Biostatistics (2006), **7**, 2, *pp.* 235–251
doi:10.1093/biostatistics/kxj004
Advance Access publication on October 26, 2005

Covariate-adjusted varying coefficient models

DAMLA ŞENTÜRK*

Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
dsenturk@stat.psu.edu

Summary

Covariate-adjusted regression was recently proposed for situations where both predictors and response in a regression model are not directly observed, but are observed after being contaminated by unknown functions of a common observable covariate. The method has been appealing because of its flexibility in targeting the regression coefficients under different forms of distortion. We extend this methodology proposed for regression into the framework of varying coefficient models, where the goal is to target the covariate-adjusted relationship between longitudinal variables. The proposed method of covariate-adjusted varying coefficient model (CAVCM) is illustrated with an analysis of a longitudinal data set containing calcium absorbtion and intake measurements on 188 subjects. We estimate the age-dependent relationship between these two variables adjusted for the covariate body surface area. Simulation studies demonstrate the flexibility of CAVCM in handling different forms of distortion in the longitudinal setting.

Keywords: Covariate-adjusted regression; Local polynomial regression; Longitudinal data; Multiplicative effects; Smoothing.

1. Introduction

Covariate-adjusted regression (CAR) has been proposed by Şentürk and Müller (2005a,b) as an adjustment for the multiplicative effects of a covariate on the response and the predictor, in a regression setting. The methodology was motivated by a study on hemodialysis patients where the regression relation between plasma fibrinogen concentration and serum transferrin level was of interest (Kaysen *et al.*, 2003). Body mass index (kg/m^2) was identified as a common covariate affecting both the variables in this study. One way of adjustment used for body mass index is dividing the variables of interest by it. Normalization through dividing by a covariate is common in medical studies, where usually some body configuration measurement like body weight, height, or body mass index is considered as a covariate with multiplicative effects on the variables of interest. Even though the distortion is thought to be multiplicative, there is still uncertainty about the exact form of distortion in most cases. Şentürk and Müller proposed a general adjustment for such data, retaining the multiplicative form but still reflecting the uncertainty by modeling the effects of the covariate through unknown smooth functions. The observed response and predictor values for the *i*th subject are denoted by

$$\tilde{Y}_i = \psi(U_i)Y_i \quad \text{and} \quad \tilde{X}_i = \phi(U_i)X_i, \qquad i = 1, \ldots, n,$$

*To whom correspondence should be addressed.

for a sample of n subjects, where $\psi(\cdot)$ and $\phi(\cdot)$ are unknown smooth functions of the observed covariate U. Here, Y and X denote the underlying unobserved parts of the observed response and predictor that are thought to be measured in a scale that does not depend on U. CAR uncovers the regression relationship adjusted for the covariate by giving consistent estimates for the parameters in the underlying, unobserved regression model

$$Y_i = \beta_0 + \beta_1 X_i + e_i,$$

based on the observed data $(\tilde{Y}_i, \tilde{X}_i, U_i)$, $i = 1, \ldots, n$.

The procedure utilizes the identifiability condition of no average distortion, i.e. $E(\tilde{Y}) = E(Y)$ and $E(\tilde{X}) = E(X)$. Under this identifiability condition, CAR gives consistent estimates regardless of the form of distortion considered as long as the distortion on the response and the predictors are of the same form (see Şentürk and Müller, 2005a, for justification). More specifically, CAR yields consistent estimates under multiplicative (i.e. $\tilde{Y}_i = \psi(U_i)Y_i$, $\tilde{X}_i = \phi(U_i)X_i$), additive (i.e. $\tilde{Y}_i = \psi(U_i) + Y_i$, $\tilde{X}_i = \phi(U_i) + X_i$), and no distortion (i.e. $\tilde{Y}_i = Y_i$, $\tilde{X}_i = X_i$). This makes CAR a very flexible adjustment where even the form of the distortion need not be known.

We propose an extension of the CAR algorithm for longitudinal data, where the measurements taken on the response and the predictor are time dependent. We focus on the simple case of a cross-sectional covariate; however, the proposed method can also be applied for the case of a longitudinal covariate as will be discussed in the Section 6. One example is a study on 188 subjects, where the longitudinal relationship between calcium intake and absorbtion is of interest (Davis, 2002). The covariate to be adjusted for is body surface area (BSA) of the subjects and a new adjustment procedure is needed to uncover the time-dependent relationship between the underlying variables adjusted for this covariate. Denote the response and the predictor measurements for the ith subject, $i = 1, \ldots, n$, taken at time t_{ij}, $j = 1, \ldots, T_i$, as

$$\tilde{Y}_i(t_{ij}) = \psi(U_i)Y_i(t_{ij}) \quad \text{and} \quad \tilde{X}_i(t_{ij}) = \phi(U_i)X_i(t_{ij}).$$

Here, Y and X denote the underlying unobserved longitudinal variables assumed to be related through the varying coefficient model

$$Y_i(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})X_i(t_{ij}) + e_i(t_{ij}) \tag{1.1}$$

(Hastie and Tibshirani, 1993). Varying coefficient models are an extension of the regression models where the coefficients are allowed to vary as smooth functions of a covariate possibly different than the predictors. They have been especially popular in applications to longitudinal data, where the coefficient functions vary as functions of time. They reduce the modeling bias with their unique structure while also avoiding the ‘curse of dimensionality’ problem. Wu and Yu (2002) give an overview of applications to longitudinal data, where the proposed estimation procedures include Hoover *et al.* (1998), Wu and Chiang (2000), Fan and Zhang (2000), and Wu *et al.* (2000) on local least squares, Hoover *et al.* (1998) and Chiang *et al.* (2001) on smoothing splines, and Huang *et al.* (2004) on basis approximations.

The central goal of this paper is the estimation of the smooth coefficient functions, $\beta_0(\cdot)$ and $\beta_1(\cdot)$ in (1.1) based on observations of the covariate U_i, and the contaminated observations on the response and the predictor $\{\tilde{Y}_i(t_{ij}), \tilde{X}_i(t_{ij})\}$. A key observation to reach this goal is that regressing $\tilde{Y}$ on $\tilde{X}$ leads to another varying coefficient model, where the coefficient functions depend both on time and the covariate U. This is illustrated in Section 2, where more details on the covariate-adjusted varying coefficient model (CAVCM) with multiple predictors are provided. A two-step procedure is proposed for estimation in the CAVCM in Section 3, motivated by the two-step procedure of Fan and Zhang (2000) proposed for estimation in varying coefficient models. As also argued by Fan and Zhang, a common feature of many longitudinal studies is that the measurements are collected at the same time points for all subjects with possibly missing values at few time points for some subjects. Let $\{t_j, j = 1, \ldots, T\}$ be the distinct time points among all t_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, T_i$. The proposed algorithm targets the raw estimates

$\hat{\beta}_0(t_j)$ and $\hat{\beta}_1(t_j)$, $j = 1, \ldots, T$, of the varying coefficient functions by fitting CAR to the data collected at each distinct time point t_j, in the first step. The CAR algorithm is appropriate for this step, since the data observed at each time point t_j is independent, collected from different subjects. The final estimates of the coefficient functions in (1.1) are obtained in the second step by smoothing the scatter plot of the raw estimates, $\{t_j, \hat{\beta}_r(t_j)\}_{j=1}^{T}$, for each component r, $r = 0, 1$, separately. The second step consists of only one-dimensional smoothing procedures, and can be carried out with any smoothing technique. Therefore, the proposed estimation procedure is fast, intuitive, and easy to implement with any standard software containing least-squares procedures.

The proposed estimation procedure for CAVCM also enjoys the same attraction as CAR in that it targets the coefficient functions regardless of the form of distortion considered under the identifiability conditions considered for CAVCM discussed in Section 2. In other words, the proposed estimation procedure targets the coefficient functions not only for the multiplicative distortion (i.e. $\tilde{Y}_i(t_j) = \psi(U_i)Y_i(t_j)$, $\tilde{X}_i(t_j) = \phi(U_i)X_i(t_j)$) but also for additive (i.e. $\tilde{Y}_i(t_j) = \psi(U_i) + Y_j(t_j)$, $\tilde{X}_i(t_j) = \phi(U_i) + X_i(t_j)$) and no distortion (i.e. $\tilde{Y}_i(t_j) = Y_i(t_j)$, $\tilde{X}_i(t_j) = X_i(t_j)$) as demonstrated through simulation studies in Section 5. Another advantage of the proposed estimation procedure shown in Section 5 is that it can handle missing values easily. If there are not enough subjects observed at a given time point t_j to fit CAR, the missing raw estimate at t_j is imputed through the smoothing in the second step. Application of the proposed method to the longitudinal calcium data can be found in Section 4.

