# Semiparametric Regression for Clustered Data Using Generalized Estimating Equations

Xihong LIN and Raymond J. CARROLL

We consider estimation in a semiparametric generalized linear model for clustered data using estimating equations. Our results apply to the case where the number of observations per cluster is finite, whereas the number of clusters is large. The mean of the outcome variable $\mu$ is of the form $g(\mu) = \mathbf{X}^T\boldsymbol{\beta} + \theta(T)$, where $g(\cdot)$ is a link function, $\mathbf{X}$ and $T$ are covariates, $\boldsymbol{\beta}$ is an unknown parameter vector, and $\theta(t)$ is an unknown smooth function. Kernel estimating equations proposed previously in the literature are used to estimate the infinite-dimensional nonparametric function $\theta(t)$, and a profile-based estimating equation is used to estimate the finite-dimensional parameter vector $\boldsymbol{\beta}$. We show that for clustered data, this conventional profile-kernel method often fails to yield a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ along with appropriate inference unless working independence is assumed or $\theta(t)$ is artificially undersmoothed, in which case asymptotic inference is possible. To gain insight into these results, we derive the semiparametric efficient score of $\boldsymbol{\beta}$, which is found to have a complicated form, and show that, unlike for independent data, the profile-kernel method does not yield a score function asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$, even when the true correlation is assumed and $\theta(t)$ is undersmoothed. We illustrate the methods with an application to infectious disease data and evaluate their finite-sample performance through a simulation study.

KEY WORDS: Asymptotics; Clustered data; Consistency; Efficiency; Generalized estimating equations; Kernel method; Longitudinal data; Nonparametric regression; Partially linear model; Profile method; Sandwich estimator; Semiparametric efficient score; Semiparametric efficiency bound.

## 1. INTRODUCTION

Clustered data arise in many fields of biomedical research, including longitudinal studies, intervention studies, and clinical trials. Parametric regression using generalized estimating equations (GEEs) (Liang and Zeger 1986) has become a popular practice for analyzing such data. It is well understood that the GEE estimators of regression coefficients are consistent when the mean function is correctly specified even when the within-cluster correlation structure is misspecified, and that the most efficient estimator is obtained by correctly specifying the within-cluster correlation. To allow for more flexible dependence of an outcome variable on covariates, there has been substantial recent interest in modeling covariate effects nonparametrically (Lin and Carroll 2000, Hoover, Rice, Wu, and Yang 1998; Wild and Yee 1996). Lin and Carroll (2000) showed that in contrast to parametric GEEs, when standard kernel methods are used, typically the most efficient estimator of the nonparametric function is obtained by completely ignoring the within-cluster correlation; correct specification of the correlation structure generally results in an asymptotically less efficient estimator.

In many instances, a semiparametric partially generalized linear regression model is more desirable than modeling every covariate effect nonparametrically. This model assumes that the mean of the outcome variable $\mu$ depends on some covariates $\mathbf{X}$ parametrically and on some other covariate $T$ nonparametrically in the form $g(\mu) = \mathbf{X}^T\boldsymbol{\beta} + \theta(T)$, where $g(\cdot)$ is a link function, $\boldsymbol{\beta}$ is an unknown parameter vector, and $\theta(\cdot)$ is an unknown smooth function. This model specification is particularly appealing when the effects of $\mathbf{X}$ (e.g., treatment) are of major interest and the effects of $T$ (e.g., confounders) are nuisance. This is because one can make inference on the effects of $\mathbf{X}$ while making minimal assumptions on the effects of $T$ using a fully nonparametric function.

One example is the longitudinal infectious disease study considered in Section 8. This study involved 275 preschool-age children who were reexamined every 3 months for 18 months for the presence of respiratory infection (yes/no) (Diggle, Liang, and Zeger 1994). The primary interest is to study the association between respiratory infection and vitamin A deficiency (yes/no), while accounting for several confounders including age. Examination of the distribution of the vertical strokes in Figure 3 suggests that the age effect departs dramatically from linearity; the vertical strokes indicate the ages for yes (top) and no (bottom).

Because the binary exposure of vitamin A deficiency is of main interest and the age effect is nuisance, we are interested in modeling the vitamin A deficiency effect while allowing the nuisance age effect to be modeled nonparametrically.

