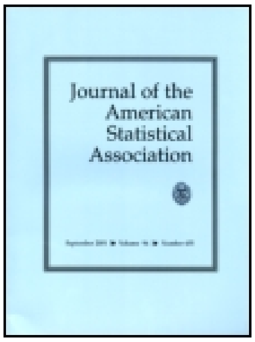

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data

Naisyin Wang, Raymond J Carroll & Xihong Lin

To cite this article: Naisyin Wang, Raymond J Carroll & Xihong Lin (2005) Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data, Journal of the American Statistical Association, 100:469, 147-157, DOI: 10.1198/016214504000000629

To link to this article: https://doi.org/10.1198/016214504000000629

Published online: 31 Dec 2011.

Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data

Naisyin WANG, Raymond J. CARROLL, and Xihong LIN

We consider marginal generalized semiparametric partially linear models for clustered data. Lin and Carroll derived the semiparametric efficient score function for this problem in the multivariate Gaussian case, but they were unable to construct a semiparametric efficient estimator that actually achieved the semiparametric information bound. Here we propose such an estimator and generalize the work to marginal generalized partially linear models. We investigate asymptotic relative efficiencies of the estimators that ignore the within-cluster correlation structure either in nonparametric curve estimation or throughout. We evaluate the finite-sample performance of these estimators through simulations and illustrate it using a longitudinal CD4 cell count dataset. Both theoretical and numerical results indicate that properly taking into account the within-subject correlation among the responses can substantially improve efficiency.

KEY WORDS: Clustered data; Generalized estimating equations; Kernel method; Longitudinal data; Marginal models; Nonparametric regression; Partially linear model; Profile method; Sandwich estimator; Semiparametric-efficient score; Semiparametric information bound; Time-dependent covariate.

1. INTRODUCTION

We consider estimation in marginal semiparametric generalized linear models for clustered data using estimating equations. These models are becoming an increasingly popular topic of research (see Zeger and Diggle 1994; Wild and Yee 1996; Pepe and Couper 1997; Hoover, Rice, Wu, and Yang 1998; Lin and Carroll 2001a,b; Lin and Ying 2001 for recent examples).

These marginal models, through general links, have predictor effects that are *partially linear*; they consist of a linear function of one set of predictors (e.g., exposure variables) with a parameter vector $\boldsymbol{\beta}$ and a completely nonparametric function of a scalar covariate (e.g., time). For uncorrelated data, Severini and Staniswalis (1994) showed how to construct a semiparametric efficient estimator of $\boldsymbol{\beta}$ using a profile-kernel method. Lin and Carroll (2001a), hereafter referred to as LC, showed that for clustered data, the conventional profile-kernel method does not yield an efficient estimator of $\boldsymbol{\beta}$ when the parametric covariate is dependent of the nonparametric covariate. In fact, such an estimated $\boldsymbol{\beta}$ could be $\sqrt{n}$-inconsistent unless either a *working independence* (WI) assumption or an undersmoothing step is adopted, here WI means that one ignores the correlation structure entirely. LC derived the semiparametric efficient score of $\boldsymbol{\beta}$ in the multivariate Gaussian case, and noted that it was a solution to a complicated Fredholm integral equation. They were unable to construct an estimator that was semiparametric efficient, however.

The purpose of this article is to propose a semiparametric efficient estimator of $\boldsymbol{\beta}$ in such marginal partially linear models, allowing the parametric and nonparametric covariates to be dependent on one another. When the nonparametric covariate is time, this implies that the parametric covariates could be time-dependent. We show that the estimator can effectively account for within-cluster correlation. It is semiparametric efficient in the Gaussian case and is more efficient than the WI estimator in non-Gaussian cases.

The outline of the article is as follows. In Section 2 we describe the model and state the major assumptions. Of particular note is that we are *not* working in the context of time series data; our asymptotics assume that the number of clusters/individuals becomes large while the number of observations per cluster/individual remains bounded. In Section 3 we describe the proposed estimator, and in Section 4 we state the main theoretical results. We present numerical studies, including an investigation of the asymptotic relative efficiencies (AREs) of two previously proposed estimators and a small simulation study, in Sections 5 and 6. In Section 7 we analyze a longitudinal dataset of CD4 cell counts of human immunodeficiency virus (HIV) seroconverters. Finally, we give some concluding remarks in Section 8.

2. THE MODEL

Suppose that the data consist of n clusters with the ith $(i = 1, \dots, n)$ cluster having m_i observations. Let Y_{ij} and $(\mathbf{X}_{ij}, T_{ij})$ be the response variable and covariates for the jth $(j = 1, \dots, m_i)$ observation in the ith cluster. Here $\mathbf{X}_{ij}$ is a $p \times 1$ vector and T_{ij} is a scalar that varies within each cluster. Let $\underline{\mathbf{Y}}_i = (Y_{i1}, \dots, Y_{im_i})^t$, and define $\underline{\mathbf{X}}_i$ and $\underline{\mathbf{T}}_i$ similarly. Our basic assumption is that the underlying distribution of the response and covariate processes are the same for all subjects, that $(\underline{\mathbf{Y}}_i, \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$ are observations of the ith randomly selected subject within, say, a fixed range of $\mathbf{T}$ such that m_i are bounded, and that

$$E(Y_{ij} | \mathbf{X}_{ij}, T_{ij}, \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i) = E(Y_{ij} | \mathbf{X}_{ij}, T_{ij}) = \mu_{ij} \tag{1}$$

(see Pepe and Couper 1997 for a discussion of this assumption). The marginal mean μ_{ij} depends on $\mathbf{X}_{ij}$ and T_{ij} through a known monotonic and differentiable link function $g(\cdot)$,

$$g(\mu_{ij}) = \mathbf{X}_{ij}^t \boldsymbol{\beta} + \theta(T_{ij}), \tag{2}$$

Naisyin Wang is Professor of Statistics and Toxicology (E-mail: *nwang@stat.tamu.edu*) and Raymond J. Carroll is Distinguished Professor of Statistics, Epidemiology and Biostatistics, Nutrition and Toxicology (E-mail: *carroll@stat.tamu.edu*), Texas, A&M University, College Station, TX 77843. Xihong Lin is Professor of Biostatistics, Department of Biostatistics, School of Public Health, Ann Arbor, MI 48109 (E-mail: *xlin@umich.edu*). Wang's research was supported by National Cancer Institute grant CA74552 and the Texas Advanced Research Program. Carroll's research was supported by National Cancer Institute grant CA57030 and by the Texas A&M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences grant P30–ES09106. Lin's research was supported by National Cancer Institute grant CA76404. The authors thank Vincent Carey for sharing his S–PLUS code, "yags," for parametric GEE. They also thank the editor, an associate editor, and three referees for their detailed and helpful comments.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000000629

where $\boldsymbol{\beta}$ is a $p \times 1$ vector and $\theta(\cdot)$ is an unknown smooth function. We thus model the effect of $\mathbf{X}$ ($p \times 1$) parametrically and the effect of $\mathbf{T}$ nonparametrically. In matrix notation, denoting $\underline{\boldsymbol{\mu}}_i = (\mu_{i1}, \ldots, \mu_{im_i})^t$ and $\mathbf{g}(\underline{\boldsymbol{\mu}}_i) = \{g(\mu_{i1}), \ldots, g(\mu_{im_i})\}^t$, we have $\mathbf{g}(\underline{\boldsymbol{\mu}}_i) = \underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\boldsymbol{\theta}}(\underline{\mathbf{T}}_i)$.

As indicated in Section 1, we allow $\mathbf{X}$ and $\mathbf{T}$ to be dependent. This in general is the case for longitudinal/clustered data. A referee has pointed out the following problem in which the original $\mathbf{X}$ and $\mathbf{T}$ are independent but yet can be reparameterized and solved using the proposed method. Specifically, suppose that one of the $\boldsymbol{\beta}$'s, say β_1, in (2) is known to be a linear function of $\mathbf{T}$ through $\beta_1(T_{ij}) = \beta_{10} + \beta_{11}T_{ij}$. It is easily seen that $\beta_1(T_{ij})X_{1ij} = \beta_{10}X_{1ij} + \beta_{11}X^*_{1ij}$, where $X^*_{1ij} = X_{1ij}T_{ij}$. Thus model (2) still holds with the added covariate $\mathbf{X}^*_1$, but the $\mathbf{X}^*_1$ is $\mathbf{T}$ dependent even if the original $\mathbf{X}_1$ is not. This reparameterization allows us to use the proposed method without modification to obtain inference on β_{10} and β_{11}.

Model (2) differs from a standard marginal generalized estimating equation (GEE) model (Liang and Zeger 1986) mainly through the nonparametric component $\boldsymbol{\theta}(\cdot)$. It is motivated by the fact that the effect of the covariate $\mathbf{T}$ (e.g., time) may be complicated and might be better modeled nonparametrically. Applications of marginal models are common (see, e.g., Diggle, Liang, and Zeger 1994; Heagerty and Zeger 2000).

Let $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$ and $\mathbf{V}_i = \mathbf{V}_i(\underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$ be the true and assumed "working" covariances of $\underline{\mathbf{Y}}_i$, where $\boldsymbol{\Sigma}_i = \mathrm{var}(\underline{\mathbf{Y}}_i | \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$ and $\mathbf{V}_i = \mathbf{S}_i^{1/2}\mathbf{R}_i\mathbf{S}_i^{1/2}$, $\mathbf{S}_i$ denotes a diagonal matrix that contains the marginal variances of the Y_{ij}'s, and $\mathbf{R}_i$ is an invertible working correlation matrix. Throughout, we assume that $\mathbf{V}_i$ can depend on a nuisance finite-dimensional parameter vector $\boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is distinct from $\boldsymbol{\beta}$.

3. THE ESTIMATION PROCEDURE

Our estimation procedure is based on profile-kernel estimating equations, where $\theta(t)$ is estimated using a kernel GEE estimator accounting for correlations proposed by Wang (2003) and $\boldsymbol{\beta}$ is estimated using a profile-type estimating equation. The proposed method differs from those proposed by Severini and Staniswalis (1994) and LC (2001a) only in the way in which $\widehat{\theta}(t, \boldsymbol{\beta})$, the estimated $\theta(t)$ for a given $\boldsymbol{\beta}$, is constructed. This is motivated by a fact shown by LC that to reach the semiparametric information bound, the within-cluster correlation must be properly accounted for in both the parametric and nonparametric estimation procedures. The conventional kernel GEE estimator of $\theta(t)$ (Lin and Carroll 2001a) fails to do so, whereas the new iterative kernel GEE estimator (Wang 2003) effectively accounts for correlation.

When the link function $\mathbf{g}$ is linear, both $\widehat{\boldsymbol{\beta}}$ and $\widehat{\theta}(t, \boldsymbol{\beta})$ are linear estimators. Closed-form expressions of the proposed estimators exist (Lin, Wang, Welsh, and Carroll 2004). To better appreciate the nature of the estimators, we present them in an iterative format. For any given $\boldsymbol{\beta}$, start with an estimator $\widetilde{\theta}(t, \boldsymbol{\beta})$ of $\theta(t)$ and an initial estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ satisfying $n^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$. Such initial estimators can be easily obtained, for example, using the WI estimator that ignores the correlation structure entirely.

We concentrate here on a local linear estimator of $\theta(t)$ proposed by Wang (2003). Let $K_h(s) = h^{-1}K(s/h)$, where K is a mean-0 symmetric density function. Define $\mathbf{G}_{ij}(t)$ to be an $m_i \times 2$ matrix with the ℓth column $\mathbf{e}_j \times \{(t - T_{ij})/h\}^{\ell-1}$ ($\ell = 1, 2$), where $\mathbf{e}_j$ is an $m_i \times 1$ vector of 0 except with the jth entry being 1. Our method starts with the WI estimator and iterates between the following steps I and II until convergence. The working covariance matrix $\mathbf{V}_i$ depends on a parameter vector $\boldsymbol{\tau}$, which is assumed to be distinct from $\boldsymbol{\beta}$ and can be estimated via the method of moments using quadratic functions of the responses.