2. Covariate-adjusted varying coefficient models

Consider the general case of an underlying varying coefficient model with p predictors,

$$Y_i(t_j) = \beta_0(t_j) + \sum_{r=1}^{p} \beta_r(t_j)X_{ri}(t_j) + e_i(t_j), \tag{2.1}$$

evaluated at T distinct time points, t_j, $j = 1, \ldots, T$. Here $e_i(t_j)$ is a zero-mean stochastic process with covariance function $\delta(t_j, t_{j'}) = \text{cov}\{e_i(t_j), e_i(t_{j'})\}$, and $\beta_0(\cdot), \beta_1(\cdot), \ldots, \beta_p(\cdot)$ are the unknown coefficient functions of interest. In the varying coefficient model (2.1), Y and X_r are not observable. Instead, one observes distorted versions $(\tilde{Y}, \tilde{X}_r)$, along with a univariate covariate U, where

$$\tilde{Y}_i(t_j) = \psi(U_i)Y_i(t_j) \quad \text{and} \quad \tilde{X}_{ri}(t_j) = \phi_r(U_i)X_{ri}(t_j), \tag{2.2}$$

for $r = 1, \ldots, p$, and ϕ_r and ψ are unknown smooth functions of U. The identifiability conditions considered are an extension of the no-average distortion condition used for CAR. They entail no average distortion at distinct time points t_j, i.e. $E\{\tilde{Y}(t_j)\} = E\{Y(t_j)\}$ and $E\{\tilde{X}(t_j)\} = E\{X(t_j)\}$, for $j = 1, \ldots, T$. The identifiability conditions can equivalently be written as conditions on the unknown smooth distortion functions as

$$E\{\psi(U_i)\} = 1, \quad E\{\phi_r(U_i)\} = 1. \tag{2.3}$$

Model (2.1)–(2.3) will be referred to as the CAVCM.

A central goal is to obtain estimators of the coefficient functions in model (2.1), given the observations of the covarite U and the distorted observations $(\tilde{Y}, \tilde{X}_r)$ in (2.2). The key to the estimation of the targeted regression functions $\{\beta_r(\cdot)\}$ is to express the regression of $\tilde{Y}$ on $\{\tilde{X}_r\}_{r=0}^{p}$ as another varying coefficient model. More precisely, under the assumption that $(e(\cdot), U, X_r(\cdot))$ $(r = 1, \ldots, p)$ are mutually independent at each fixed time point, the regression of $\tilde{Y}$ on $\{\tilde{X}_r\}_{r=0}^{p}$ can be expressed as

$$E\{\tilde{Y}(t_j)|\tilde{X}_1(t_j), \ldots, \tilde{X}_p(t_j), U\} = \gamma_0(U, t_j) + \sum_{r=1}^{p} \gamma_r(U, t_j)\tilde{X}_r(t_j),$$

where

$$\gamma_0(U, t_j) = \psi(U)\beta_0(t_j) \quad \text{and} \quad \gamma_r(U, t_j) = \beta_r(t_j)\frac{\psi(U)}{\phi_r(U)}.$$

Therefore,

$$\tilde{Y}_i(t_j) = \gamma_0(U_i, t_j) + \sum_{r=1}^{p} \gamma_r(U_i, t_j)\tilde{X}_{ri}(t_j) + \epsilon(U_i, t_j), \tag{2.4}$$

with $\epsilon(U_i, t_j) \equiv \psi(U_i)e_i(t_j)$. The assumption that the underlying predictors, $X_r(\cdot)$, and response, $Y(\cdot)$, are independent of the contaminating variable U is an assumption defining the proposed contamination setting through defining these unobserved, underlying variables, and for that matter is not one that can be checked in practice. Thus, the question to be answered in practice is whether or not these independence conditions help define interpretable latent variables of interest from their observable counterparts. In the calcium data analyzed, the interpretations of the latent variables are BSA-adjusted calcium intake and absorbtion.

In the varying coefficient model given in (2.4), the observed variables vary according to two variables instead of one, the covariate U and time, resulting in two-dimensional coefficient functions. Since the variables $\tilde{Y}_i(t_j)$, $\tilde{X}_{ri}(t_j)$, and U_i and the time points are all observable, we first target these estimable two-dimensional coefficient functions, $\gamma_r(\cdot, \cdot)$, through their one-dimensional projections at each time point t_j. The underlying one-dimensional coefficient functions of interest, $\beta_r(\cdot)$, are then targeted using estimates of $\gamma_r(\cdot, \cdot)$ and the identifiability conditions given in (2.3).

3. Two-step estimation procedure

The proposed estimation algorithm is based on a similar idea as the two-step procedure proposed by Fan and Zhang (2000), for estimation in varying coefficient models. Fan and Zhang considered estimation in (2.1), when the longitudinal response and predictors can be observed directly, free from any distorting effects. A common feature of many longitudinal studies is that measurements are collected at the same time points, $\{t_j, j = 1, \ldots, T\}$, for all subjects with possibly missing values at few time points for some subjects. Noting that a different linear regression between the response and the predictors applies for each time point in a varying coefficient model, they regress the response on the predictors at a fixed time point t_j to obtain the raw estimates for the smooth coefficient functions $\beta_0(t_j), \beta_1(t_j), \ldots, \beta_p(t_j)$ in the first step. In the second step, the scatter plots of raw estimates for the rth coefficient function are smoothed versus the time points, for each component r separately, to obtain the final smooth estimates for the coefficient functions.

Even though this two-step estimation procedure is easy to implement, involving only linear regression fits and one-dimensional smoothing procedures, it will not be applicable when the longitudinal response Y and predictors X_r are not observed directly. In addition, regressing the observed distorted response $\tilde{Y}(t_j)$ on the predictors $\tilde{X}_r(t_j)$ in the first step of the algorithm will not target $\beta_r(t_j)$ of the underlying varying coefficient model under the CAVCM. More specifically, it follows from equation (2.4) of Şentürk and Müller (2005a) that under the multiplicative distorting effects given in (2.2), the raw estimates of Fan and Zhang for $\beta_1(t_j)$ evaluated at each time point t_j in the simple case of one predictor targets

$$\xi = \frac{E\{\phi(U)\psi(U)\}[\beta_0(t_j)E\{X_1(t_j)\} + \beta_1(t_j)E\{X_1^2(t_j)\}] - \beta_1(t_j)[E\{X_1(t_j)\}]^2 - \beta_0(t_j)E\{X_1(t_j)\}}{E\{\phi^2(U)\}E\{X_1^2(t_j)\} - [E\{X_1(t_j)\}]^2} \tag{3.1}$$

instead of $\beta_1(t_j)$. It is also shown that ξ can assume any real value, resulting possibly in arbitrarily large biases for the raw estimates of Fan and Zhang if the distortion covariate is ignored within the CAVCM.

The bias created by the final smooth estimates of Fan and Zhang will be investigated in Section 5, through simulation studies.

Our proposal is geared towards handling distortions on the response and predictors, as commonly encountered in medical data. Rather than fitting a linear regression between $\tilde{Y}$ and $\tilde{X}_r$ observed at each time point t_j, we fit CAR between $\tilde{Y}$ and $\tilde{X}_r$, adjusting for U, to obtain the raw estimates $\hat{\beta}_0(t_j), \hat{\beta}_1(t_j), \ldots, \hat{\beta}_p(t_j)$ in the first step. This is motivated by the fact that a different one-dimensional varying coefficient model holds at each time point t_j in the two-dimensional varying coefficient model given in (2.4). This one-dimensional varying coefficient model can be expressed as,

$$\tilde{Y}_i(t_j) = \gamma_{0j}(U_i) + \sum_{r=1}^{p} \gamma_{rj}(U_i)\tilde{X}_{ri}(t_j) + \epsilon_j(U_i), \tag{3.2}$$

where only data observed at a fixed time point t_j is used, and the coefficient functions vary depending on U only. Here the one-dimensional coefficient functions will be related to the constants $\beta_0(t_j), \beta_1(t_j), \ldots, \beta_p(t_j)$ through the equations

$$\gamma_{0j}(U_i) = \psi(U_i)\beta_0(t_j) \quad \text{and} \quad \gamma_{rj}(U_i) = \beta_r(t_j)\frac{\psi(U_i)}{\phi_r(U_i)}, \tag{3.3}$$

where not necessarily all n subjects but say n_j subjects may be observed at time t_j, $i = 1, \ldots, n_j$. We first target the coefficient functions $\gamma_{rj}(\cdot)$ in (3.2) through local linear fits, and then arrive at the raw estimate of $\beta_r(t_j)$ through a weighted average of the estimates of $\gamma_{rj}(\cdot)$, making use of the relations in (3.3) and the identifiability conditions. We use local polynomial regressions to fit CAR at each time point in the first step of the algorithm, as it is shown through simulation studies to have the best small sample performance yielding the smallest mean squared error compared to other binning algorithms (Şentürk and Nguyen, 2005). The second step similarly consists of smoothing the scatter plot of each coefficient component $\{t_j, \hat{\beta}_r(t_j)\}_{j=1}^{T}$ to obtain the final estimates $\tilde{\beta}_0(t_j), \tilde{\beta}_1(t_j), \ldots, \tilde{\beta}_p(t_j)$.