Several authors have considered such semiparametric regression models. A key challenge of estimation in this model is that it is composed of a finite-dimensional parameter vector $\boldsymbol{\beta}$ and an infinite dimensional parameter $\theta(\cdot)$. Estimation for independent nonclustered data has been considered by Carroll, Fan, Gijbels, and Wand (1997), Hastie and Tibshirani (1990), and Severini and Staniswalis (1994). These authors used the kernel method to estimate $\theta(t)$ and the profile likelihood–based method to estimate $\boldsymbol{\beta}$. They showed that the estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent and semiparametric efficient (Bickel, Klaassen, Ritov, and Wellner 1993). For longitudinal data, Zeger and Diggle (1994) considered a semiparametric model with a nonparametric time trajectory and parametric covariate effects. They estimated $\theta(t)$ using a kernel method by ignoring the within-cluster correlation, and estimated $\boldsymbol{\beta}$ using weighted least squares by accounting for the within-cluster

Xihong Lin is Associate Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *xlin@sph.umich.edu*). Her research was supported by National Cancer Institute grant CA–76404. Raymond J. Carroll is Distinguished Professor, Departments of Statistics and Biostatistics and Epidemiology, Texas A&M University, College Station TX 77843 (E-mail: *carroll@stat.tamu.edu*). His research was supported by National Cancer Institute grant CA–57030, and by the Texas A&M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences grant P30-ES09106. The authors thank the editor, the associate editor, and two referees for their helpful comments and suggestions.

correlation. They did not study the asymptotic properties of their method. Severini and Staniswalis (1994) extended their independent data results to clustered data using profile-kernel GEEs. They claimed that the estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent for any working correlation matrix specification. Zhang, Lin Raz, and Sowers (1998) considered a semiparametric linear mixed model and estimated the nonparametric function using a smoothing spline.

In this article we consider a marginal semiparametric regression model for clustered data with $\theta(t)$ estimated using kernel estimating equations and $\boldsymbol{\beta}$ estimated using profile-based estimating equations. Our estimating equations are similar to those of Severini and Staniswalis (1994) except that different working correlation matrices are allowed in the two sets of estimating equations, and local linear regression is used instead of local average kernel regression. The main focus of this article is to investigate whether it is possible to construct a $\sqrt{n}$-consistent and efficient estimator of $\boldsymbol{\beta}$ using the profile-kernel method. This work is motivated by our observation of the diametrically opposed asymptotic properties of parametric and certain nonparametric GEEs in terms of how to obtain the most efficient estimators, the former requiring correctly specifying the correlation and the latter requiring completely ignoring the correlation. Hence we are interested in investigating whether such different asymptotic behavior affects consistency and efficiency of the estimator of $\boldsymbol{\beta}$ in the semiparametric model using the conventional profile-kernel method. In particular, does correct specification of the within-cluster correlation still yield a $\sqrt{n}$-consistent and semiparametric efficient estimator of $\boldsymbol{\beta}$?

The results that we have obtained are surprising. To obtain a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ using the conventional profile-kernel method, one generally must either artificially undersmooth $\theta(t)$ or completely ignore the within-cluster correlation by assuming working independence in the profile-kernel estimating equations. Thus, if one accounts for within-cluster correlation using the profile-kernel method, then the standard bandwidth selection methods used for estimating $\theta(t)$, such as cross-validation, fail, the sandwich covariance estimator of the estimator of $\boldsymbol{\beta}$ fails, and the conventional hypothesis tests on $\boldsymbol{\beta}$ such as the Wald and Score tests fail. With undersmoothing or working independence, asymptotically correct inference about $\boldsymbol{\beta}$ becomes possible. To gain insight into these results, we derive the semiparametric efficient score of $\boldsymbol{\beta}$, which is found to have a complicated form, and show that unlike for independent data, the profile-kernel method does not yield a score function that is asymptotically equivalent to the semiparametric efficient score for $\boldsymbol{\beta}$, even when the true correlation is assumed and $\theta(t)$ is undersmoothed. Our main conclusion is that, unlike for independent data, the conventional profile-kernel method is not semiparametric efficient and must be modified in ad hoc ways (undersmoothing) or to be made less efficient (working independence) to even be made $\sqrt{n}$ consistent.

The article is organized as follows. In Section 2 we state the semiparametric model for clustered data and in Section 3 discuss estimation of $\theta(t)$ using kernel estimating equations previously proposed in the literature and of $\boldsymbol{\beta}$ using profile estimating equations. In Section 4 we study the asymptotic properties of the profile-kernel estimators of $\boldsymbol{\beta}$ and $\theta(t)$. In Section 5 we derive the semiparametric efficient score of $\boldsymbol{\beta}$ within a likelihood framework, and show that the conventional profile-kernel estimating equations of $\boldsymbol{\beta}$ often do not yield a score equation that is asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$. In Section 6 we discuss practical implications of our results. We illustrate the methods with a simulation study in Section 7 and an application to infectious disease data in Section 8. We conclude with a discussion is Section 9.