- *Step I.* Let $\widetilde{\theta}(\cdot)$ be the current estimator of $\theta(\cdot)$. Given $\boldsymbol{\beta}$, let $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(t, \boldsymbol{\beta}) = \{\widehat{\alpha}_0(t, \boldsymbol{\beta}), \widehat{\alpha}_1(t, \boldsymbol{\beta})\}^t$ be the solution to the kernel equation

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t - T_{ij})\mu_{ij}^{(1)}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})\mathbf{G}_{ij}^t(t)\mathbf{V}_i^{-1} \times \bigl[\underline{\mathbf{Y}}_i - \boldsymbol{\mu}^*\bigl\{t, \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\theta}}(\underline{\mathbf{T}}_i; \boldsymbol{\beta})\bigr\}\bigr] = 0, \quad (3)$$

where the ℓth element of $\boldsymbol{\mu}^*\{t, \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\theta}}(\underline{\mathbf{T}}_i, \boldsymbol{\beta})\}$ is

$$\mu\bigl[\mathbf{X}_{i\ell}^t\boldsymbol{\beta} + I(\ell = j)\{\widehat{\alpha}_0 + \widehat{\alpha}_1(t - T_{ij})/h\} + I(\ell \neq j)\widetilde{\theta}(T_{i\ell}, \boldsymbol{\beta})\bigr]$$

and $\mu_{ij}^{(1)}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})$ is the first derivative of the function $\boldsymbol{\mu}(\cdot) = \mathbf{g}^{-1}(\cdot)$ evaluated at $\mathbf{X}_{ij}^t\boldsymbol{\beta} + \widehat{\alpha}_0 + \widehat{\alpha}_1(t - T_{ij})/h$. The updated estimator of $\theta(t)$ is $\widehat{\theta}(t, \boldsymbol{\beta}) = \widehat{\alpha}_0(t, \boldsymbol{\beta})$.
- *Step II.* Find $\widehat{\boldsymbol{\beta}}$ to solve the profile-type estimating equation

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}\{\underline{\mathbf{X}}_i\boldsymbol{\beta} + \widehat{\boldsymbol{\theta}}(\underline{\mathbf{T}}_i, \boldsymbol{\beta})\}^t}{\partial \boldsymbol{\beta}}\mathbf{V}_i^{-1} \times \bigl[\underline{\mathbf{Y}}_i - \boldsymbol{\mu}\{\underline{\mathbf{X}}_i\boldsymbol{\beta} + \widehat{\boldsymbol{\theta}}(\underline{\mathbf{T}}_i, \boldsymbol{\beta})\}\bigr] = 0. \quad (4)$$

Denote by $\{\widehat{\theta}(t), \widehat{\boldsymbol{\beta}}\}$ the estimates at convergence with $\widehat{\theta}(t) = \widehat{\theta}(t, \widehat{\boldsymbol{\beta}})$. As mentioned earlier, our algorithm differs from those proposed in step I by replacing the original kernel GEE estimator by one that uses the correlations, whereas step II is the same. This modification turns out to be the key to constructing a semiparametric-efficient estimator of $\boldsymbol{\beta}$.

As shown in Section 4 and later illustrated in the simulation, the proposed $\widehat{\boldsymbol{\beta}}$ is insensitive to the choice of bandwidth. For example, any bandwidth h of order $O_p(n^{-1/5})$, as in the univariate case, can be used. The same asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ will be obtained. Results for higher-order local polynomials can be easily obtained by following the derivations given by Ruppert and Wand (1994).

4. THEORETICAL RESULTS

We emphasize that we assume in our asymptotic theory that the number of clusters, n, goes to ∞ while the cluster sizes, m_i, remains bounded. If m_i also tends to ∞, then the problem is different, as pointed out by LC. We assume that the regularity conditions specified in the Appendix hold. Denote by $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0(\mathbf{t})$ the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}(\mathbf{t})$. Let $d^{(r)}(\cdot)$ be the rth derivative of any function $d(\cdot)$, let $\alpha_\ell = h^\ell\theta^{(\ell)}(\mathbf{t})/\ell!$, let $v_i^{j\ell}$ be the (j, ℓ)th element of a matrix $\mathbf{V}_i^{-1}$, and let $f_j(\mathbf{t})$ be the marginal density of the T_{ij}.

Our results concerning $\widehat{\boldsymbol{\theta}}(\mathbf{t})$ are simple and coincide exactly with those of Wang (2003), who showed that $\widehat{\boldsymbol{\theta}}(\mathbf{t})$ can effectively account for the within-cluster correlation and is asymptotically more efficient than the WI estimator. The main focus of this article is on the properties of $\widehat{\boldsymbol{\beta}}$.

LC gave the semiparametric efficient score for a multivariate normal partially linear model with a known $\boldsymbol{\Sigma}$ for all i. We use the same setup but allow the conditional mean to be as given in (2) and the conditional variance to be $\boldsymbol{\Sigma}_i(\underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$ with known $\boldsymbol{\tau}$. The results remain the same with estimated $\boldsymbol{\tau}$. We discuss this point in Remark 5.

A referee has also suggested that we provide a link of our work to the least-favorable direction principle of Bickel, Klaassen, Ritov, and Wellner (1993). This link, described in Appendix Section A.1, provides an alternative derivation of the efficient score than that given by LC, and also allows extension to the unequal m_i case.

Under the assumed multivariate normal structure, we show that $\widehat{\boldsymbol{\beta}}$ is semiparametric efficient (Proposition 1 and Corollary 1). We then extend the results to model (2) for general outcomes with a working covariance $\mathbf{V}$ and without distributional assumptions (Proposition 2).

In what follows in this section, for simplicity of notation, we consider the case where $m_i \equiv m$ and $(\underline{\mathbf{Y}}_i, \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$ are iid. Let $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\underline{\mathbf{X}}, \underline{\mathbf{T}})$ be a diagonal matrix, with the diagonal element being the first derivative of $\boldsymbol{\mu}$. We show in Section A.1 that the semiparametric information bound for $\boldsymbol{\beta}$ under the multivariate normal assumption is

$$E\big[\{\underline{\mathbf{X}} - \boldsymbol{\varphi}_{\text{eff}}(\underline{\mathbf{T}})\}^t \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\{\underline{\mathbf{X}} - \boldsymbol{\varphi}_{\text{eff}}(\underline{\mathbf{T}})\}\big],$$

where $\boldsymbol{\varphi}_{\text{eff}}(\underline{\mathbf{T}})$ is an $m \times p$ matrix whose jth row is $\boldsymbol{\varphi}_{\text{eff}}(\mathbf{T}_j)$, and $\boldsymbol{\varphi}_{\text{eff}}(\mathbf{T}_j) = \{\varphi_{\text{eff},1}(\mathbf{T}_j), \ldots, \varphi_{\text{eff},p}(\mathbf{T}_j)\}^T$. The function $\boldsymbol{\varphi}_{\text{eff}}(\mathbf{t})$ solves

$$\sum_{j=1}^{m}\sum_{\ell=1}^{m} E\big[\Delta_{jj}\sigma^{j\ell}\Delta_{\ell\ell}\{\mathbf{X}_\ell - \boldsymbol{\varphi}_{\text{eff}}(\mathbf{T}_\ell)\}\big|\mathbf{T}_j = \mathbf{t}\big]f_j(\mathbf{t}) = 0, \qquad (5)$$

where $\sigma^{j\ell}$ is the (j,ℓ)th element of $\boldsymbol{\Sigma}^{-1}$ and Δ_{jj} is the (j,j)th element of $\boldsymbol{\Delta}$. Equation (5) corresponds to a Fredholm integral equation of the second kind (Kress 1989, chap. 1), namely

$$\boldsymbol{\varphi}_{\text{eff}}(\mathbf{t}) - \left\{\mathbf{q}(\mathbf{t}) - \int H(\mathbf{t},\mathbf{s})\boldsymbol{\varphi}_{\text{eff}}(\mathbf{s})\,d\mathbf{s}\right\} = 0, \qquad (6)$$

where, letting $f_{\ell j}(\cdot,\cdot)$ denote the joint density function of $(\mathbf{T}_\ell, \mathbf{T}_j)$, $H(\mathbf{t},\mathbf{s})$ and $\mathbf{q}(\mathbf{t})$ are defined as

$$H(\mathbf{t},\mathbf{s}) = \frac{\sum_j \sum_{\ell \neq j} E(\Delta_{jj}\sigma^{j\ell}\Delta_{\ell\ell}|\mathbf{T}_\ell = \mathbf{s}, \mathbf{T}_j = \mathbf{t}) f_{\ell j}(\mathbf{s},\mathbf{t})}{\sum_{j=1}^{m} E(\sigma^{jj}\Delta_{jj}^2|\mathbf{T}_j = \mathbf{t}) f_j(\mathbf{t})} \qquad (7)$$

and

$$\mathbf{q}(\mathbf{t}) = \frac{\sum_{j=1}^{m}\sum_{\ell=1}^{m} E(\Delta_{jj}\sigma^{j\ell}\Delta_{\ell\ell}\mathbf{X}_\ell|\mathbf{T}_j = \mathbf{t}) f_j(\mathbf{t})}{\sum_{j=1}^{m} E(\sigma^{jj}\Delta_{jj}^2|\mathbf{T}_j = \mathbf{t}) f_j(\mathbf{t})}. \qquad (8)$$

Equivalent derivations in the linear case were given by LC.

We next study the asymptotic distribution of our estimator $\widehat{\boldsymbol{\beta}}$ and show that it reaches the foregoing semiparametric information bound. Define $\widehat{\boldsymbol{\varphi}}(\mathbf{t}, \boldsymbol{\beta}) = -\partial\widehat{\boldsymbol{\theta}}(\mathbf{t}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}^t$. We first show that $\widehat{\boldsymbol{\varphi}}(\mathbf{t}, \boldsymbol{\beta})$ converges to $\boldsymbol{\varphi}_{\text{eff}}(\mathbf{t})$, which is a crucial theoretical result of this article and the key to investigation of the asymptotic properties of $\widehat{\boldsymbol{\beta}}$. We also use it later to justify why the proposed estimator does not require undersmoothing of $\boldsymbol{\theta}(\mathbf{t})$ to obtain a $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\beta}}$.

Proposition 1. Let $\widehat{\boldsymbol{\varphi}}$ be the partial derivative of the final estimator of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\beta}$ as defined earlier, and let $\boldsymbol{\varphi}$ be its limit as $n \to \infty$. Then $\boldsymbol{\varphi}$ satisfies (6); that is, $\boldsymbol{\varphi}(\mathbf{t}) = \boldsymbol{\varphi}_{\text{eff}}(\mathbf{t})$.

Corollary 1. Under the assumed multivariate normal structure, with $h \to 0, n \to \infty$, at the rate that $nh^8 \to 0$, and $nh/\log(1/h) \to \infty$, we have

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ \to \text{normal}\big(\mathbf{0}, E\big[\{\underline{\mathbf{X}} - \boldsymbol{\varphi}_{\text{eff}}(\underline{\mathbf{T}})\}^t\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\{\underline{\mathbf{X}} - \boldsymbol{\varphi}_{\text{eff}}(\underline{\mathbf{T}})\}\big]\big),$$

in distribution; that is, $\widehat{\boldsymbol{\beta}}$ reaches the semiparametric information bound.