3.1 *Step 1: obtaining the raw estimates*

Denote the available (observed) data at time t_j by $\{U_i, \tilde{\mathbf{X}}_i(t_j), \tilde{Y}_i(t_j)\}$, $i = 1, \ldots, n_j$, for a sample of size n_j, where $\tilde{\mathbf{X}}_i(t_j) = (\tilde{X}_{1i}(t_j), \ldots, \tilde{X}_{pi}(t_j))^{\mathrm{T}}$ are the p-dimensional predictors. The function $\gamma_{rj}(U)$ in (3.2) can be approximated based on local polynomial modeling as

$$\gamma_{rj}(U) \approx \sum_{k=0}^{q} \frac{1}{k!}\gamma_{rj}^{(k)}(u)(U-u)^k, \quad r = 0, 1, \ldots, p,$$

for U in a neighborhood of u. Here, $\gamma_{rj}^{(k)}$ denotes the kth derivative of $\gamma_{rj}(\cdot)$. Consider the local linear least-squares estimator of γ_{rj} through the minimization of

$$\sum_{i=1}^{n}\left[\tilde{Y}_i(t_j) - \sum_{r=0}^{p}\{\alpha_{rj,0} + \alpha_{rj,1}(U_i - u)\}\tilde{X}_{ri}(t_j)\right]^2 K_h(U_i - u) \tag{3.4}$$

with respect to $\alpha_{rj,0}$ and $\alpha_{rj,1}$ for a specified kernel function K with bandwidth h where $K_h(\cdot) = K(\cdot/h)/h$. We choose to consider local linear fits for computational simplicity, as they are comparable to local cubic fits for all practical purposes in implementation. Note that $\tilde{X}_{0i}(t_j) = 1$, corresponding to the intercept function $\gamma_{0j}(U)$. Minimization of criterion (3.4) is a weighted least-squares problem. Assuming that $\mathcal{X}^{\mathrm{T}}(t_j)\mathbf{W}\mathcal{X}(t_j)$ is nonsingular, the solution is

$$\hat{\boldsymbol{\alpha}}_j = (\mathcal{X}^{\mathrm{T}}(t_j)\mathbf{W}\mathcal{X}(t_j))^{-1}\mathcal{X}^{\mathrm{T}}(t_j)\mathbf{W}\tilde{\mathbf{Y}}(t_j),$$

where $\mathcal{X}(t_j)$ is the following $n_j \times 2(p+1)$ matrix

$$\mathcal{X}(t_j) = \begin{bmatrix} 1 & (U_1 - u) & \tilde{X}_{11}(t_j) & (U_1 - u)\tilde{X}_{11}(t_j) & \cdots & \tilde{X}_{p1}(t_j) & (U_1 - u)\tilde{X}_{p1}(t_j) \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & (U_n - u) & \tilde{X}_{1n}(t_j) & (U_n - u)\tilde{X}_{1n}(t_j) & \cdots & \tilde{X}_{pn}(t_j) & (U_n - u)\tilde{X}_{pn}(t_j) \end{bmatrix},$$

$$\mathbf{W} = \mathrm{diag}\{K_h(U_1 - u), \ldots, K_h(U_n - u)\} \quad \text{and} \quad \tilde{\mathbf{Y}}(t_j) = (\tilde{Y}_1(t_j), \ldots, \tilde{Y}_n(t_j))^{\mathrm{T}}.$$

The local least-squares estimator of $\gamma_{rj}(u)$ is given by

$$\hat{\gamma}_{rj}(u) = \mathbf{e}^{\mathrm{T}}_{2r+1,2(p+1)}\hat{\boldsymbol{\alpha}}_j = \mathbf{e}^{\mathrm{T}}_{2r+1,2(p+1)}(\mathcal{X}^{\mathrm{T}}(t_j)\mathbf{W}\mathcal{X}(t_j))^{-1}\mathcal{X}^{\mathrm{T}}(t_j)\mathbf{W}\tilde{\mathbf{Y}}(t_j), \quad r = 0, \ldots, p,$$

where $\mathbf{e}_{2r+1,2(p+1)}$ is a unit vector of length $2(p+1)$ with 1 in position $2r+1$.

The estimators of the targeted regression parameters, $\{\beta_r(t_j)\}_{r=0}^{p}$, are then obtained by averaging over the raw estimates, $\hat{\gamma}_{rj}(U_i)$, this time evaluated at the original observations of the covariate $(U_i)_{i=1}^{n}$. More precisely,

$$\hat{\gamma}_{rj}(U_i) = \mathbf{e}^{\mathrm{T}}_{2r+1,2(p+1)}(\mathcal{X}_i^{\mathrm{T}}(t_j)\mathbf{W}_i\mathcal{X}_i(t_j))^{-1}\mathcal{X}_i^{\mathrm{T}}(t_j)\mathbf{W}_i\tilde{\mathbf{Y}}(t_j),$$

where $\mathcal{X}_i(t_j)$ and $\mathbf{W}_i$ are $\mathcal{X}(t_j)$ and $\mathbf{W}$ with $u = U_i$. This leads to the following raw estimates:

$$\hat{\beta}_0(t_j) = n_j^{-1}\sum_{i=1}^{n}\hat{\gamma}_{0j}(U_i) \quad \text{and} \quad \hat{\beta}_r(t_j) = \frac{1}{\bar{\tilde{X}}_r(t_j)}\sum_{i=1}^{n_j}\frac{1}{n_j}\hat{\gamma}_{rj}(U_i)\tilde{X}_{ri}(t_j), \tag{3.5}$$

where $\bar{\tilde{X}}_r(t_j) = n_j^{-1}\sum_{i=1}^{n_j}\tilde{X}_{ri}(t_j)$. The estimates are motivated by the relations $E\{\gamma_{0j}(U)\} = \beta_0(t_j)$ and $E\{\gamma_{rj}(U)\tilde{X}_r(t_j)\} = \beta_r(t_j)E\{\psi(U)X_r(t_j)\} = \beta_r(t_j)E\{X_r(t_j)\} = \beta_r(t_j)E\{\tilde{X}_r(t_j)\}$ that follow directly from (2.3) and (3.3). An important assumption here is that $E\{\tilde{X}_r(t_j)\}$ or equivalently $E\{X_r(t_j)\}$ is not equal to 0, since it is targeted by the denominator of $\hat{\beta}_r(t_j)$ in (3.5). The consistency of $\{\hat{\beta}_r(t_j)\}_{r=0}^{p}$ for $\{\beta_r(t_j)\}_{r=0}^{p}$ has been shown in Şentürk and Nguyen (2005). As outlined by Şentürk and Nguyen (2005), we utilize the generalized cross-validation (GCV) criterion proposed by Wahba (1977) and Craven and Wahba (1979) for the selection of the bandwidth h in the first step of the proposed algorithm with local polynomial modeling. Note that other criteria can be used for bandwidth selection at this step. The literature includes studies of Zhang *et al.* (1998) and Zhang (2004) that utilized the double-penalized quasi-likelihood approach to estimate the smoothing parameters and the nonparametric functions simultaneously in a mixed model framework.

