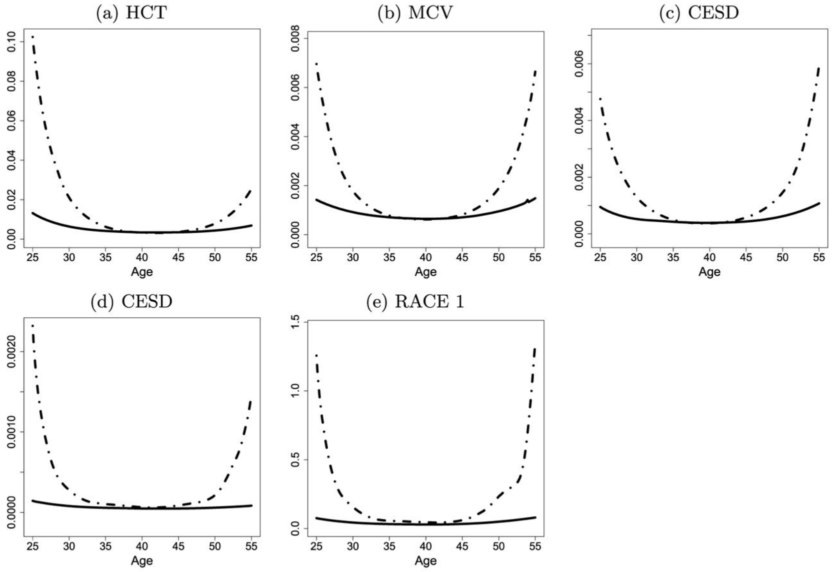

Figure 7.

Comparison of standard errors for our joint analysis and separate univariate analyses of CD4 cell percentage and smoking status. In each univariate analysis, the other outcome was included as an additional predictor. Panels (a), (b), and (c) show results for CD4 cell percentage, and panels (d) and (e) show results for smoking status. For each panel, the solid curve and dashed-dotted curve show the standard errors for the joint and univariate models, respectively.

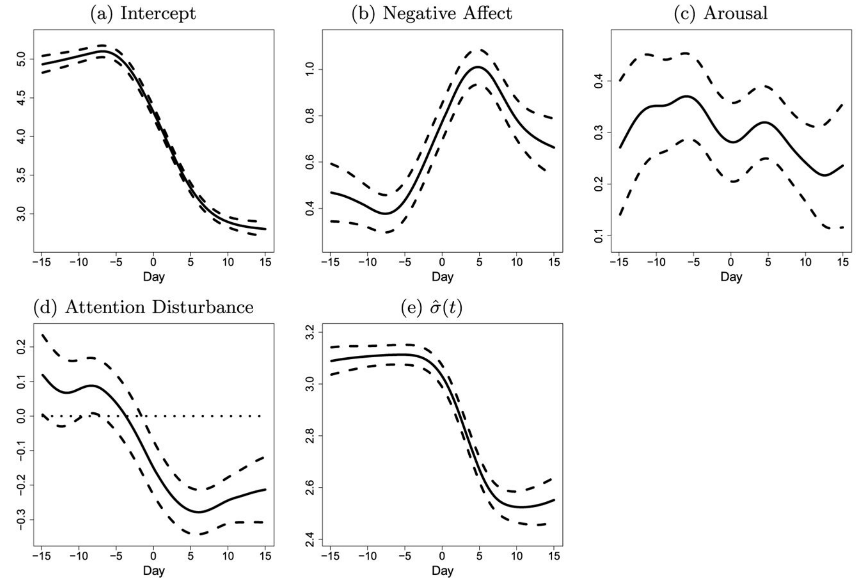

Figure 8.

The results of our data analysis for the continuous response, urge to smoke. For each panel, the solid curve shows the estimate, the dashed curves show the estimated 95% pointwise confidence band, and the dotted line marks zero.

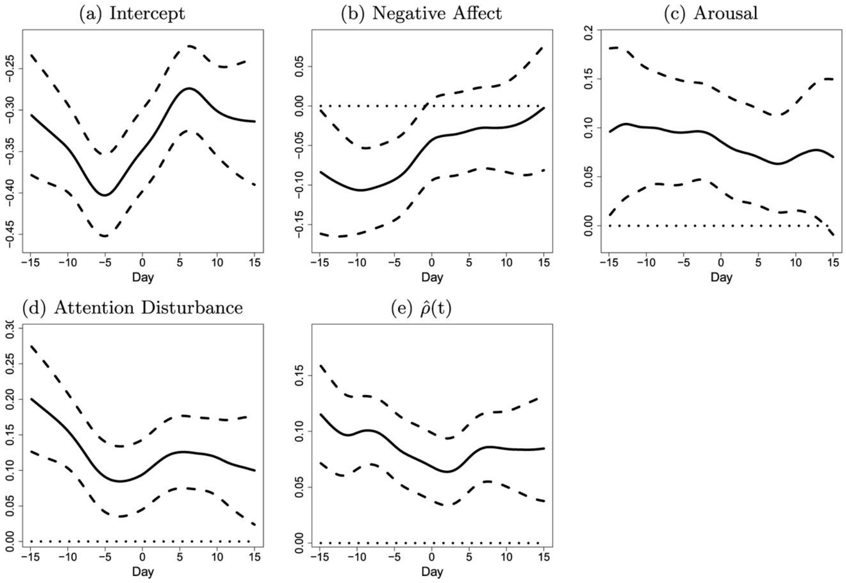

Figure 9.

The results of our data analysis for the binary response (potential factors leading to smoking), and the estimated time-varying association between this binary response and urge to smoke. For each panel, the solid curve shows the estimate, the dashed curves show the estimated 95% pointwise confidence band, and the dotted line marks zero.