The proofs of Proposition 1 and Corollary 1 are given in Sections A.3 and A.4. We now consider the properties of $\widehat{\boldsymbol{\beta}}$ in model (2) for general outcomes assuming a working covariance matrix $\mathbf{V}$ and without the normality assumption. The asymptotic properties of $\widehat{\boldsymbol{\beta}}$ are presented in Proposition 2. Recall that $\boldsymbol{\Delta}_i = \boldsymbol{\Delta}(\underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i) = \text{diag}\{\mu_{ij}^{(1)}\}$, and define $\widetilde{\underline{\mathbf{X}}}_i = \underline{\mathbf{X}}_i - \boldsymbol{\varphi}(\underline{\mathbf{T}}_i, \boldsymbol{\beta}_0)$.

Proposition 2. Let $\widetilde{\mathbf{A}}(\mathbf{V}) = E(\widetilde{\underline{\mathbf{X}}}^t\boldsymbol{\Delta}\mathbf{V}^{-1}\boldsymbol{\Delta}\widetilde{\underline{\mathbf{X}}})$ and $\widetilde{\mathbf{B}}(\mathbf{V}, \boldsymbol{\Sigma}) = E(\widetilde{\underline{\mathbf{X}}}^t\boldsymbol{\Delta}\mathbf{V}^{-1}\boldsymbol{\Sigma}\mathbf{V}^{-1}\boldsymbol{\Delta}\widetilde{\underline{\mathbf{X}}})$. With $h \to 0, n \to \infty$, at the rate that $nh^8 \to 0$ and $nh/\log(1/h) \to \infty$, we have

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to \text{normal}\{\mathbf{0}, \boldsymbol{\Omega}(\mathbf{V}, \boldsymbol{\Sigma})\},$$

where

$$\boldsymbol{\Omega}(\mathbf{V}, \boldsymbol{\Sigma}) = \{\widetilde{\mathbf{A}}(\mathbf{V})\}^{-1}\widetilde{\mathbf{B}}(\mathbf{V}, \boldsymbol{\Sigma})\{\widetilde{\mathbf{A}}(\mathbf{V})\}^{-1}. \qquad (9)$$

The asymptotic covariance $\boldsymbol{\Omega}(\mathbf{V}, \boldsymbol{\Sigma})$ is minimized by $\mathbf{V} = \boldsymbol{\Sigma}$, and in this case equals $\widetilde{\mathbf{A}}^{-1}(\boldsymbol{\Sigma})$.

The proof of (9) is sketched in Appendix A.4. Several remarks about our theoretical results are now in order.

Remark 1. LC showed that when the conventional profile-kernel method is used, except in the special case where WI is assumed, a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ can be obtained only if one artificially undersmooths the nonparametric estimator to eliminate an unwanted bias term. That is, either the bandwidth must be chosen so that $nh^4 \to 0$, or a nonparametric regression algorithm with smaller bias than standard kernel regression must be used (e.g., the twicing method). Even when $\boldsymbol{\theta}(\mathbf{t})$ is undersmoothed, the conventional profile-kernel estimator assuming the true correlation is still inefficient. However, using our method, not only is such undersmoothing unnecessary, but also the resulting estimator of $\boldsymbol{\beta}$ is semiparametric efficient.

Remark 2. Under local linear smoothing, Corollary 1 and Proposition 2 hold for a wide range of bandwidths. That is, $\widehat{\boldsymbol{\beta}}$ is quite insensitive to the choice of bandwidth. For example, any data-driven method of order $O_P(n^{-1/5})$ fulfills the requirements. With m_i being finite, Wang (2003) illustrated the following asymptotic phenomenon, namely that the proposed estimate of θ is essentially a locally weighted estimate of clusterwise pseudoresponses, where the ith pseudoresponse is formed by a linear combination of responses in the ith cluster. This implies that the derivations that justify use of plug-in method or cross-validation under the independent data scenario can be equivalently carried out here, and that the asymptotic bias and variance given by Wang can be used to show that the resulting choice of h is of order $n^{-1/5}$, where n is the number of subject.

Consequently, there are a host of data-driven bandwidths can be used to obtain the same asymptotic distribution of $\widehat{\boldsymbol\beta}$. Examples include the plug-in bandwidth discussed by Wang (2003) and the leave-one-subject-out cross-validation using either the WI or the proposed $\widehat{\boldsymbol\theta}$. That is, with the proposed estimating procedure, which data-driven bandwidth to use is not of particular concern, at least asymptotically. An illustration of this phenomenon is provided through a simulation study in Section 6.

Remark 3. Lin and Carroll (2001a) allowed different working covariances in their estimating equations of $\boldsymbol\theta(\mathbf{t})$ and $\boldsymbol\beta$. This is motivated by the fact that the most efficient conventional kernel estimator requires ignoring correlation, whereas a more efficient estimator of $\boldsymbol\beta$ requires accounting for correlation. A similar approach can be adopted in our method by simply using V_{1i} and V_{2i} to replace $\mathbf{V}_i$ in (3) and (4). The dependence of the result on $\mathbf{V}_1$ is implicitly embedded in $\widetilde{\mathbf{X}}$, whereas $\boldsymbol\Omega(\mathbf{V},\boldsymbol\Sigma)$ needs to be replaced by $\boldsymbol\Omega(\mathbf{V}_2,\boldsymbol\Sigma)$. However, there is no advantage of doing this in our framework, because our results show that when $\mathbf{V}_1=\mathbf{V}_2=\boldsymbol\Sigma$, our method gives the most efficient estimators of both $\boldsymbol\theta(\mathbf{t})$ and $\boldsymbol\beta$. Further, if we allow $\mathbf{V}_1$ and $\mathbf{V}_2$ to be different in our method, then a consistent estimator of $\boldsymbol\beta$ requires undersmoothing.

The estimation framework considered by Zeger and Diggle (1994) is a special case of the conventional profile-kernel estimating structure of LC. Specifically, they assumed working independence $\mathbf{V}_1$ for estimating $\boldsymbol\theta(\mathbf{t})$ and a nondiagonal $\mathbf{V}_2$ that accounts for the within-cluster correlation for estimating $\boldsymbol\beta$. LC referred to an extension of their estimator as the "undersmoothed profile-kernel estimator," where $\boldsymbol\theta(\mathbf{t})$ is undersmoothed to guarantee that the estimator of $\boldsymbol\beta$ is $\sqrt{n}$-consistent. Even though Zeger and Diggle carried out their calculations with a backfitting method and we concentrate on kernel profile estimation, hereafter we refer to the estimator of $\boldsymbol\beta$ using working independence $\mathbf{V}_1$ and estimated $\mathbf{V}_2$ as the Zeger–Diggle (ZD) estimator to credit these authors for their original idea of choosing the pair $(\mathbf{V}_1,\mathbf{V}_2)$.

Note that in theory, this undersmoothed estimator is still not semiparametric efficient even when assuming $\mathbf{V}_2=\boldsymbol\Sigma$. Our numerical ARE study in Section 5 suggests that this estimator has a high relative efficiency. However, as pointed out by LC, the consistency of the ZD estimator of $\boldsymbol\beta$ requires undersmoothing in estimating $\boldsymbol\theta(\mathbf{t})$. Further, as also pointed out by LC, a practical drawback of this estimator is that a "regular" sandwich variance estimator cannot be used. A variance estimator would involve either empirically estimating the complicated $\mathbf{Z}_1$ and $\mathbf{Z}_2$ terms given in Section A.5 or using a bootstrap method.

Remark 4. Because $\boldsymbol\Omega(\mathbf{V},\boldsymbol\Sigma)$ is minimized when $\mathbf{V}=\boldsymbol\Sigma$ (i.e., when the correct covariance structure is specified), Proposition 2 implies that the proposed estimator is more efficient than the WI estimator that uses $\mathbf{V}=\mathbf{I}$, the identity matrix.

Remark 5. Parallel to standard GEEs (Liang and Zeger 1986), it can be shown that our estimator is still consistent when the working covariance matrix $\mathbf{V}$ is misspecified and is most efficient when $\mathbf{V}$ is correctly specified. Obviously, more efficiency can be gained by adopting a more complicated estimating equation for $\boldsymbol\beta$ under certain special models, such as those with part of $\boldsymbol\tau$ being $\boldsymbol\beta$; that is, there is information for $\boldsymbol\beta$ beyond the mean. This has been done in parametric cases (see, e.g., Prentice and Zhao 1991; Crowder 2001). As pointed out in the literature, and also shared by our own experience, little information is gained from the added complexity. A relevant discussion was given by Crowder (2001) for parametric cases. In this respect, issues to be considered for the proposed estimator are the same as those in parametric models.

For simplicity, we assume in our asymptotic work that the working correlation parameter vector $\boldsymbol\tau$ in $\mathbf{V}$ is known. It can be estimated via the method of moments using a quadratic function of $\mathbf{Y}$'s. Following Lin and Carroll (2001b), it can be shown that once such an estimator of $\boldsymbol\tau$ converges in probability to some $\boldsymbol\tau^*$ at a $\sqrt{n}$ rate, then there is no asymptotic effect on our estimators of $\boldsymbol\beta$ and $\boldsymbol\theta(\cdot)$ due to estimation of $\boldsymbol\tau$; that is, Proposition 2 still holds. In addition, with $\boldsymbol\tau$ being of finite dimension, following remark 3.2 of Begun, Hall, Huang, and Wellner (1983), it can be shown that the semiparametric information bound given earlier remains the same whether $\boldsymbol\tau$ is known or estimated. Consequently, in the case that we consider, our estimator of $\boldsymbol\beta$ is semiparametric efficient when $\widehat{\boldsymbol\tau}$ is $\sqrt{n}$-consistent to certain $\boldsymbol\tau^*$. Following Carroll, Wu, and Ruppert (1988), one could iteratively update the estimated τ, and no more than three or four iterations of this process should suffice even for second-order purposes.

5. ASYMPTOTIC RELATIVE EFFICIENCY OF ESTIMATED β

In this section we study numerically the AREs of the WI and ZD estimators with respect to the semiparametric-efficient estimator. We concentrate on the efficiencies of the estimators of β. The AREs the WI estimator and the proposed estimator of $\boldsymbol\theta(\mathbf{t})$ are the same as reported by Wang (2003).

We consider the case where the cluster size is constant ($m_i=m$) and X_{ij} is a scalar Gaussian covariate. The underlying model is $Y_{ij}=X_{ij}\beta+\theta(T_{ij})+\epsilon_{ij}$. Let $\underline{\boldsymbol\epsilon}_i=(\epsilon_{i1},\ldots,\epsilon_{im})^T$ with $\underline{\boldsymbol\epsilon}_i\sim\mathrm{N}(\mathbf{0},\boldsymbol\Sigma)$, and let σ_{jj} be the jth diagonal element of $\boldsymbol\Sigma$. To simplify calculations, we also assume that T_{ij} is Gaussian, even though this violates the assumption that $f_j(\mathbf{t})$ has to be bounded away from 0. One can view the resulting efficiency calculation as an approximation of that when T_{ij} follows a truncated normal. The advantage of assuming normality is that the integral equation (6) has a closed-form solution and thus the semiparametric information bound has a closed form. Specifically, we assume that both $\underline{\mathbf{X}}_i$ ($m\times 1$) and $\underline{\mathbf{T}}_i$ ($m\times 1$) are centered multivariate normal random variables with mean $\mathbf{0}$ and covariances $\mathrm{cov}(\underline{\mathbf{X}}_i)=\sigma_{\mathbf{X}}^2\boldsymbol\Gamma_{\mathbf{XX}}$, $\mathrm{cov}(\underline{\mathbf{T}}_i)=\sigma_{\mathbf{T}}^2\boldsymbol\Gamma_{\mathbf{TT}}$, and $\mathrm{cov}(\underline{\mathbf{X}}_i,\underline{\mathbf{T}}_i)=\sigma_{\mathbf{X}}\sigma_{\mathbf{T}}\boldsymbol\Gamma_{\mathbf{XT}}$, where $\boldsymbol\Gamma_{\mathbf{XX}}=\{\rho_{jk}^{\mathbf{XX}}\}$, $\boldsymbol\Gamma_{\mathbf{TT}}=\{\rho_{jk}^{\mathbf{TT}}\}$, and $\boldsymbol\Gamma_{\mathbf{XT}}=\{\rho_{jk}^{\mathbf{XT}}\}$ are the correlation matrices. That is, we assume that all X_{ij}'s share the same marginal distribution with a common variance $\sigma_{\mathbf{X}}^2$. Similarly, all T_{ij}'s share the same marginal distribution with a common variance $\sigma_{\mathbf{T}}^2$.