3.2 *Step 2: obtaining the final smooth estimates*

The final smooth estimates of $\beta_r(t)$ are computed in the second step of the algorithm for each component r, $r = 0, 1, \ldots, p$, separately as

$$\tilde{\beta}_r(t) = \sum_{j=1}^{T} w_r(t_j, t)\hat{\beta}_r(t_j).$$

The weights $w_r(t_j, t)$ in the above expression can be obtained from any linear smoothing technique, such as local polynomial smoothing used by Fan and Zhang (2000) or spline smoothing used by Wu *et al.* (2000). This additional smoothing step is needed to bring in information from neighboring time points, improving on the efficiency of the estimates. It can be easily carried out, using any convenient software, as it only involves one-dimensional smoothing. Another benefit of this one-dimensional smoothing procedure is that it would be easier to choose a suitable bandwidth for each component separately through visualization of the data. This second step is also crucial in providing flexibility in dealing with missing values. If there are not enough patients observed at a particular time point to fit CAR (a minimum of roughly 20–30 observations are needed to fit CAR with one predictor as determined through simulation studies), raw estimates at that time point will be considered missing. These missing values can be imputed in the second smoothing step.

4. Application to longitudinal data: calcium absorbtion

The relationship between calcium absorbtion and calcium intake is of interest in addressing the problem of calcium deficiency. Heaney *et al.* (1989) have shown a complex inverse relation between the two variables, where age is among the variables that have a significant influence on calcium absorbtion efficiency. Other variables found to affect this relationship are body configuration measures such as body mass index or BSA that are found to be negatively correlated with calcium intake (Heaney, 2003). In order to uncover the age-dependent regression relation of absorbtion on intake adjusted for BSA, we analyze data from a longitudinal study on 188 subjects, conducted primarily to search for predictors of calcium absorbtion (Davis, 2002, p. 336). All the subjects were between 35 and 45 years of age at the beginning of the study (1967), where repeated measurements per subject were taken in 5-year intervals with the number of repeated measurements ranging from 1 to 4. Longitudinal measurements were taken on absorbtion and intake among others.

The coefficient functions from the underlying varying coefficient model

$$\text{absorbtion} = \beta_0(\text{age}) + \beta_1(\text{age})\text{intake} + e(\text{age}) \tag{4.1}$$

have been estimated adjusted for BSA through fitting a CAVCM to the longitudinal measurements of absorbtion and intake as proposed in Section 3. Three subjects have been removed before analysis, as their BSA values were outliers. A total of 20 age points, $36, 39, 41, 42, \ldots, 55, 56, 58, 61$, have been considered to fit the CAVCM, where the data observed at ages $(35, 36, 37)$, $(38, 39, 40)$, $(57, 58, 59)$, and $(60, 61, 62)$ have been collapsed to groups concentrated at 36, 39, 58, and 61, respectively. This grouping was carried out in order to have enough subjects observed at each age point to fit CAR. The number of subjects per age point ranged from 20 to 39, where all observations came from different subjects even after the collapsing of age points, since longitudinal measurements per subject were taken in 5-year intervals. The raw estimates $\hat{\beta}_0(\cdot)$ and $\hat{\beta}_1(\cdot)$ of $\beta_0(\cdot)$ and $\beta_1(\cdot)$ given in Figure 1 (top panels, dots) have been obtained using the weighted averaging described in Section 3.1, using a GCV bandwidth choice of 0.2. Overlaying the raw estimates are the proposed smooth estimates $\tilde{\beta}_0(\cdot)$, $\tilde{\beta}_1(\cdot)$ (solid) and Fan and Zhang's smooth estimates (dashed) of the two coefficient functions, both obtained through local polynomial smoothing with a bandwidth choice of 7. The 90% pointwise bootstrap confidence intervals (dotted) are also displayed in Figure 1.

The bootstrap confidence intervals in Figure 1 are based on the $(\alpha/2)B$th and $(1-\alpha/2)B$th percentiles of the bootstrap estimates, $\tilde{\beta}_0(t_j)^{(b)}$ and $\tilde{\beta}_1(t_j)^{(b)}$, obtained from $B = 1000$ bootstrap samples generated from the original data. The bootstrap estimates are obtained through the same two-step procedure used for CAVCM estimates. In the first step, the raw bootstrap estimates for each time point are computed through CAR based on the bootstrap sample. In the second step, smooth bootstrap estimates are obtained as linear

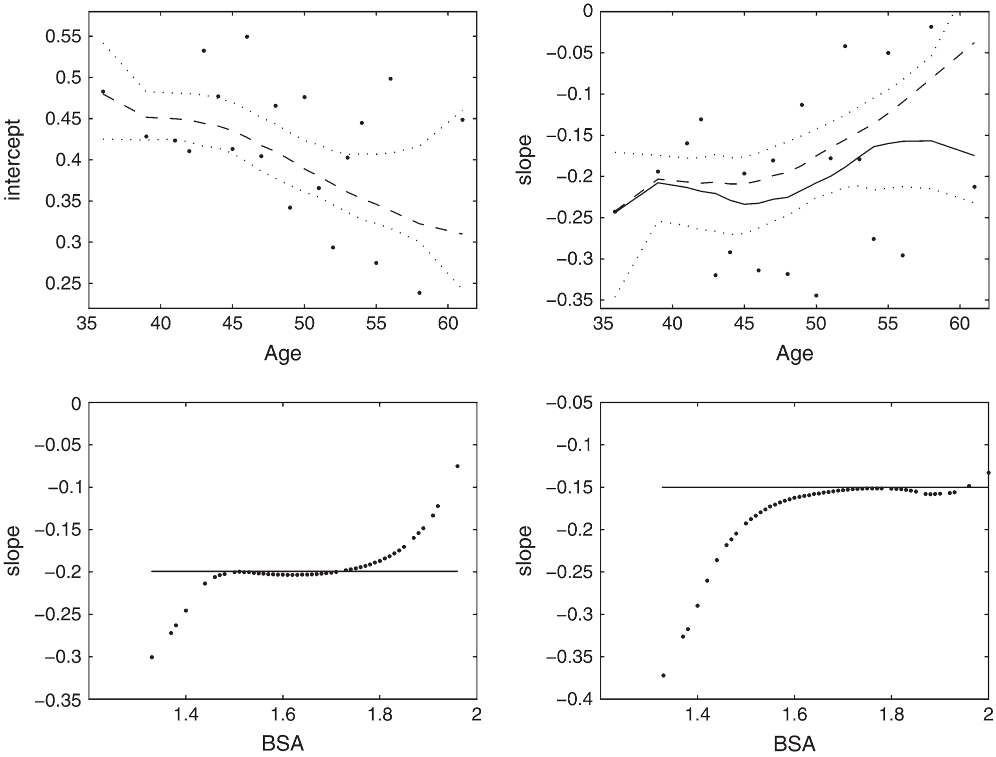

Fig. 1. Plots of the estimated smooth coefficient functions $\tilde{\beta}_0(\cdot)$ (top left panel) and $\tilde{\beta}_1(\cdot)$ (top right panel) for the underlying varying coefficient model, absorbtion $= \beta_0(\text{age}) + \beta_1(\text{age})\text{intake} + e(\text{age})$, adjusted for BSA. The proposed smooth coefficient function estimates (solid) and Fan and Zhang's unadjusted smooth estimates (dashed) are both obtained with local polynomial smoothing with a bandwidth choice of 7 for both functions. Overlaying the smooth estimates are the raw estimates $\hat{\beta}_0(\cdot)$, $\hat{\beta}_1(\cdot)$ (dots) along with 90% pointwise bootstrap confidence intervals (dotted). The bottom two plots correspond to slope functions (dots) from the varying coefficient models, absorbtion $= \gamma_0(\text{BSA}) + \gamma_1(\text{BSA})\text{intake} + \epsilon(\text{BSA})$, for subjects of age less than (bottom left panel) and greater than 45 (bottom right panel). Overlaying the two slope functions are the slope estimates (solid) from the linear regressions of absorbtion on intake, unadjusted for BSA, for subjects in the two age groups. Total number of repeated measurements is 515 collected on $n = 185$ subjects.

combinations of the raw bootstrap estimates $\hat{\beta}_r(t_j)^{(b)}$,

$$\tilde{\beta}_r(t_{j'})^{(b)} = \sum_{j=1}^{20} w_r(t_j, t_{j'})\hat{\beta}_r(t_j)^{(b)},$$

where the weights $w_r(t_j, t_{j'})$ are the ones used in obtaining the CAVCM smooth estimates, $\tilde{\beta}_r(t_{j'})$. The estimated nonparametric densities of the standardized 1000 bootstrap estimates of $\beta_0(t_j)$ (top panel) and $\beta_1(t_j)$ (bottom panel), evaluated at the 20 time points t_j, $j = 1, \ldots, 20$, are given in Figure 2, overlaying the standard normal density. The estimates at all the time points seem to be reasonably close to normal for both functions, enabling the use of the percentile bootstrap method. We also examine the estimated coverage levels of the proposed bootstrap confidence intervals in the simulation setting described in Section 5, Model I. One thousand data sets have been generated under the varying coefficient model given