## 2. A SEMIPARAMETRIC MARGINAL MODEL

In this section we present the semiparametric regression model for clustered data. Suppose that the data consist of $n$ clusters with the $i$th ($i = 1, \ldots, n$) cluster having $m_i$ observations. Let $Y_{ij}$ and $(\mathbf{X}_{ij}, T_{ij})$ be the response variable and the covariates of the $j$th ($j = 1, \ldots, m_i$) observation in the $i$th cluster, where $\mathbf{X}_{ij}$ is a $p \times 1$ vector and $T_{ij}$ is a scalar. Given the covariates $\mathbf{X}_{ij}$ and $T_{ij}$, the mean and the variance of the outcome variable $Y_{ij}$ are $E(Y_{ij}) = \mu_{ij}$ and $\text{var}(Y_{ij}) = \phi w_{ij}^{-1} V(\mu_{ij})$, where $\phi$ is a scale parameter, $w_{ij}$ is a known weight, and $V(\cdot)$ is a known variance function. The marginal mean $\mu_{ij}$ depends on $\mathbf{X}_{ij}$ and $T_{ij}$ through a known monotonic and differentiable link function $g(\cdot)$,

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \theta(T_{ij}), \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector and $\theta(\cdot)$ is an unknown smooth function. We model the effects of $\mathbf{X}$ ($p \times 1$) parametrically and the effects of $T$ nonparametrically, and treat the within-cluster correlation parameters as nuisance parameters. In particular, it is important to note the assumption (Pepe and Couper 1997) that

$$E(Y_{ij} | \mathbf{X}_{ij}, T_{ij}) = E\{Y_{ij} | \mathbf{X}_{ij}, T_{ij}, (\mathbf{X}_{ij}, T_{ik})_{k \neq j}\}, \tag{2}$$

an assumption also made implicitly by Lin and Carroll (2000). In matrix notation, denoting by $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^T$, $g(\boldsymbol{\mu}_i) = \{g(\mu_{i1}), \ldots, g(\mu_{im_i})\}^T$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$, and $\mathbf{X}_i$, and $\boldsymbol{T}_i$ similarly, we have $g(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\theta}(\boldsymbol{T}_i)$. If model (1) does not include $\theta(T_{ij})$, then it reduces to the parametric generalized linear model considered by Liang and Zeger (1986). If model (1) does not include $\mathbf{X}_{ij}^T \boldsymbol{\beta}$, then it reduces to the nonparametric model considered by Lin and Carroll (2000). Severini and Staniswalis (1994) considered a model similar to (1)–(2).

It is important to emphasize that we are considering a marginal model for the clustered data through specification of mean and variance functions. This is in the spirit of GEE-type models (Liang and Zeger 1986). Except for Gaussian data, our marginal models need not be a full semiparametric likelihood specification.

## 3. PROFILE-KERNEL ESTIMATING EQUATIONS

In this section we develop kernel estimating equations for $\theta(t)$ and profile estimating equations for $\boldsymbol{\beta}$. The formulation of the profile estimating equation is similar to the score equation calculated using the conventional profile likelihood approach in parametric regression. We give the motivation of these estimating equations in Section 3.1, and describe their forms in Section 3.2.

### 3.1 Motivation of the Profile-Kernel Estimating Equations

To motivate the profile-kernel estimating equations for $\boldsymbol{\beta}$ and $\theta(t)$ under the semiparametric model (1), we first consider the GEEs for the parametric model

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T\boldsymbol{\beta}. \tag{3}$$

Of course, (3) is a special case of (1) when $\theta(t) = 0$. Liang and Zeger (1986) proposed estimating $\boldsymbol{\beta}$ using the estimating equations

$$\sum_{i=1}^{n} \frac{\partial \mu(\mathbf{X}_i\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \sum_{i=1}^{n} \mathbf{X}_i^T \boldsymbol{\Delta}_i \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{4}$$

where $\boldsymbol{\mu}_i = E(\mathbf{Y}_i) = \mu(\mathbf{X}_i\boldsymbol{\beta})$ with the $j$th component $\mu_{ij} = \mu(\mathbf{X}_{ij}^T\boldsymbol{\beta}) = g^{-1}(\mathbf{X}_{ij}^T\boldsymbol{\beta})$, $\boldsymbol{\Delta}_i = \operatorname{diag}\{\mu_{ij}^{(1)}\}$, $\mu^{(1)}(\cdot)$ is the first derivative of $\mu(\cdot)$, $\mathbf{V}_i = \mathbf{S}_i^{1/2}\mathbf{R}_i(\boldsymbol{\tau})\mathbf{S}_i^{1/2}$, $\mathbf{S}_i = \operatorname{diag}[\phi w_{ij}^{-1} V\{\mu_{ij}\}]$ contains the marginal variances of the $Y_{ij}$, and $\mathbf{R}_i$ is an invertible working correlation matrix, possibly depending a parameter vector $\boldsymbol{\tau}$, which can be estimated using the method of moments. Liang and Zeger (1986) showed that the GEE estimator $\hat{\boldsymbol{\beta}}$ is asymptotically consistent if the mean function $\mu_{ij}$ is correctly specified even when the working correlation matrix $\mathbf{R}_i$ is misspecified. The efficient kernal estimator of $\boldsymbol{\beta}$ is obtained by specifying $\mathbf{R}_i$ as the true correlation matrix.