We first calculate the semiparametric-efficient score by solving the integral equation (6). For simplicity, we suppress the subscript i in the following discussion. Under the foregoing multivariate normal assumption of $\underline{\mathbf{X}}$ and $\underline{\mathbf{T}}$, calculations sketched in Section A.5 show that (6) has a closed-form solution,

$$\boldsymbol{\varphi}_{\mathrm{eff}}(\mathbf{t}) = \frac{\sigma_{\mathbf{X}} \sum_{j=1}^{m} \sum_{k=1}^{m} \sigma^{jk} \rho_{jk}^{\mathbf{XT}}}{\sigma_{\mathbf{T}} \sum_{j=1}^{m} \sum_{k=1}^{m} \sigma^{jk} \rho_{jk}^{\mathbf{TT}}} \mathbf{t}, \tag{10}$$

where σ^{jk} is the (j,k)th element of $\boldsymbol{\Sigma}^{-1}$. Let $\widetilde{\underline{\mathbf{X}}}_{\mathrm{eff}} = \underline{\mathbf{X}} - \boldsymbol{\varphi}_{\mathrm{eff}}(\underline{\mathbf{T}})$. The semiparametric information bound of β is $I_{\mathrm{eff}}(\beta) = E(\widetilde{\underline{\mathbf{X}}}_{\mathrm{eff}}^T \boldsymbol{\Sigma}^{-1} \widetilde{\underline{\mathbf{X}}}_{\mathrm{eff}})$.

For the WI estimator, we assume that the working covariance matrix is $\boldsymbol{\Sigma}_d = \mathrm{diag}(\boldsymbol{\Sigma})$. Using result 1 of LC (2001a), the asymptotic information matrices of the WI and ZD estimators are

$$I_{\mathrm{WI}}(\beta) = E(\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}^T \boldsymbol{\Sigma}_d^{-1} \widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}) E(\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}^T \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_d^{-1} \widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}})^{-1} \times E(\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}^T \boldsymbol{\Sigma}_d^{-1} \widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}})$$

and

$$I_{\mathrm{ZD}}(\beta) = E(\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}^T \boldsymbol{\Sigma}^{-1} \widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}) E\{(\underline{\mathbf{Z}}_1 - \underline{\mathbf{Z}}_2)^T \boldsymbol{\Sigma} (\underline{\mathbf{Z}}_1 - \underline{\mathbf{Z}}_2)\}^{-1} \times E(\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}^T \boldsymbol{\Sigma}^{-1} \widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}),$$

where $\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}} = \underline{\mathbf{X}} - \boldsymbol{\varphi}_{\mathrm{WI}}(\underline{\mathbf{T}})$, $\boldsymbol{\varphi}_{\mathrm{WI}}(\mathbf{t}) = \{\sigma_{\mathbf{X}} \sum_{j=1}^{m} \sigma_{jj}^{-1} \rho_{jj}^{\mathbf{XT}}\}/\{\sigma_T \sum_{j=1}^{m} \sigma_{jj}^{-1}\}\mathbf{t}$, and $\underline{\mathbf{Z}}_1$ and $\underline{\mathbf{Z}}_2$ are as defined in Section A.5. The two AREs of interest are

$$ARE_{\mathrm{WI}}(\beta) = \frac{I_{\mathrm{WI}}(\beta)}{I_{\mathrm{eff}}(\beta)} \qquad \text{and} \qquad ARE_{\mathrm{ZD}}(\beta) = \frac{I_{\mathrm{ZD}}(\beta)}{I_{\mathrm{eff}}(\beta)}.$$

It can be easily seen that these ARE are free of the marginal variances of $\mathbf{X}$ and $\mathbf{T}$, that is, of $\sigma_{\mathbf{X}}^2$ and $\sigma_{\mathbf{T}}^2$.

We performed a numerical ARE study by assuming an exchangeable correlation structure on $\underline{\mathbf{Y}}$, $\underline{\mathbf{X}}$, and $\underline{\mathbf{T}}$; that is, with $\mathbf{I}$ being the identity matrix and $\mathbf{J}$ being a matrix with all elements equal to 1,

$$\boldsymbol{\Sigma} = \sigma^2\{(1-\rho)\mathbf{I} + \rho\mathbf{J}\},$$

$$\boldsymbol{\Gamma}_{\mathbf{XX}} = (1-\rho_{\mathbf{X}})\mathbf{I} + \rho_{\mathbf{X}}\mathbf{J},$$

$$\boldsymbol{\Gamma}_{\mathbf{TT}} = (1-\rho_{\mathbf{T}})\mathbf{I} + \rho_{\mathbf{T}}\mathbf{J}.$$

Furthermore, we let $\boldsymbol{\Gamma}_{\mathbf{XT}} = \rho_{\mathbf{XT}}\{(1-\delta)\mathbf{I} + \delta\mathbf{J}\}$, with $0 < \delta \le 1$. That is, $\mathrm{corr}(X_{ij}, T_{ij}) = \rho_{\mathbf{XT}}$ and for $j \ne k$, $\mathrm{corr}(X_{ij}, T_{ik}) = \delta\rho_{\mathbf{XT}}$, which could be smaller than the correlation between the paired X_{ij} and T_{ij} measured at the same time. Throughout, we set $\delta = .6$ and $\rho_{\mathbf{T}} = .3$.

Assuming the cluster size $m = 4$, Figures 1(a) and 1(b) displays the asymptotic relative efficiencies ARE_{WI} and ARE_{ZD} as functions of ρ, the correlation among the outcome $\underline{\mathbf{Y}}$. We assume the correlations among $\underline{\mathbf{X}}$ and between $\underline{\mathbf{X}}$ and $\underline{\mathbf{T}}$ to be $\rho_{\mathbf{X}} = \rho_{\mathbf{XT}} = .3$ and .6, which represent low and moderate levels of correlation. The results are depicted by the solid (for .3) and dotted (for .6) curves.

Figure 1 shows that the WI estimator is subject to a moderate amount of efficiency loss even when the correlation among the outcomes $\mathbf{Y}$ is modest. The loss of efficiency becomes substantial when the correlation among the outcomes $\mathbf{Y}$ becomes large. The ZD estimator, which assumes the true correlation in estimating β, has a much higher relative efficiency compared with the WI estimator. For example, when $\rho = .6$, $\rho_{\mathbf{X}} = \rho_{\mathbf{XT}} = .3$ and .6, ARE_{WI} and ARE_{ZD} are 44.7% and 98.8%, and 47.6% and 92.8%. Considerable loss of efficiency is found in the ZD estimator only when ρ is very large.

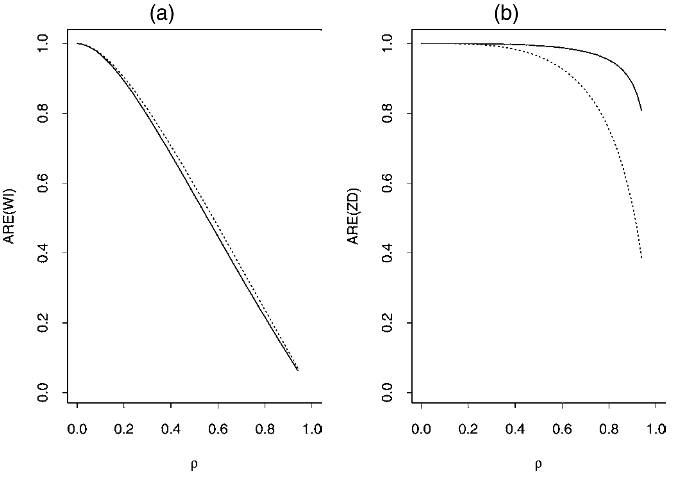

Figure 1. AREs of the (a) WI Estimator and (b) the ZD Estimator for Different $\rho_{\mathbf{X}}$ and $\rho_{\mathbf{XT}}$. The solid curves correspond to a scenario with $\rho_{\mathbf{X}} = \rho_{\mathbf{XT}} = .3$, whereas the dotted curves correspond to $\rho_{\mathbf{X}} = \rho_{\mathbf{XT}} = .6$. For both scenarios, the cluster size $m = 4$, $\rho_{\mathbf{T}} = .3$, and $\delta = .6$.

The exchange of the relative positions of the two curves in Figure 1 suggests that the relationship between the ARE and the level of correlation within and among $\underline{\mathbf{X}}$ and $\underline{\mathbf{T}}$ differs between the WI and ZD estimators. For example, as the correlation between $\underline{\mathbf{X}}$ and $\underline{\mathbf{T}}$ increases, the loss of efficiency of the ZD estimator increases, whereas that of the WI estimator decreases slightly.

Concentrating on the scenario with $\rho_{\mathbf{X}} = \rho_{\mathbf{XT}} = .6$, Figure 2 illustrates the changes in relative efficiencies with the cluster size m. As in Figure 1, panels (a) and (b) display the AREs ARE_{WI} and ARE_{ZD} as functions of ρ. The curves, from top to bottom, correspond to $m = 3, 4, 5$, and 6. For both the

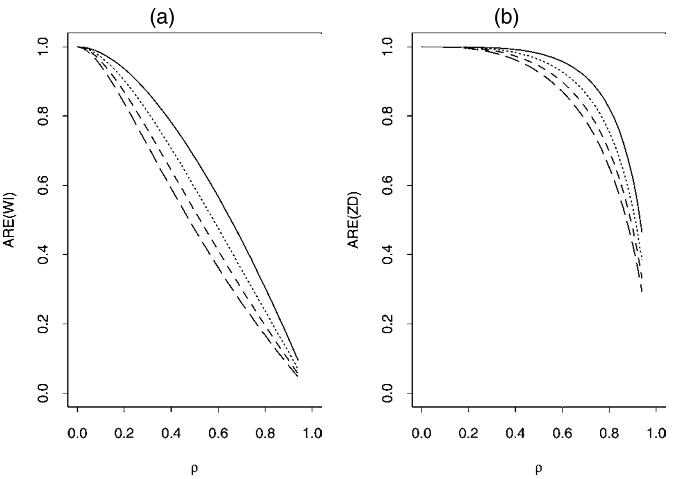

Figure 2. AREs of the (a) WI Estimator and (b) the ZD Estimator for Different Cluster Sizes, m. The curves from top to bottom correspond to $m = 3, 4, 5$, and 6.

WI and ZD estimators, the loss of efficiency increases with increasing cluster size m.