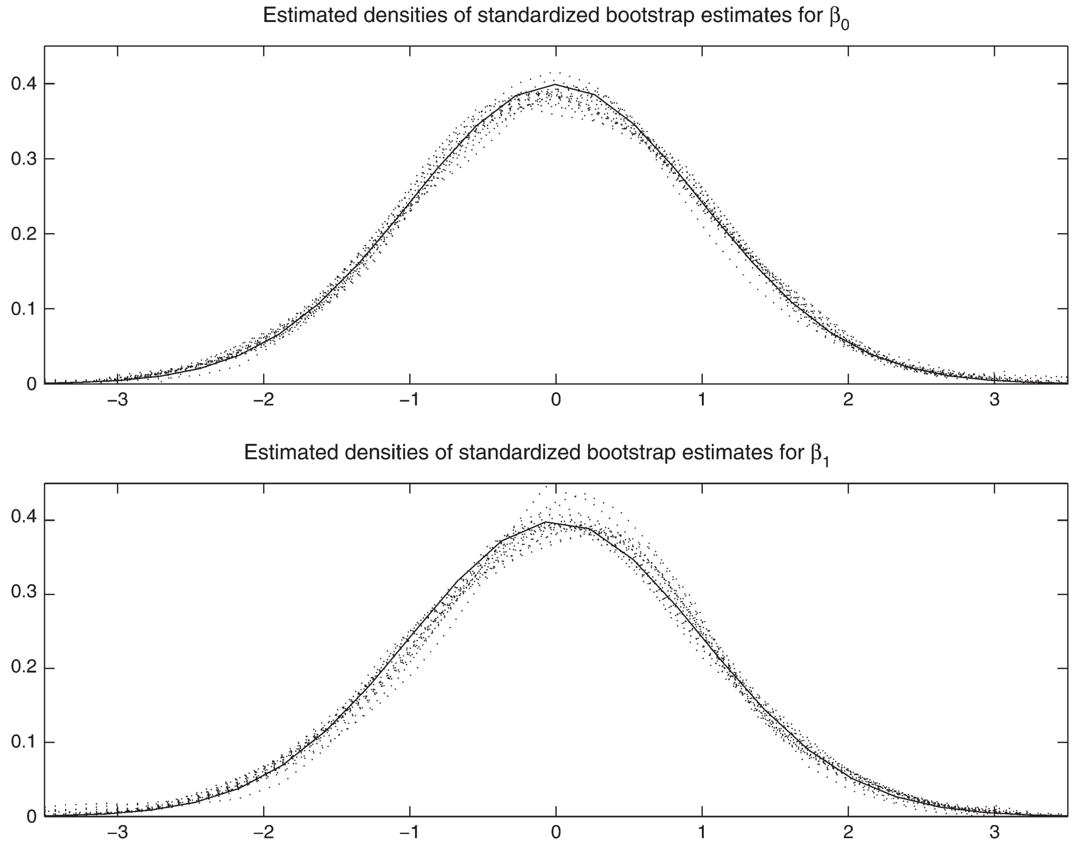

Fig. 2. Plot of the estimated nonparametric densities of 1000 standardized bootstrap estimates $\tilde{\beta}_0(t_j)^{(b)}$ (dotted, upper panel) and $\tilde{\beta}_1(t_j)^{(b)}$ (dotted, lower panel) used in forming 90% pointwise confidence intervals of the varying coefficient functions in the analysis of calcium absorbtion data. For both coefficients, 20 densities are presented corresponding to the 20 time points t_j that the bootstrap estimates are evaluated at. The standard normal density (solid) is also given in both panels. A fine binning procedure is followed by local least-squares fits with bandwidth choices of 0.5 to obtain the nonparametric densities.

in (5.1) below, where 1000 bootstrap samples have been generated from each data set. Figure 3 gives the estimated coverage values of the proposed confidence intervals for the three coefficient functions in (5.1) at each time point, corresponding to significance levels of 0.80 and 0.90. The estimated coverage values are roughly on target. The cross-sectional mean estimates and mean confidence intervals for the three coefficient functions, averaged over the 1000 Monte Carlo runs, are also shown in Figure 3, overlaying the true coefficient functions.

As is seen from Fan and Zhang's estimates, when unadjusted for BSA, the inverse effect of calcium intake on absorbtion seems to be declining with age. However, when adjusted for BSA with CAVCM, the inverse effect of calcium intake seems to be staying at about the same level as age increases. Even though the bootstrap confidence intervals from the CAVCM include the unadjusted estimates as well, the smooth estimates for BSA-adjusted and unadjusted models seem to differ, especially after the age of 45. In order to discover the precise nature of the effect of BSA on the relationship between calcium intake and absorbtion, we fitted varying coefficient models to the two groups of data observed at ages before and

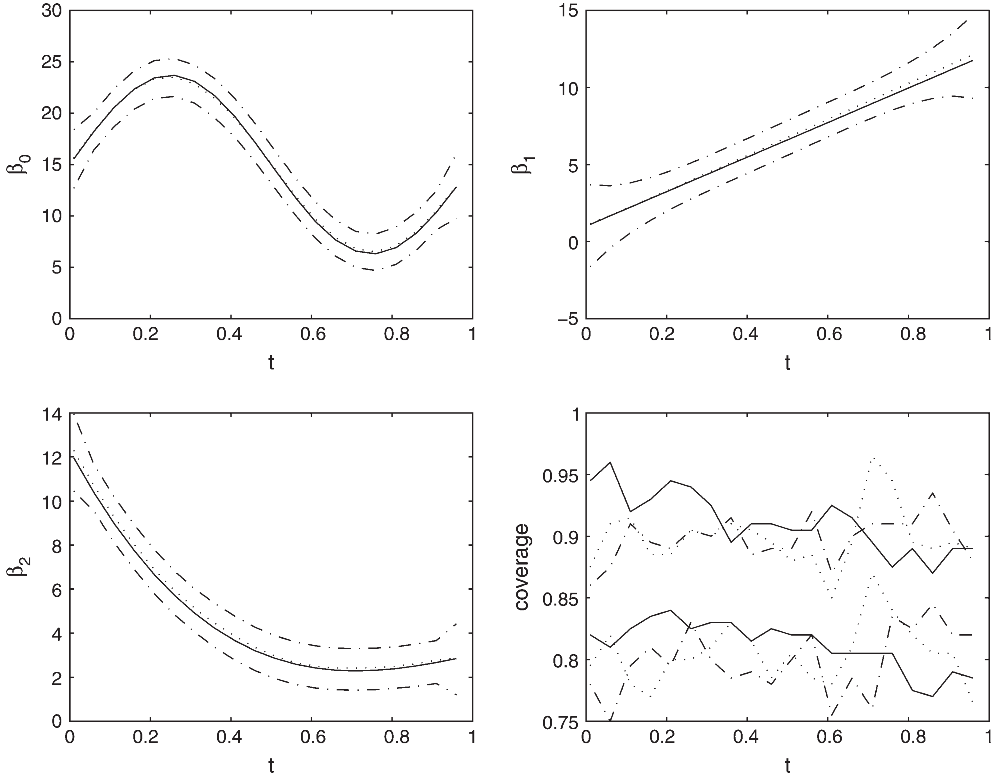

Fig. 3. Cross-sectional mean functions obtained from 1000 Monte Carlo runs under Model I (Section 5) of the proposed estimates (dotted). Also given are the mean bootstrap confidence intervals (dash-dotted) for the coefficient functions $\beta_0(t)$ (upper left panel), $\beta_1(t)$ (upper right panel), and $\beta_2(t)$ (lower left panel), overlaying the true coefficient functions (solid), in (5.1). Estimated coverage levels of the proposed pointwise confidence intervals for $\beta_0(t)$ (dotted), $\beta_1(t)$ (solid), and $\beta_2(t)$ (dash-dotted) at significance levels of 0.90 and 0.80 are plotted (lower right panel) against time t_j.

after 45. The two models fitted have the following form

$$\widetilde{\text{absorbtion}} = \gamma_0(\text{BSA}) + \gamma_1(\text{BSA})\widetilde{\text{intake}} + \epsilon(\text{BSA}), \tag{4.2}$$

where the longitudinal measurements of intake and absorbtion in the two age groups are collapsed together and treated as independent for the sake of the argument.

For both age groups, the inverse effect of intake on absorbtion declines in general as BSA increases as seen in Figure 1 (bottom panels). Nevertheless, one difference between the two age groups is that for those subjects over 45 with BSA greater than 1.6, the general inverse effect stays constant, whereas for those under 45, the effect keeps declining after BSA of 1.8. Also given in Figure 1 are the slope estimates from the linear regressions of $\widetilde{\text{absorbtion}}$ on $\widetilde{\text{intake}}$, unadjusted for BSA. Notice that a general CAR model targeting the underlying coefficients in the regression model

$$\text{absorbtion} = \eta_0 + \eta_1 \text{intake} + e$$

adjusted for BSA would get its estimates by averaging the smooth estimates of the coefficients in model (4.2). An important observation is that this average and therefore the estimates of such a CAR model would have lower values than their unadjusted linear regression counterparts for subjects over 45, but roughly at the same level for those less than 45. This also explains the fact that the BSA-adjusted varying slope estimates from model (4.1) are lower than the unadjusted estimates for ages over 45. Hence, adjusting for or stratifying by BSA, the stronger inverse effect of intake on absorbtion for subjects with lower BSA values becomes more influential in forming the BSA-adjusted regression coefficients. Thus, adjusted for BSA, this inverse effect does not decline, but stays at about the same level as the age of the subject considered increases.

5. Simulation study

In this section we compare the performance of CAVCM and Fan and Zhang's estimation procedure under three distortion models: multiplicative distortion, additive distortion, and no distortion.

We consider a simulation setup that mimics the calcium absorbtion data, by having 185 subjects with up to four repeated measurements. We consider the following underlying varying coefficient model

$$Y_i(t_j) = \beta_0(t_j) + \beta_1(t_j)X_{1i}(t_j) + \beta_2(t_j)X_{2i}(t_j) + e_i(t_j), \tag{5.1}$$

where $i = 1, \dots, 185$ and the time points t_j, $j = 1, \dots, 20$, are chosen to be equidistant between 0 and 1. The number of repeated measurements for each subject is chosen randomly to be $1, 2, 3$, or 4 with probabilities $0.025, 0.025, 0.05$, and 0.90, respectively. Thus, there are unequal number of observations taken on each subject where 80% or more of the data is missing. This yields 20–45 subjects observed at each time point on average, where the expected number of observations per time point is around 35. In (5.1), the three targeted coefficient functions are chosen to represent three different types of curves; $\beta_0(t) = 15 + 8.7\sin(2\pi t)$, $\beta_1(t) = 1 + 11.2t$, and $\beta_2(t) = 1 + 2t^2 + 11.3(1-t)^3$. The predictor $X_1(t)$ is a uniform random variable over the time-dependent interval $[t/4, 0.6 + t/4]$, and $X_2(t)$, when conditioning on $X_1(t)$, is a normal random variable with mean 1.5, and conditional variance $\text{var}\{X_2(t)|X_1(t)\} = \{1 + X_1(t)\}/\{8 + X_1(t)\}$. The error process $e(t)$ is sampled independently from the predictors from a stationary Gaussian process with mean zero and a decaying exponential covariance function $\delta(t_j, t_{j'}) = 5.27\exp(-0.5|t_j - t_{j'}|)$. The covariate U is generated from a uniform $[0, 1]$ distribution.

For the first considered multiplicative distortion model, the observed response and predictors are modeled as

$$\tilde{Y}_i(t_j) = \psi(U_i)Y_i(t_j) \quad \text{and} \quad \tilde{X}_{ri}(t_j) = \phi_r(U_i)X_{ri}(t_j), \quad r = 1, 2, \tag{Model I}$$

where the distorting functions considered are

$$\psi(U) = (U+3)^2/a, \quad \phi_1(U) = \exp(U)/b, \quad \phi_2(U) = (U+1.5)^2/c.$$

The constants $a = 12.33$, $b = 1.71$, and $c = 4.08$ are chosen such that the distorting functions satisfy the identifiability constraint of no average distortion in (2.3), namely $E\{\psi(U_i)\} = 1$ and $E\{\phi_r(U_i)\} = 1$.

For the second considered additive distortion model, the observed response and predictors are modeled as

$$\tilde{Y}_i(t_j) = \psi(U_i) + Y_i(t_j) \quad \text{and} \quad \tilde{X}_{ri}(t_j) = \phi_r(U_i) + X_{ri}(t_j), \quad r = 1, 2, \tag{Model II}$$

where the distorting functions are

$$\psi(U) = (U+3)^2 - a, \quad \phi_1(U) = \exp(U) - b, \quad \phi_2(U) = (U+1.5)^2 - c.$$

The constants a, b, and c have the same values as in Model I, but this time are subtracted from the specified functions of U, so that the distorting functions satisfy the identifiability constraint of no average distortion,

$E\{\tilde{Y}(t_j)\} = E\{Y(t_j)\}$ and $E\{\tilde{X}_r(t_j)\} = E\{X_r(t_j)\}$. The identifiability condition entails $E\{\psi(U_i)\} = 0$, and $E\{\phi_r(U_i)\} = 0$ in the additive distortion model.

As the last model, we consider no distortion in which case the observed and underlying response and predictors are the same:

$$\tilde{Y}_i(t_j) = Y_i(t_j) \quad \text{and} \quad \tilde{X}_{ri}(t_j) = X_{ri}(t_j), \quad r = 1, 2. \qquad \text{(Model III)}$$

For all the above models, the proposed CAVCM smooth estimates $\beta_0(\cdot)$, $\beta_1(\cdot)$, and $\beta_2(\cdot)$ have been obtained through local polynomial smoothing with cross-validation bandwidth choices of 0.12, 0.20, and 0.14, respectively. The GCV bandwidth choice was 0.5 in obtaining the CAVCM raw estimates in the first step. Fan and Zhang's smooth estimates have also been obtained for the three coefficient functions in the above three models, through local polynomial smoothing with cross-validation bandwidth choices of 0.14, 0.20, and 0.14, respectively. Smooth estimates of both methods from a single Monte Carlo run are displayed overlaying the true coefficient functions in Figures 4, 5, and 6 for Models I, II, and III, respectively. The bias of Fan and Zhang's raw estimates under the multiplicative distortion model have been given in (3.1). Thus, even though the smooth estimates of Fan and Zhang are on target for Model III of no distortion, they have considerable bias as seen in Figures 4 and 5 for the distortion Models I

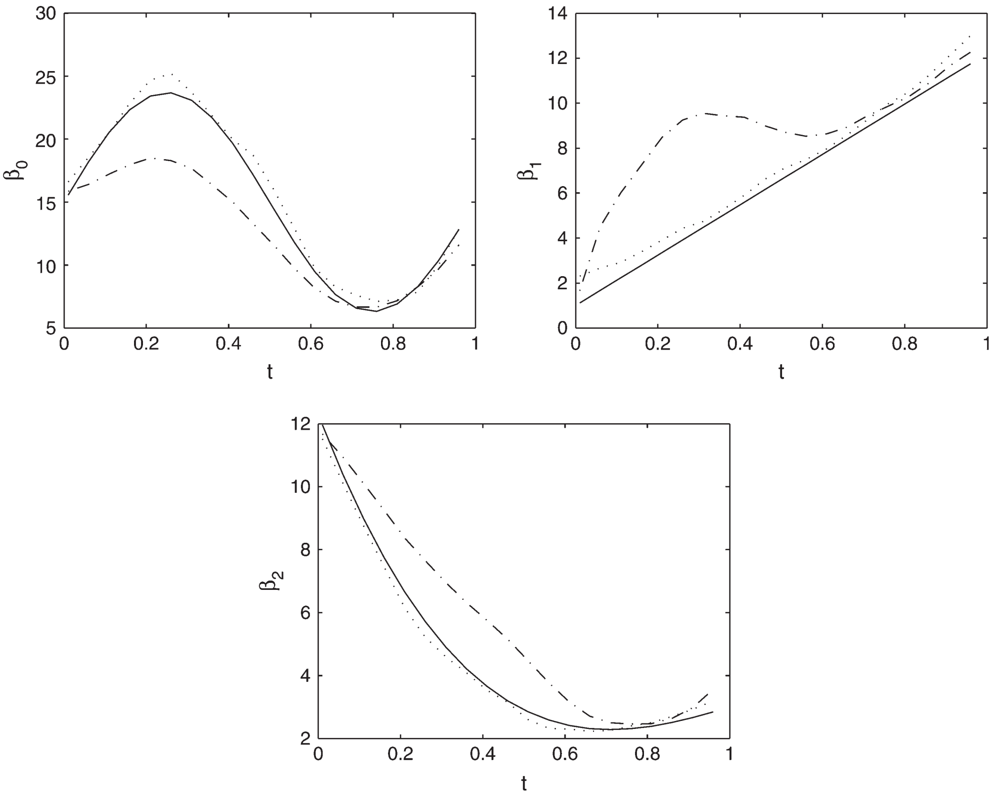

Fig. 4. Proposed CAVCM estimates (dotted) and Fan and Zhang's smooth estimates (dash-dotted) for the true coefficient functions $\beta_0(t)$ (solid, top left panel), $\beta_1(t)$ (solid, top right panel), and $\beta_2(t)$ (solid, bottom left panel) for the multiplicative distortion model, Model I. Both estimates are obtained through local polynomial smoothing in the second step with cross-validation bandwidth choices of 0.12, 0.20, and 0.14 (CAVCM) and 0.14, 0.20, and 0.14 (Fan and Zhang) for β_0, β_1, and β_2, respectively.