6. A SIMULATION STUDY

In this section we report a simulation study to investigate the finite-sample performance of the proposed estimator and compare it with the WI and ZD estimators. We consider the following longitudinal scenario. For each subject, we generated the time-varying covariates $\underline{\mathbf{T}}$ and $\underline{\mathbf{X}}_1$ as sums of independent uniform $[-1, 1]$ random variables and a common uniform $[0, 1]$ random variable. This made each $\mathbf{X}_1$ and $\mathbf{T}$ dependent and yielded $\rho_{\mathbf{X}} = \rho_{\mathbf{T}} = \rho_{\mathbf{XT}} = .2$ and $\delta = 1$. We also included a time-independent covariate $\underline{\mathbf{X}}_2$, which equals 0 for half of the subjects and 1 for the other half and mimics a treatment indicator. We generated the response Y_{ij} assuming a conditional mean $E(Y_{ij}|\mathbf{X}_{ij}) = \sin(2T_{ij}) + \beta_1 X_{1ij} + \beta_2 X_{2ij}$, a common variance 1 and an exchangeable correlation structure with $\rho = .6$. We let $\beta_1 = \beta_2 = 1$, $n = 100$, and $m = 4$. We generated 250 datasets with $\underline{\mathbf{X}}_1$ and $\underline{\mathbf{T}}$ regenerated each time. An exchangeable correlation structure was assumed with ρ being estimated using the method of moments. All estimates, including the profile iterative kernel, WI, and ZD methods, were computed using the Epanechnikov kernel for K.

To understand how insensitive the proposed estimator is to the choice of the bandwidth h, we did the following. For the first 50 datasets, we estimated the bandwidth using a method mimicking the idea of the empirical bias bandwidth selection method (Ruppert 1997; Wang 2003) and the leave-one-subject-out cross-validation methods in nonparametric estimation with the WI and the proposed $\widehat{\theta}$. We then further expanded the range of selected bandwidths to [.35, .65]. We evaluated the performance of the three estimators for seven bandwidths equally spaced between .35 and .65. For the ZD estimate, each bandwidth was further multiplied by $(n \times m)^{-2/15}$, an undersmoothing required for $\sqrt{n}$-consistency of $\boldsymbol{\beta}$ estimation. The ratios of the resulting Monte Carlo variances and mean squared errors (MSEs) of each of the three estimates relative to the proposed efficient estimate as functions of bandwidths are displayed in Figure 3. Panels (a) and (b) show the estimates of β_1, and (c) and (d) show the estimates of β_2. The solid, dotted, and dashed curves correspond to the proposed, WI, and ZD estimates. These results demonstrate that the estimates of $\boldsymbol{\beta}$ perform about equally well in this specified range containing multiple data-driven bandwidths for all datasets. This implies that any reasonable data-driven bandwidth can work well here. As a representative illustration, Table 1 summarizes the averaged biases, standard errors (SEs), and MSEs of the estimates of $\boldsymbol{\beta}$ for bandwidth $h = .45$.

The three estimates of β_2 had very similar performance. This is consistent with the theory. Because $\underline{\mathbf{X}}_2$ and $\underline{\mathbf{T}}$ are independent and $\underline{\mathbf{X}}_2$ is balanced among all subjects, it is expected theoretically that the three estimates should be equivalent at least up to the first order. The relative ratios of the absolute biases among three estimators ranged from .7 to 1.55. No one estimator was consistently better than another.

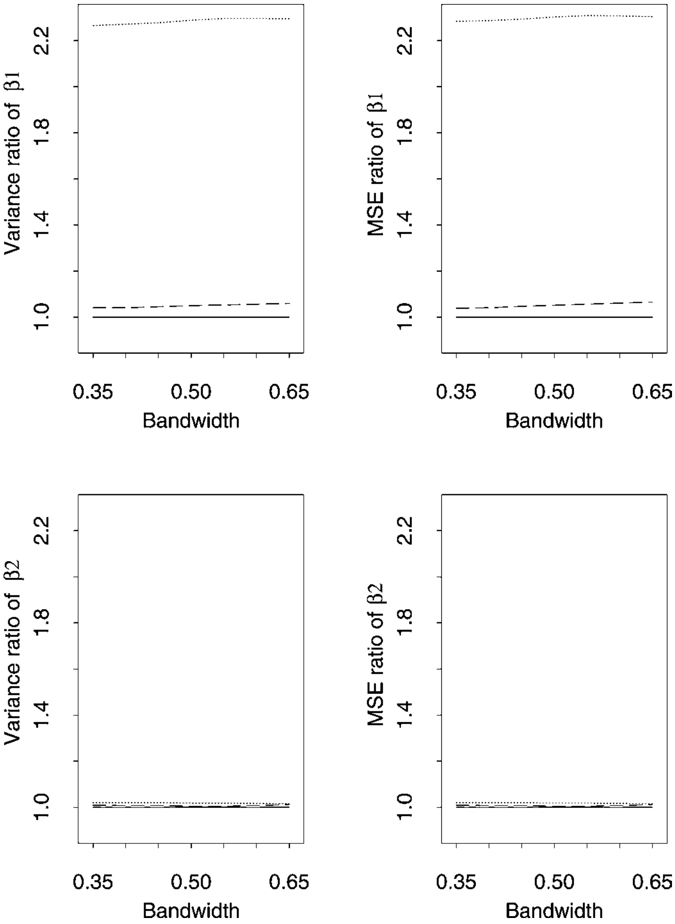

Figure 3. The Ratios of Monte Carlo Variances and MSEs of Each of the Three Estimates of $\widehat{\boldsymbol{\beta}}$ Relative to the Proposed Efficient Estimate as Functions of Bandwidths. The solid, dotted, and dashed curves correspond to the proposed, WI, and ZD estimates. The top two panels are for $\widehat{\beta}_1$, while the bottom two are for $\widehat{\beta}_2$.

For $\widehat{\beta}_1$, both the variance and absolute bias of the proposed estimate were uniformly smaller than that of the ZD estimate for each bandwidth considered, and the variances and absolute biases of both estimates were uniformly much smaller than those of the WI estimate. Compared with the WI estimate, the proposed and ZD estimates reduced the variances by more than 50%. The range of the absolute bias ratios of the WI estimate over the proposed estimate varied from 1.29 to 11.18. Nonetheless, Table 1 and a comparison between the variance and MSE plots in Figure 1 suggest that the bias is not of concern and the variance is a dominating factor when comparing the MSEs among the three estimates. Selecting the bandwidth equal to .45, the sandwich SE estimates of the proposed method agreed well with the empirical SEs. For example, when $h = .45$, the Monte Carlo SE of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ were .0567 and .1632 for the proposed method, whereas the averages of the corresponding sandwich estimated SEs were .0551 and .1612. Finally, the new nonparametric estimate of $\theta(t)$ was more efficient than the WI

Table 1. Summary of Simulation Study Results From 250 Replications

Parameter	*Method*	*Bias*	*SE*	*MSE*
$\beta_1 = 1$	WI	.0732	.8564	.0739
	ZD	.0222	.5803	.0337
	NEW	.0118	.5675	.0322
$\beta_2 = 1$	WI	.0135	1.6486	.2718
	ZD	.0134	1.6379	.2683
	NEW	.0107	1.6324	.2665

NOTE: The bandwidth is .45. Each entry equals the original value multiplied by 10.

estimate. The average MSEs of the WI estimates over the range of the bandwidths we considered were about 1.6 times of those using the proposed method.

7. APPLICATION TO THE LONGITUDINAL CD4 CELL COUNT DATA

We applied the semiparametric model given in (1) and (2) to the longitudinal CD4 cell count data among HIV seroconverters previously analyzed by Zeger and Diggle (1994). This study involved 369 subjects whose CD4 counts were measured during a period ranging from 3 years before to 6 years after seroconversion. A total of 2,376 CD4 measurements were available, and the number of CD4 observations per subject varied from 1 to 12, with most of the subjects having between 4 and 10 observations. It was of interest to estimate the average time course of CD4 counts and the effects of other covariates. These covariates included age, smoking status measured by packs of cigarettes, drug use (yes, 1; no, 0), number of sex partners, and depression status measured by the CESD scale (large values indicating more depression symptoms). (See Zeger and Diggle 1994 for a more detailed description of the data.)

Let $\underline{\mathbf{T}}$ be years since seroconversion. We conducted an analysis on the square root–transformed CD4 counts using the WI estimator and the proposed efficient estimator. The purpose behind the transformation is to reduce skewness of the original CD4 measurements, as indicated by Zeger and Diggle (1994). Our results in Section 6 indicate that neither estimator is sensitive to the choice of bandwidth. Therefore, we simply used a "partial" leave-one-subject-out cross-validation that dropped 50 randomly selected subjects one at a time in nonparametric estimation to select a bandwidth of 1.86. To ensure that our data analysis was indeed insensitive to the bandwidth selection, we repeated the analysis by reducing and increasing the bandwidth by 50%. The changes in coefficients and SEs were minimal. We hence report the results using the bandwidth 1.86.

For the proposed estimator, we used a working covariance structure described by ZD as "random intercept plus serial correlation and measurement error." More precisely, we assumed a random intercept and an exponential decay serial correlation by specifying the covariance structure as $\tau^2\mathbf{I} + \nu^2\mathbf{J} + \omega^2\mathbf{H}$, where $\mathbf{J}$ is a matrix of 1's and $H(j,k) = \exp(-\alpha|T_{ij} - T_{ik}|)$. The covariance estimates obtained by the ZD estimator were $\widehat{\boldsymbol{\xi}} = (\widehat{\tau}^2, \widehat{\nu}^2, \widehat{\omega}^2, \widehat{\alpha}^2) = (14.1, 6.9, 16.1, .22)$. By leaving out residuals in the boundary and coupling a least squares method in variogram analysis and a moment variance estimation approach, we obtained a slightly different set of estimates, $\widehat{\boldsymbol{\xi}} = (11.32, 3.26, 22.15, .23)$. Table 2 refers to our and the ZD working covariances as "scenario I" and "scenario II."

Table 2 gives the regression coefficient estimates of the parametric covariates using the WI and proposed efficient methods. The SEs were all calculated using the sandwich method. Based on the new method, smoking and the number of sex partners were significantly positively associated with the CD4 counts, whereas age, drug use and depression had no significant effects. Note some fairly large numerical differences between the WI and the proposed estimates for smoking and drug use, a change of sign and statistical significance for number of sex partners, and overall much smaller SEs for our method.

The decrease in SEs is in accordance with our theory; the other phenomena are more difficult to explain. Nonetheless, they are not unique to semiparametric GEE methods. Similar discrepant outcomes occurred in parametric GEE estimation in which $\boldsymbol{\theta}(\mathbf{t})$ was replaced by a cubic regression function in time. Furthermore, we simulated data using the observed covariates but with responses generated from the multivariate normal with mean equal to the fitted mean in the parametric correlated GEE estimation and with correlation given in scenario II. The level of divergence between two sets of results in the simulated data was fairly consistent with what appeared in Table 2. For example, among the first 25 generated datasets, 3 had different signs in sex partners and 7 had the scale of drug use coefficient obtained by WI 1.8 times or larger than what obtained by the proposed method. When we applied the proposed method to 100 such generated datasets, the means of Monte Carlo estimated SEs and the Monte Carlo SEs are fairly close to each other. In fact, the ratios of these two sets of values are practically identical to the equivalent ratios obtained using parametric GEE.

Comparing the estimates obtained in the two scenarios accounting for correlations, we note that using a slightly different covariance estimate does not change the outcome much, even though the estimates are quite different with or without considering correlations.

The nonparametric curve estimates using the WI (dotted line) and proposed (solid line) estimators are plotted in Figure 4(a). The CD4 counts were stable before seroconversion and sharply decreased after seroconversion. By accounting for correlation, our method suggests that the decreasing trend remained after 2 years. The estimated SEs are given in Figure 4(b). The SE of the proposed curve estimate is uniformly smaller than that of the WI estimate. The results agree with the theory.