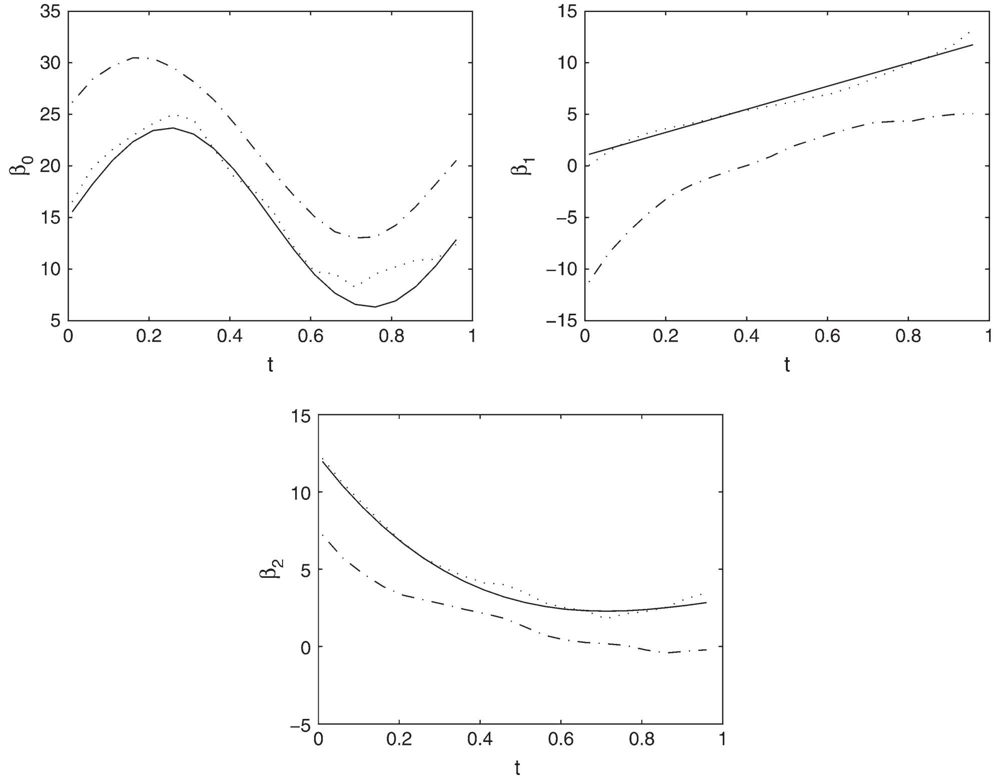

Fig. 5. Proposed CAVCM estimates (dotted) and Fan and Zhang's smooth estimates (dash-dotted) for the true coefficient functions $\beta_0(t)$ (solid, top left panel), $\beta_1(t)$ (solid, top right panel), and $\beta_2(t)$ (solid, bottom left panel) for the additive distortion model, Model II. Both estimates are obtained through local polynomial smoothing in the second step with cross-validation bandwidth choices of 0.12, 0.20, and 0.14 (CAVCM) and 0.14, 0.20, and 0.14 (Fan and Zhang) for β_0, β_1, and β_2, respectively.

and II, as expected. The CAVCM smooth estimates are on target for all three models. This shows that the CAVCM method is a very flexible adjustment method, where the form or even the existence of the distortion need not be known.

Another comparative measure of the performance of the fits obtained by the two methods is mean absolute deviation error (MADE), or weighted average-squared error (WASE), defined as

$$\text{MADE} = (3T)^{-1}\sum_{r=0}^{2}\sum_{j}^{T}\frac{|\beta_r(t_j)-\tilde{\beta}_r(t_j)|}{\text{range}(\beta_r)} \quad \text{and} \quad \text{WASE} = (3T)^{-1}\sum_{r=0}^{2}\sum_{j}^{T}\frac{\{\beta_r(t_j)-\tilde{\beta}_r(t_j)\}^2}{\text{range}^2(\beta_r)},$$

where $\tilde{\beta}_r(t_j)$ are the smooth estimates for both methods and range(β_r) is the range of the function $\beta_r(t)$. We also consider unweighted average-squared error (UASE) which is defined in the same way as WASE, but without any weights in the denominator. The box plots of the MADE, WASE, and UASE ratios of the proposed CAVCM method over Fan and Zhang's estimates from 1000 Monte Carlo runs are given in Figure 7, upper panel for Model I, middle panel for Model II, and lower panel for Model III. These plots indicate that the proposed estimates indeed handle the multiplicative and additive distortion models much better than Fan and Zhang's estimates. Even though CAVCM estimates target the true coefficient functions also under the case of no distortion, they are outperformed by Fan and Zhang's estimates in case

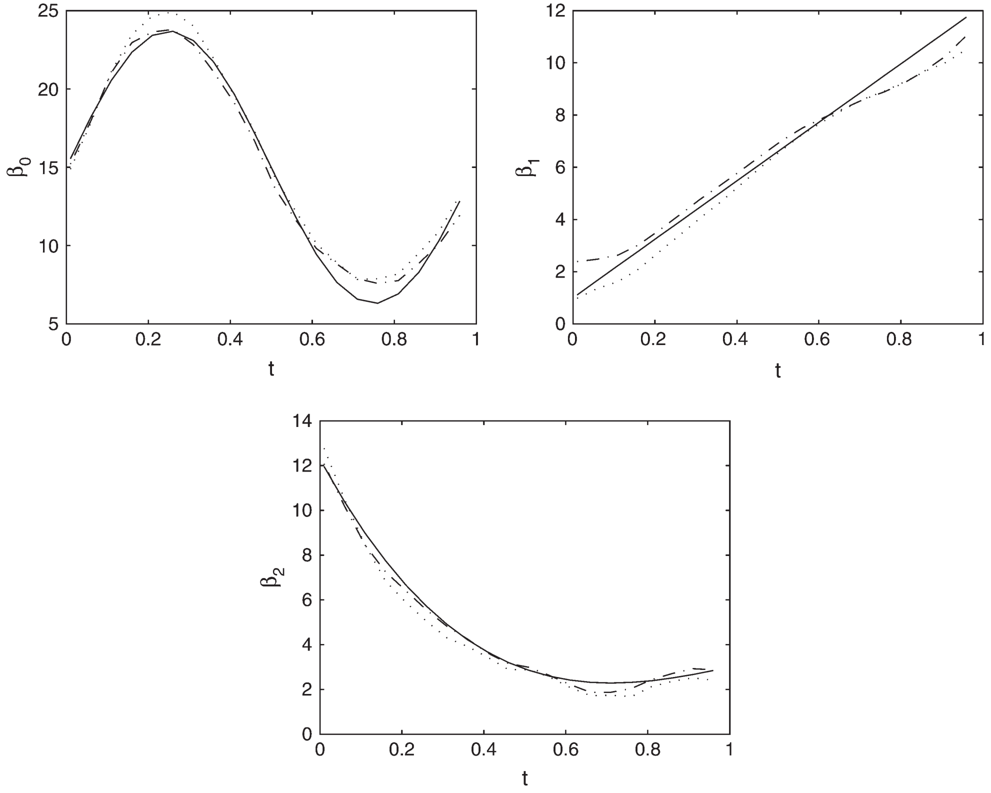

Fig. 6. Proposed CAVCM estimates (dotted) and Fan and Zhang's smooth estimates (dash-dotted) for the true coefficient functions $\beta_0(t)$ (solid, top left panel), $\beta_1(t)$ (solid, top right panel), and $\beta_2(t)$ (solid, bottom left panel) for the no distortion model, Model III. Both estimates are obtained through local polynomial smoothing in the second step with cross-validation bandwidth choices of 0.12, 0.20, and 0.14 (CAVCM) and 0.14, 0.20, and 0.14 (Fan and Zhang) for β_0, β_1, and β_2, respectively.

of Model III. This result is not surprising since the simple linear regression fits utilized in the first step of Fan and Zhang's algorithm are more efficient than CAR estimates in obtaining the raw estimates at each time point, for the specific case of no distortion.