Table 2. Regression Coefficients in the CD4 Cell Counts Study in HIV Seroconverters Using the Semiparametric Efficient and the Working Independence Estimate

	Working independence		*Semiparametric efficient scenario I*		*Semiparametric efficient scenario II*	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Age	.014	.035	.010	.033	.008	.032
Smoking	.984	.182	.549	.144	.579	.139
Drug	1.049	.526	.584	.331	.584	.335
Sex partners	−.054	.059	.080	.038	.078	.039
Depression	−.033	.021	−.045	.013	−.046	.014

NOTE: For the semiparametric efficient estimates, the working covariance parameter, $\widehat{\xi} = (11.32, 3.26, 22.15, .23)$ for scenario I, and $\widehat{\xi} = (14.1, 6.9, 16.1, .22)$, for scenario II.

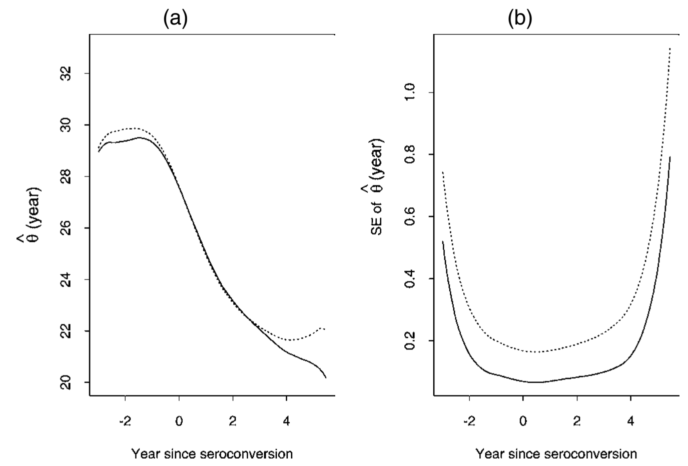

Figure 4. Two $\widehat{\theta}(t)$'s (a) and Their Estimated Pointwise SEs (b). The solid and dotted curves correspond to the proposed estimates and the WI estimates.

8. DISCUSSION

We have considered the marginal semiparametric partial generalized linear model previously discussed by LC (2001a) for clustered data, where the effects of some covariates $\underline{\mathbf{X}}$ are modeled parametrically and the effect of a covariate $\underline{\mathbf{T}}$ is modeled nonparametrically as $\theta(t)$. LC showed that the conventional profile-kernel method failed to yield a semiparametric efficient estimator of $\boldsymbol{\beta}$. By simply replacing the nonparametric estimator in LC's original profile-kernel method by the newly proposed iterative kernel estimator $\widehat{\theta}(t)$, we were able to construct a semiparametric efficient estimator of $\boldsymbol{\beta}$ under the same multivariate normal scenario given by LC.

Unlike in the LC model, a regular bandwidth can be used, and undersmoothing is no longer needed to construct $\sqrt{n}$-consistent estimates of $\boldsymbol{\beta}$ when accounting for correlations. In addition, the proposed $\widehat{\theta}(t)$ has less variation than the WI estimator. Our numerical results suggest that the proposed method performs well in finite samples and outperforms the WI method. They also suggest that the proposed $\widehat{\boldsymbol{\beta}}$ is relatively insensitive to the choice of bandwidth. Most important, we have shown that properly accounting for the within-subject correlation can reduce variation of parameter estimates in the general semiparametric model (2), just as in parametric models.

APPENDIX: CONDITIONS AND PROOFS

Assume that each f_j has a compact support and that on its support, f_j is bounded away from 0. Throughout this appendix, we assume that the equivalent convexity conditions given by Carroll, Fan, Gijbels, and Wand (1997) hold. These conditions ensure that the $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ obtained in (3) and (4) exist uniquely and lie in a compact set. We also assume that conditions equivalent to condition 2 of Carroll et al. (1997) hold. The purpose behind these assumptions is to establish uniform convergence of $\widehat{\theta}$ and $\widehat{\boldsymbol{\varphi}}$. The structure of the proof has been given by Mack and Silverman (1982) and the proof of lemma A.1 and equation (A.5) of Carroll et al. (1997). We further assume that $\iint H^2(t,s)\,dt\,ds < 1$. This condition ensures the existence and uniqueness of a solution to (6) (see Kress 1989, chap. 2). With a linear link and iid $\mathbf{T}_1,\ldots,\mathbf{T}_m$, this condition is equivalent to a constraint on dependent structure of the responses from the same subject. Except in the first half of Section A.1, we concentrate on $m_i \equiv m$. We let $n \to \infty$ and $h \to 0$ at the rate that $nh^8 \to 0$ and $nh/\log(1/h) \to \infty$.

A.1 Semiparametric Efficient Score

Under the multivariate normal model, the joint density of $\{\underline{\mathbf{Y}}_i, \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i\}$ is

$$f_{\mathbf{Y}|\mathbf{X},\mathbf{T}}\big(\underline{\mathbf{y}}_1,\ldots,\underline{\mathbf{y}}_n\big|\{\underline{\mathbf{x}}_i,\underline{\mathbf{t}}_i\}\big)f_{\mathbf{X},\mathbf{T}}(\underline{\mathbf{x}}_1,\ldots,\underline{\mathbf{x}}_n,\underline{\mathbf{t}}_1,\ldots,\underline{\mathbf{t}}_n), \qquad (\mathrm{A.1})$$

where $f_{\mathbf{X},\mathbf{T}}$ is the joint density of $\{\underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i\}$ and $f_{\mathbf{Y}|\mathbf{X},\mathbf{T}}$ is multivariate normal with conditional means specified in (2) and conditional variance $\boldsymbol{\Sigma}_i(\underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$. Following Bickel et al. (1993, chap. 3), we first define the following submodels:

P_1: {Model (A.1) with known $\theta_0(\cdot)$ and $f_{\mathbf{X},\mathbf{T}}(\cdot)$},
P_2: {Model (A.1) with known $\boldsymbol{\beta}_0$ and $f_{\mathbf{X},\mathbf{T}}(\cdot)$},
and
P_3: {Model (A.1) with known $\boldsymbol{\beta}_0$ and $\theta_0(\cdot)$}.

For the parametric family P_1, the score function for $\boldsymbol{\beta}_0$ is

$$\mathbf{S}_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \underline{\mathbf{X}}_i^t \boldsymbol{\Delta}_i \boldsymbol{\Sigma}_i^{-1}\big[\underline{\mathbf{Y}}_i - \boldsymbol{\mu}\{\underline{\mathbf{X}}_i\boldsymbol{\beta}_0 + \theta_0(\underline{\mathbf{T}}_i)\}\big].$$

By linearly spanning the score functions of parametric submodels of P_2 with $\theta(\cdot)$ replaced by $\theta(\eta,\cdot)$, the tangent space of P_2 is

$$\dot{P}_2 = \left\{\sum_{i=1}^{n} \boldsymbol{\varphi}^t(\underline{\mathbf{T}}_i)\boldsymbol{\Delta}_i\boldsymbol{\Sigma}_i^{-1}\big[\underline{\mathbf{Y}}_i - \boldsymbol{\mu}\{\underline{\mathbf{X}}_i\boldsymbol{\beta}_0 + \theta_0(\underline{\mathbf{T}}_i)\}\big], \quad \text{where } \boldsymbol{\varphi}(\cdot) \in L_2\right\}.$$

We need only concentrate on $\dot{P}_2$, because it is easy to see that $\mathbf{S}_{\boldsymbol{\beta}}$ and any member in $\dot{P}_2$ are orthogonal to the score function in any parametric submodel of P_3; consequently, they are orthogonal to $\dot{P}_3$.

By theorem 1 of Bickel et al. (1993, sec. 3.4), the efficient score $\mathbf{S}^*_{\boldsymbol{\beta}} = \mathbf{S}_{\boldsymbol{\beta}} - \prod(\mathbf{S}_{\boldsymbol{\beta}}|\dot{P}_2)$ is $\sum_i\{\underline{\mathbf{X}}_i - \boldsymbol{\varphi}_{\mathrm{eff}}(\underline{\mathbf{T}}_i)\}^t\boldsymbol{\Delta}_i\boldsymbol{\Sigma}_i^{-1}[\underline{\mathbf{Y}}_i - \boldsymbol{\mu}\{\underline{\mathbf{X}}_i\boldsymbol{\beta}_0 + \theta_0(\underline{\mathbf{T}}_i)\}]$, where $\boldsymbol{\varphi}_{\mathrm{eff}}(\cdot)$ satisfies the requirement that $\mathbf{S}^*_{\boldsymbol{\beta}}$ is orthogonal to any member in $\dot{P}_2$. That is, $\boldsymbol{\varphi}_{\mathrm{eff}}$ needs to satisfy $\sum_i E[\{\underline{\mathbf{X}}_i - \boldsymbol{\varphi}_{\mathrm{eff}}(\underline{\mathbf{T}}_i)\}^t\boldsymbol{\Delta}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Delta}_i\boldsymbol{\varphi}(\underline{\mathbf{T}}_i)] = 0$ for any $\boldsymbol{\varphi}$. This is equivalent to

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{\ell=1}^{m_i} E\big[\Delta_{ijj}\sigma^{ij\ell}\Delta_{i\ell\ell}\{\mathbf{X}_{i\ell} - \boldsymbol{\varphi}_{\mathrm{eff}}(T_{i\ell})\}\boldsymbol{\varphi}(T_{ij})\big] = 0. \qquad (\mathrm{A.2})$$

Up to this point, we have not assumed that $m_i \equiv m$. The results in the main text are reported with this assumption for the purpose of a clean presentation. Without this assumption, but with a bound of the m_i, in the limit as $n \to \infty$, the efficient score (A.2) becomes the obvious weighted average of the interior sums $j, \ell = 1,\ldots,m_i$.

When $m_i \equiv m$ so that $(\underline{\mathbf{Y}}_i, \underline{\mathbf{X}}_i, \underline{\mathbf{T}}_i)$ are iid, (A.2) is equivalent to

$$\sum_{j=1}^{m}\sum_{\ell=1}^{m} E\big[\Delta_{jj}\sigma^{j\ell}\Delta_{\ell\ell}\{\mathbf{X}_\ell - \boldsymbol{\varphi}_{\mathrm{eff}}(\mathbf{T}_\ell)\}\boldsymbol{\varphi}(\mathbf{T}_j)\big] = 0,$$

which leads directly to (5).

A.2 Asymptotic Structure of $\widehat{\theta}$

A condition guaranteeing a unique solution to (6) was given at the beginning of the Appendix. The purpose of this section is to provide an asymptotic expansion for $\widehat{\theta}_{[k]}(t,\boldsymbol{\beta}_0)$ at the kth iteration and at the convergence, when the iterations converge. As described in Section 3, the WI estimator, denoted by $\widehat{\theta}_{[0]}(t,\boldsymbol{\beta}_0)$, is used as an initial estimator. Its asymptotic expansion is

$$\begin{aligned}\widehat{\theta}_{[0]}&(t,\boldsymbol{\beta}_0) - \theta(t)\\ &= \frac{1}{2}b_{[0]}(t)h^2 + W_2^{-1}(t)n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}^{(1)}v_i^{jj}K_h(T_{ij}-t)(Y_{ij}-\mu_{ij})\\ &\quad + o_p\big(h^2 + \{\log(n)/nh\}^{1/2} + n^{-1/2}\big), \end{aligned} \qquad (\mathrm{A.3})$$

where

$$b_{[0]}(\mathbf{t}) = \theta^{(2)}(\mathbf{t}) \qquad \text{and}$$
$$W_2(\mathbf{t}) = \sum_{j=1}^{m} E\{\Delta_{jj}^2 v^{jj} | \mathbf{T}_j = \mathbf{t}\} f_j(\mathbf{t}). \tag{A.4}$$

Recall that $f_{j\ell}(\mathbf{t}, \mathbf{s})$ denotes the joint density of $(\mathbf{T}_j, \mathbf{T}_\ell)$ evaluated at $(\mathbf{t}, \mathbf{s})$. Define

$$Q(\mathbf{t}, \mathbf{s}) = \sum_j \sum_{\ell \neq j} E\big[\Delta_{jj} v^{j\ell} \Delta_{\ell\ell} \{W_2(\mathbf{T}_\ell)\}^{-1} \big| \mathbf{T}_j = \mathbf{t}, \mathbf{T}_\ell = \mathbf{s}\big] f_{j\ell}(\mathbf{t}, \mathbf{s}) \tag{A.5}$$

and

$$b_{[k]}(\mathbf{t}) = b_{[0]}(\mathbf{t}) - W_2^{-1}(\mathbf{t}) \times \sum_j \sum_{\ell \neq j} E\big\{\Delta_{jj} v^{j\ell} \Delta_{\ell\ell} b_{[k-1]}(\mathbf{T}_\ell) \big| \mathbf{T}_j = \mathbf{t}\big\} f_j(\mathbf{t}), \tag{A.6}$$

with $b_{[0]}(\mathbf{t})$ defined in (A.4). It can be shown that the estimated $\theta(\cdot)$ after one step update of $\widehat{\theta}_{[0]}$ has the expansion

$$\begin{aligned}
&\widehat{\theta}_{[1]}(t) - \theta(t) \\
&\quad = \frac{1}{2} b_{[1]}(t) h^2 \\
&\qquad + W_2^{-1}(t) n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij}^{(1)} K_h(T_{ij} - t) \left\{ \sum_{\ell=1}^{m} v_i^{jl} (Y_{i\ell} - \mu_{i\ell}) \right\} \\
&\qquad - W_2^{-1}(t) n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij}^{(1)} v_i^{jj} Q(t, T_{ij}) (Y_{ij} - \mu_{ij}) \\
&\qquad + o_p\big(h^2 + \{\log(n)/nh\}^{1/2} + n^{-1/2}\big).
\end{aligned}$$

Further, define an integration operator $\ddot{\mathcal{A}}\{B(\cdot, \cdot); \mathbf{t}, \mathbf{s}\}$,

$$\ddot{\mathcal{A}}(B; \mathbf{t}, \mathbf{s}) = -\sum_j \sum_{\ell \neq j} E\big[\Delta_{jj} v^{j\ell} \Delta_{\ell\ell} \{W_2(\mathbf{T}_\ell)\}^{-1} B(\mathbf{T}_\ell, \mathbf{s}) \big| \mathbf{T}_j = \mathbf{t}\big] f_j(\mathbf{t}). \tag{A.7}$$

For iteration $k \geq 2$, we have

$$\begin{aligned}
&\widehat{\theta}_{[k]}(t, \boldsymbol{\beta}_0) - \theta(t) \\
&\quad = \frac{1}{2} b_{[k]}(t) h^2 \\
&\qquad + W_2^{-1}(t) n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij}^{(1)} K_h(T_{ij} - t) \left\{ \sum_{\ell=1}^{m} v_i^{jl} (Y_{i\ell} - \mu_{i\ell}) \right\} \\
&\qquad + W_2^{-1}(t) n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij}^{(1)} Q_{1,[k]}(t, T_{ij}) \left\{ \sum_{\ell=1}^{m} v_i^{jl} (Y_{i\ell} - \mu_{i\ell}) \right\} \\
&\qquad + W_2^{-1}(t) n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij}^{(1)} v_i^{jj} Q_{2,[k]}(t, T_{ij}) (Y_{ij} - \mu_{ij}) \\
&\qquad + o_p\big(h^2 + \{\log(n)/nh\}^{1/2} + n^{-1/2}\big),
\end{aligned} \tag{A.8}$$

where $b_{[k]}(\mathbf{t})$ is defined in (A.6), $Q_{1,[1]}(\mathbf{t}, \mathbf{s}) \equiv 0$, $Q_{2,[1]}(\mathbf{t}, \mathbf{s}) = -Q(\mathbf{t}, \mathbf{s})$,

$$Q_{1,[k]}(\mathbf{t}, \mathbf{s}) = -Q(\mathbf{t}, \mathbf{s}) + \ddot{\mathcal{A}}\big(Q_{1,[k-1]}; \mathbf{t}, \mathbf{s}\big)$$

and

$$Q_{2,[k]}(\mathbf{t}, \mathbf{s}) = \ddot{\mathcal{A}}\big(Q_{2,[k-1]}; \mathbf{t}, \mathbf{s}\big).$$

At convergence, $\widehat{\theta}_*(\mathbf{t}) - \theta(\mathbf{t})$ shares the same asymptotic structure as in (A.8) except that $b_{[k]}$, $Q_{1,[k]}$, and $Q_{2,[k]}$ are replaced by b_*, $Q_{1,*}$, and $Q_{2,*}$, where $\ddot{\mathcal{A}}$ is given in (A.7), and b_*, $Q_{1,*}$, and $Q_{2,*}$ satisfy the corresponding integration equations,

$$b_*(\mathbf{t}) = \theta^{(2)}(\mathbf{t}) - W_2^{-1}(\mathbf{t}) \sum_j \sum_{\ell \neq j} E\{\Delta_{jj} v^{j\ell} \Delta_{\ell\ell} b_*(\mathbf{T}_\ell) | \mathbf{T}_j = \mathbf{t}\} f_j(\mathbf{t}),$$

$$Q_{1,*}(\mathbf{t}, \mathbf{s}) = -Q(\mathbf{t}, \mathbf{s}) + \ddot{\mathcal{A}}(Q_{1,*}; \mathbf{t}, \mathbf{s}),$$

and

$$Q_{2,*}(\mathbf{t}, \mathbf{s}) = \ddot{\mathcal{A}}(Q_{2,*}; \mathbf{t}, \mathbf{s}).$$

A.3 Proof of Proposition 1

In this section we sketch a proof to show that in the Gaussian case assuming a linear link, $\boldsymbol{\varphi} = \boldsymbol{\varphi}_{\text{eff}}$, the latter given by (6) with $\boldsymbol{\Delta}$ the identity matrix. For the proof of the general case, simply place elements in $\boldsymbol{\Delta}$ (the first derivative of $\boldsymbol{\mu}$) at the right places. At convergence, using (3), it can be shown that for a given $\boldsymbol{\beta}$, we have

$$\begin{aligned}
0 = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} & K_h(t - T_{ij}) \\
\times \Bigg(& \sigma_i^{jj} \big[Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta} - \widehat{\theta}(t, \boldsymbol{\beta}) - \{(T_{ij} - t)/h\} \widehat{\alpha}_1(t, \boldsymbol{\beta})\big] \\
& + \sum_{l \neq j} \sigma_i^{jl} \{Y_{il} - \mathbf{X}_{il}\boldsymbol{\beta} - \widehat{\theta}(T_{il}, \boldsymbol{\beta})\} \Bigg).
\end{aligned}$$

Taking derivatives with respect to $\boldsymbol{\beta}$ on both sides, direct derivations lead to

$$\begin{aligned}
\mathbf{0} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(t - T_{ij}) \Bigg[& \sigma_i^{jj} \widehat{\boldsymbol{\varphi}}(t) - \sigma_i^{jj} \{(T_{ij} - t)/h\} \, \partial \widehat{\alpha}_1(\mathbf{t}, \boldsymbol{\beta})/\partial \boldsymbol{\beta} \\
& - \sum_{\ell} \sigma_i^{j\ell} \mathbf{X}_{il} + \sum_{\ell \neq j} \sigma_i^{j\ell} \widehat{\boldsymbol{\varphi}}(T_{i\ell}) \Bigg].
\end{aligned}$$

It is straightforward to show that $n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(t - T_{ij}) \times \sigma_i^{jj} \{(T_{ij} - t)/h\} = o_p(1)$,

$$n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \sigma_i^{jj} K_h(t - T_{ij}) = \left\{ \sum_{j=1}^{m} E(\sigma^{jj} | \mathbf{T}_j = \mathbf{t}) f_j(\mathbf{t}) \right\} \{1 + o_p(1)\}, \tag{A.9}$$

and

$$n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(t - T_{ij}) \sum_{\ell=1}^{m} \sigma_i^{j\ell} X_{i\ell} = \left\{ \sum_{j=1}^{m} \sum_{\ell=1}^{m} E(\sigma^{j\ell} \mathbf{X}_\ell | \mathbf{T}_j = \mathbf{t}) f_j(\mathbf{t}) \right\} \{1 + o_p(1)\}. \tag{A.10}$$

In addition, we have that

$$\begin{aligned}
& n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(t - T_{ij}) \left\{ \sum_{\ell \neq j} \sigma^{j\ell} \widehat{\boldsymbol{\varphi}}(T_{i\ell}) \right\} \\
&\quad = \sum_{j=1}^{m} \sum_{\ell \neq j}^{m} \int E(\sigma^{j\ell} | \mathbf{T}_j = \mathbf{t}) \widehat{\boldsymbol{\varphi}}(t_\ell) f_{\ell j}(t_\ell, t) \, dt_\ell \, \{1 + o_p(1)\}.
\end{aligned} \tag{A.11}$$

The combination of (A.9), (A.10), and (A.11) leads to

$$\begin{aligned}&\sum_{j=1}^{m} E(\sigma^{jj}|\mathbf{T}_j=\mathbf{t})f_j(\mathbf{t})\widehat{\boldsymbol{\varphi}}(\mathbf{t}) - \sum_{j=1}^{m}\sum_{\ell=1}^{m} E(\sigma^{j\ell}\mathbf{X}_\ell|\mathbf{T}_j=\mathbf{t})f_j(\mathbf{t}) \\ &\qquad - \sum_{j=1}^{m}\sum_{\ell\neq j}^{m}\int E(\sigma^{j\ell}|\mathbf{T}_j=\mathbf{t})\widehat{\boldsymbol{\varphi}}(t_\ell)f_{\ell j}(t_\ell,\mathbf{t})\,dt_\ell = o_p(1), \qquad \text{(A.12)}\end{aligned}$$

uniformly on $\mathbf{t}$. Dividing (A.12) by $\sum_j E(\sigma^{jj}|\mathbf{T}_j=\mathbf{t})f_j(\mathbf{t})$, noting that $\widehat{\boldsymbol{\varphi}}(\mathbf{t})$ uniformly converges to $\boldsymbol{\varphi}(\mathbf{t})$, and comparing the second and third terms to (7) and (8) with elements in $\boldsymbol{\Delta}\equiv\mathbf{1}$, we can directly establish Proposition 1 by letting $\widehat{\boldsymbol{\varphi}}(\mathbf{t})$ converges to $\boldsymbol{\varphi}(\mathbf{t})$ in (A.12) and noting that $\boldsymbol{\varphi}(\mathbf{t})$ fulfills (6).

A.4 Proof of (9)

In this section we derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$. We first construct the following lemma, which is a consequence of Proposition 1.