6. Remarks

The proposed method of CAVCM provides a covariate-adjusted analysis for the regression relation between longitudinal variables. The two-step procedure is especially flexible in two different ways. It is flexible in handling different forms of distortion as illustrated through simulation studies. Not only the form but also the existence of the distortion need not be known. This nature of the algorithm is particularly appealing in case of a multiple varying coefficient model, where different predictors may be believed to have different relations with the covariate. Note, however, that there are some restrictions if the form of confounding on the response and predictors are not of the same form. More specifically, CAR yields consistent estimates under a model having an additive distortion in the response with predictors having either multiplicative or additive distortion. On the other hand, it will not give consistent estimates under a model with multiplicative distortion on the response with additive distorted predictors.

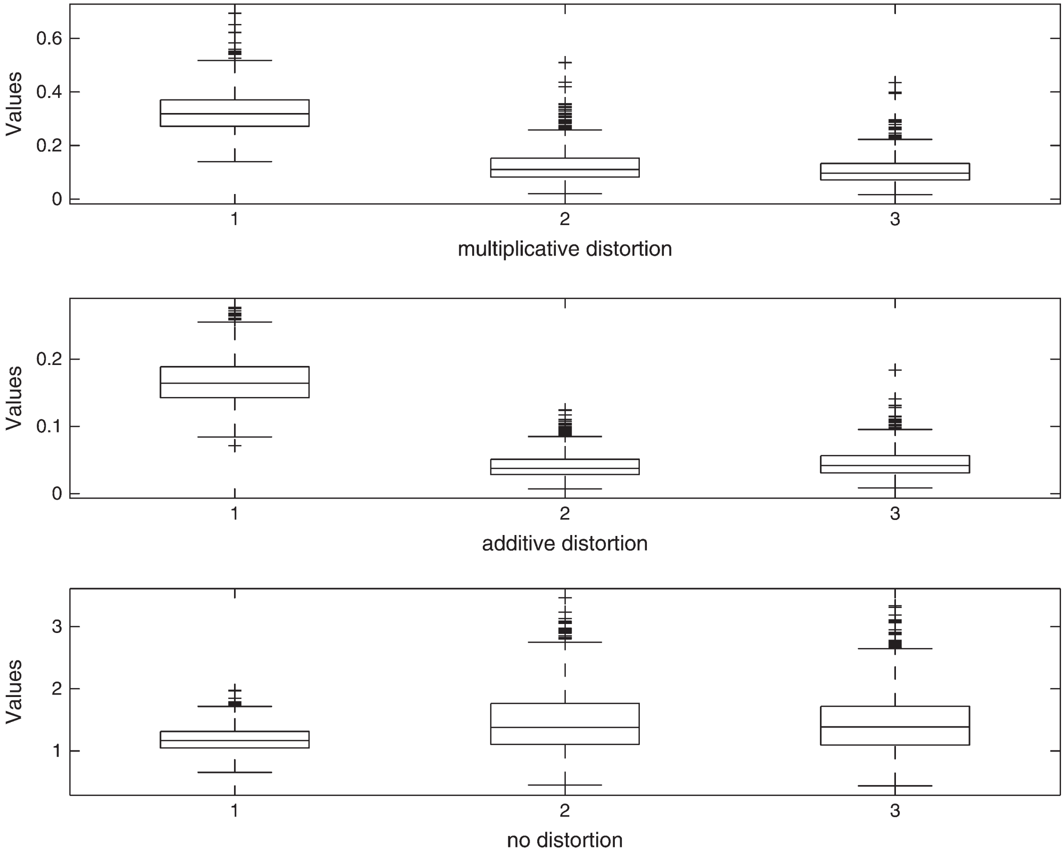

Fig. 7. Box plots for the ratios of error measures for proposed CAVCM estimates over Fan and Zhang's smooth estimates, for MADE (column 1), WASE (column 2), UASE (column 3), for Model I (upper panel), Model II (middle panel), and Model III (lower panel). Quotients smaller than 1 show that the proposed method is superior in the presence of distortion. The box plots are based on ratios obtained from 1000 Monte Carlo runs.

The second flexibility of the proposed method is its applicability to most longitudinal data structures. Assuming that the data is collected on the same set of time points for different subjects, the proposed methodology can handle a great percentage of missing values including those cases with only one measurement per some subjects. The only limitation comes from the use of the CAR algorithm in the first step, entailing the need of more than 20 subjects observed at most of the time points considered. However, the subjects observed at different time points do not need to be the same or of the same number. The values of the estimated coefficient functions at those time points where there are not enough subjects to fit CAR are imputed through the smoothing procedure in the second step.

The CAVCM algorithm can be applied to cases where longitudinal measurements are collected on the covariate as well. The methodology has been described for the case of cross-sectional covariate so far for simplicity of notation. For the case of a cross-sectional covariate, a fixed subject will have one reading on the covariate variable across time, whereas for the case of a longitudinal covariate, the readings will vary across time even for the same subject. However, in both cases, the observed measurements at a fixed time point come from different subjects which is a key observation enabling the application of CAR in the first step of the proposed estimation procedure. Thus, there is no change in the algorithm in case of a longitudinal covariate. The only difference between the two cases is one in model assumptions that the identifiability conditions on the distorting functions given in (2.3) need to hold at each time point for the case of the longitudinal covariate.

Fan and Zhang (2000) provide expressions for the asymptotic bias and variance of their smooth estimates obtained in the second step conditional on the predictor processes and the observed time points. These results give some insight to the optimal choice of bandwidth for the second smoothing step. The asymptotic bias and variance of the smooth CAVCM estimates can similarly be obtained once the bias and variance expressions for the CAR estimates in the first step of the estimation procedure are worked out. Another idea for future research would be to look for ways of incorporating the correlation structure of the longitudinal data into the proposed covariate-adjusted varying coefficient estimator to improve its efficiency.

Acknowledgments

We are extremely grateful to an anonymous referee, the associate editor, and the editors for many helpful remarks that improved the exposition of the paper. We also want to acknowledge Danh Nguyen for his valuable feedback during the preparation of the manuscript.

References

Chiang, C., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association* **96**, 605–617.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics* **31**, 377–403.

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer, p. 336.

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

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

Heaney, R. P. (2003). Normalizing calcium intake: projected population effects for body weight. *Journal of Nutrition* **133**, 268S–270S.

Heaney, R. P., Recker, R. R., Stegman, M. R. and Moy, A. J. (1989). Calcium absorption in women: relationships to calcium intake, estrogen status, age. *Journal of Bone and Mineral Research* **4**, 469–475.

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. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* **14**, 763–788.

Kaysen, G. A., Dubin, J. A., Müller, H. G., Mitch, W. E., Rosales, L. M., Levin, N. W. and the Hemo Study Group (2003). Relationship among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. *Kidney International* **61**, 2240–2249.

Şentürk, D. and Müller, H. G. (2005a). Covariate adjusted regression. *Biometrika* **92**, 59–74.

Şentürk, D. and Müller, H. G. (2005b). Inference for covariate adjusted regression via varying coefficient models. *Annals of Statistics* (in press).

Şentürk, D. and Nguyen, D. V. (2005). Estimation in covariate adjusted regression. *Computational Statistics and Data Analysis* (in press).

WAHBA, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In Krisnaiah, P. R. (ed.), *Applications of Statistics*. North Holland: Amsterdam, pp. 507–523.

WU, C. O. AND CHIANG, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica* **10**, 433–456.

WU, C. O. AND YU, K. F. (2002). Nonparametric varying coefficient models for the analysis of longitudinal data. *International Statistical Review* **70**, 373–393.

WU, C. O., YU, K. F. AND CHIANG, C. T. (2000). A two-step smoothing method for varying-coefficient models with repeated measurements. *Annals of the Institute of Statistical Mathematics* **25**, 519–543.

ZHANG, D. (2004). Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics* **60**, 8–15.

ZHANG, D., LIN, X., RAZ, J. AND SOWERS, M. F. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**, 710–719.

[*Received April 26, 2005; first revision September 2, 2005; second revision October 5, 2005; accepted for publication October 21, 2005*]