Lemma A.1. For any function $\mathbf{A}(\cdot)$,

$$\sum_j\sum_k E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{jk}\Delta_{kk}\mathbf{A}(\mathbf{T}_k)|\mathbf{T}_k=\mathbf{t}\}f_k(\mathbf{t})=0. \qquad \text{(A.13)}$$

Furthermore,

$$\sum_j\sum_k E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{jk}\Delta_{kk}\mathbf{A}(\mathbf{T}_k)\}=0. \qquad \text{(A.14)}$$

Proof. We rewrite (5) by

$$\sum_j\sum_k E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{jk}\Delta_{kk}|\mathbf{T}_k=\mathbf{t}\}f_k(\mathbf{t})=0. \qquad \text{(A.15)}$$

Equation (A.13) is established by multiplying both sides of (A.15) by $\mathbf{A}(\mathbf{t})$ and noting that

$$E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{jk}\Delta_{kk}\mathbf{A}(\mathbf{T}_k)|\mathbf{T}_k=\mathbf{t}\}=E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{jk}\Delta_{kk}|\mathbf{T}_k=\mathbf{t}\}\mathbf{A}(\mathbf{t}).$$

Equation (A.14) follows directly from (A.13).

To prove (9), following the same derivations as used by LC and keeping only the essential terms, some tedious calculations lead to the following asymptotic expansion for the profile estimator $\widehat{\boldsymbol{\beta}}$:

$$n^{1/2}\{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\}=\{\widetilde{\mathbf{A}}(\mathbf{V})\}^{-1}\{B_n+C_{1n}-C_{2n}\}+o_p(1),$$

where $\widetilde{\mathbf{A}}(\mathbf{V})$ is defined in Proposition 2,

$$\begin{aligned}B_n &= 1/2(nh^4)^{1/2} \\ &\quad\times\Bigg[n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{m}\mu_{ij}^{(1)}v_i^{j\ell}\mu_{i\ell}^{(1)}\widetilde{\mathbf{X}}_{ij} \\ &\qquad\times\{b_*(T_{i\ell})+hb_{*1}(T_{i\ell})+O_p(h^2)\}\Bigg]\{1+o_p(1)\},\end{aligned}$$

$$C_{1n}=n^{-1/2}\sum_{i=1}^{n}\underline{\widetilde{\mathbf{X}}}_i^t\boldsymbol{\Delta}_i(\mathbf{V}_i)^{-1}(\underline{\mathbf{Y}}_i-\underline{\boldsymbol{\mu}}_i),$$

and

$$\begin{aligned}C_{2n} &= \Bigg\{n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_i^{j\ell}\mu_{i\ell}^{(1)} \\ &\quad\times\Bigg(W_2^{-1}(T_{i\ell})n^{-1}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}(v_{i'})^{j'j'} \\ &\quad\times\Bigg[K_h(T_{i'j'}-T_{i\ell})\Bigg\{\sum_l(v_{i'})^{j'l}(Y_{i'l}-\mu_{i'l})\Bigg\} \\ &\qquad+Q_{2,*}(T_{i\ell},T_{i'j'})(Y_{i'j'}-\mu_{i'j'}) \\ &\qquad+Q_{1,*}(T_{i\ell},T_{i'j'})\Bigg\{\sum_l(v_{i'})^{j'l}(Y_{i'l}-\mu_{i'l})\Bigg\}\Bigg]\Bigg)\Bigg\}\{1+o_p(1)\},\end{aligned}$$

with b_*, $Q_{1,*}$, and $Q_{2,*}$ as defined in Appendix A.2 and b_{*1} as the next order term in a higher-order bias expansion of $\widehat{\theta}$ following the equivalent derivations in theorem 1 of Fan, Gijbels, Hu, and Huang (1995). In general, for an estimator of θ (e.g., the ZD estimator), the terms inside the square bracket of B_n is of order $O_p(1)$. Thus one needs nh^4 goes to 0 to eliminate the bias term in B_n. For the proposed estimator, as a consequence of (A.14), the first and second terms inside the square bracket of B_n are of order $o_p(1)$; thus $B_n=o_p(1)$, provided that nh^8 goes to 0.

We now proceed to show that C_{2n} is of order $o_p(1)$. Write $C_{2n}=\sum_{\ell=1}^{3}C_{2\ell n}$, where

$$\begin{aligned}C_{21n} &= \frac{1}{\sqrt{n}}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}(v_{i'})^{j'j'} \\ &\quad\times\Bigg\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{m}K_h(T_{i'j'}-T_{i\ell})\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_i^{j\ell}\mu_{i\ell}^{(1)}W_2^{-1}(T_{i\ell})\Bigg\} \\ &\quad\times\Bigg\{\sum_l(v_{i'})^{j'l}(Y_{i'l}-\mu_{i'l})\Bigg\}\{1+o_p(1)\},\end{aligned}$$

where the term inside the first $\{\cdot\}$ is asymptotically equivalent to

$$\sum_j\sum_\ell E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{j\ell}\Delta_{\ell\ell}W_2^{-1}(\mathbf{T}_\ell)|\mathbf{T}_\ell=\mathbf{t}\}f_\ell(\mathbf{t})\Big|_{t=T_{i'j'}}. \qquad \text{(A.16)}$$

Similarly, we have

$$\begin{aligned}C_{22n} &= \frac{1}{\sqrt{n}}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}(v_{i'})^{j'j'} \\ &\quad\times[E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{j\ell}\Delta_{\ell\ell}W_2^{-1}(\mathbf{T}_\ell)Q_{2,*}(\mathbf{T}_\ell,\mathbf{t})\}|t=T_{i'j'}] \\ &\quad\times(Y_{i'j'}-\mu_{i'j'})\{1+o_p(1)\}\end{aligned}$$

and

$$\begin{aligned}C_{23n} &= \frac{1}{\sqrt{n}}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}(v_{i'})^{j'j'} \\ &\quad\times[E\{\widetilde{\mathbf{X}}_j\Delta_{jj}v^{j\ell}\Delta_{\ell\ell}W_2^{-1}(\mathbf{T}_\ell)Q_{1,*}(\mathbf{T}_\ell,\mathbf{t})\}|t=T_{i'j'}] \\ &\quad\times\Bigg\{\sum_l(v_{i'})^{j'l}(Y_{i'l}-\mu_{i'l})\Bigg\}\{1+o_p(1)\}.\end{aligned}$$

By Lemma A.1, we observe that the expectation terms in (A.16), C_{22n} and C_{23n} all equal 0. They are also only functions of $\mathbf{T}$ and $\mathbf{X}$. Thus, with $h\to 0$ and $n\to\infty$ at the rates that $nh^8\to 0$ and $nh/\log(1/h)\to\infty$, we obtain that B_n, C_{21n}, and C_{22n} are all of order $o_p(1)$.

It follows that

$$n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\{\widetilde{\mathbf{A}}(\mathbf{V})\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\widetilde{\mathbf{X}}}_i^t\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}(\underline{\mathbf{Y}}_i-\underline{\boldsymbol{\mu}}_i)\{1+o_p(1)\}, \qquad \text{(A.17)}$$

which implies that $n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\to \text{normal}\{\mathbf{0},\boldsymbol{\Omega}(\mathbf{V},\boldsymbol{\Sigma})\}$, where $\boldsymbol{\Omega}(\mathbf{V},\boldsymbol{\Sigma})$ is defined in Proposition 2. An application of an extended Cauchy–Schwartz inequality given by Johnson and Wichern (1982, sec. 2.7) indicates that the best choice of $\mathbf{V}$ is $\boldsymbol{\Sigma}$, which gives $\boldsymbol{\Omega}(\mathbf{V},\boldsymbol{\Sigma})=\widetilde{\mathbf{A}}^{-1}(\boldsymbol{\Sigma})$.

A.5 Derivation of (10) and Structure of $\underline{\mathbf{Z}}_1$ and $\underline{\mathbf{Z}}_2$

Under the normality assumption of $\underline{\mathbf{X}}$ and $\underline{\mathbf{T}}$, $E(\mathbf{X}_\ell|\mathbf{T}_j = \mathbf{t}) = (\sigma_{\mathbf{X}}\rho_{\mathbf{XT}}/\sigma_{\mathbf{T}})\mathbf{t}$. This equation and the structure of (6) imply that $\boldsymbol{\varphi}_{\mathrm{eff}}(\mathbf{t}) = c(\sigma_{\mathbf{X}}/\sigma_{\mathbf{T}})\mathbf{t}$, where c is some constant. Plugging this into (6), it is straightforward to solve for c and obtain (10). Similar calculations can be used to obtain $I_{\mathrm{WI}}(\beta)$ and $I_{\mathrm{ZD}}(\beta)$ given in Section 5, where using result 1 of LC, $\underline{\mathbf{Z}}_1$ and $\underline{\mathbf{Z}}_2$ in $I_{\mathrm{ZD}}(\beta)$ are $\underline{\mathbf{Z}}_1 = \boldsymbol{\Sigma}^{-1}\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI}}$ and the kth row of $\underline{\mathbf{Z}}_2$ is

$$\frac{\sigma_{kk}^{-1}\sum_{j=1}^{m}\sum_{\ell=1}^{m}\sigma^{j\ell}E(\widetilde{\underline{\mathbf{X}}}_{\mathrm{WI},j}|\mathbf{T}_\ell = \mathbf{T}_k)}{\sum_{j=1}^{m}\sigma_{jj}^{-1}}.$$

[Received May 2002. Revised March 2004.]

REFERENCES

Begun, J. H., Hall, W. J., Huang, W. M., and Wellner, J. A. (1983), "Information and Asymptotic Efficiency in Parametric–Nonparametric Models," *The Annals of Statistics*, 11, 432–452.

Bickel, P. J., Klaassen, A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models*, Baltimore, MD: Johns Hopkins University Press.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Carroll, R. J., Wu, C. F. J., and Ruppert, D. (1988), "The Effect of Estimating Weights in Linear Regression," *Journal of the American Statistical Association*, 83, 1045–1054.

Crowder, M. (2001), "On Repeated Measures Analysis With Misspecified Covariance Structure," *Journal of the Royal Statistical Society*, Ser. B, 63, 55–62.

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

Fan, J., Gijbels, I., Hu, T. C., and Huang, L. S. (1995), "A Study of Variable Bandwidth Selection for Local Polynomial Regression," *Statistica Sinica*, 6, 113–127.

Heagerty, P. J., and Zeger, S. L. (2000), "Marginalized Multilevel Models and Likelihood Inference," *Statistical Science*, 15, 1–26.

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

Johnson, R. A., and Wichern, D. W. (1982), *Applied Multivariate Statistical Analysis*, New York: Prentice-Hall.

Kress, R. (1989), *Linear Integral Equations,* Berlin: Springer-Verlag.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, X., and Carroll, R. J. (2001a), "Semiparametric Regression for Clustered Data Using Generalized Estimating Equations," *Journal of the American Statistical Association*, 96, 1045–1056.

——— (2001b), "Semiparametric Regression for Clustered Data," *Biometrika*, 88, 1179–1865.

Lin, X., Wang, N., Welsh, A., and Carroll, R. J. (2004), "Equivalent Kernels of Smoothing Splines in Nonparametric Regression for Clustered Data," *Biometrika*, 91, 177–193.

Lin, D. Y., and Ying, Z. (2001), "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data" (with discussion), *Journal of the American Statistical Association*, 96, 103–126.

Mack, Y., and Silverman, B. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, 405–415.

Pepe, M. S., and Couper, D. (1997), "Modeling Partly Conditional Means With Longitudinal Data," *Journal of the American Statistical Association*, 92, 991–998.

Prentice, R., and Zhao, L. P. (1991), "Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses," *Biometrics*, 47, 825–839.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049–1062.

Ruppert, D., and Wand, M. P. (1994), "Multivariate Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Wang, N. (2003), "Marginal Nonparametric Kernel Regression Accounting for Within-Subject Correlation," *Biometrika*, 90, 43–52.

Wild, C. J., and Yee, T. W. (1996). "Additive Extensions to Generalized Estimating Equation Methods," *Journal of the Royal Statistical Society*, Ser. B, 58, 711–725.

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