Now consider kernel estimating equations for the nonparametric model

$$g(\mu_{ij}) = \theta(T_{ij}). \tag{5}$$

Lin and Carroll (2000) considered the $p$th local polynomial kernel estimating equations for $\theta(t)$. We consider here the local linear kernel estimator, that is, $p = 1$. Let $h$ denote the bandwidth parameter, and let $K(\cdot)$ denote the symmetric kernel density function. Let $K_h(v) = h^{-1}K(v/h)$ and $\boldsymbol{T}_i(t)$ be an $m_i \times 2$ matrix with the $j$th row $\{1, (T_{ij} - t)/h\}$. Lin and Carroll (2000) considered two kernel (symmetric and asymmetric) estimating equations for $\theta(t)$ at any $t$,

$$\sum_{i=1}^{n} \mathbf{T}_i(t)^T \boldsymbol{\Delta}_i(t)\mathbf{K}_{ih}^{1/2}(t)\mathbf{V}_i^{-1}(t)\mathbf{K}_{ih}^{1/2}(t)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(t)\} = 0 \tag{6}$$

and

$$\sum_{i=1}^{n} \mathbf{T}_i(t)^T \boldsymbol{\Delta}_i(t)\mathbf{V}_i^{-1}(t)\mathbf{K}_{ih}(t)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(t)\} = 0, \tag{7}$$

where $\mathbf{K}_{ih}(t) = \operatorname{diag}\{K_h(T_{ij} - t)\}$ and $[\boldsymbol{\mu}_i(t), \boldsymbol{\Delta}_i(t), \mathbf{V}_i(t), \mathbf{S}_i(t)\}$ are the same as those defined in (4) except that they are evaluated at $\mu_{ij}(t) = \mu\{\alpha_0 + \alpha_1(T_{ij} - t)/h\}$, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$ is a $2 \times 1$ vector of unknown parameters. Equation (7) was also considered by Severini and Staniswalis (1994) using the local average kernel ($p = 0$). Having estimated $\boldsymbol{\alpha}$ at $t$ as $\hat{\boldsymbol{\alpha}}$, the kernel estimator of $\theta(t)$ is $\hat{\theta}(t) = \hat{\alpha}_0$. The working correlation matrix $\mathbf{R}_i$ in $\mathbf{V}_i(t)$ may again depend on a parameter vector $\boldsymbol{\tau}$, which again can be estimated using the method of moments. The kernel estimators under (6) and (7) are different except when working independence is assumed; that is, $\mathbf{R}_i = \mathbf{I}$.

Lin and Carroll (2000) showed that the two estimators under (6) and (7) have different asymptotic properties; asymptotic properties of the kernel estimator under (7) are much harder to study. The most important results of Lin and Carroll (2000) are that, unlike the parametric GEE estimator in (4), typically the asymptotically most efficient kernel estimator of the nonparametric function $\hat{\theta}(t)$ using (6) and (7) is obtained by entirely ignoring the within-cluster correlation and pretending that the observations within the same cluster were independent; that is, assuming working independence $\mathbf{R}_i = \mathbf{I}$. Correctly specifying the correlation matrix in fact typically has adverse effects and results in an asymptotically less efficient estimator of $\theta(t)$.

In view of the opposite asymptotic behaviors of parametric and nonparametric regression, we are led to ask whether using the conventional kernel method to estimate $\theta(t)$ will affect $\sqrt{n}$ consistency and efficiency of the estimation of $\boldsymbol{\beta}$. For example, is it still possible to specify an appropriate working correlation matrix in estimating equations in the semiparametric model (1) to obtain consistent and efficient estimators of $\boldsymbol{\beta}$ and $\theta(t)$? The various combinations of working independence and true correlation structure can be entertained for the separate estimating equations for $\boldsymbol{\beta}$ and $\theta(t)$. We pursue this question using profile likelihood ideas. We propose the profile-kernel estimating equations for the semiparametric model (1) in the next section, and answer these questions in Section 4 by performing asymptotic analysis.

### 3.2 Profile-Kernel Estimating Equations for Semiparametric Model (1)

In this section we develop estimating equations for $\boldsymbol{\beta}$ and $\theta(t)$ in the semiparametric model (1). A main feature of (1) is that $\boldsymbol{\beta}$ is a finite-dimensional parameter vector and $\theta(t)$ is an infinite-dimensional parameter. For independent data when the mean and variance functions determine a distribution, (e.g., generalized linear models), if the kernel method is used to estimate $\theta(t)$, then the profile method yields a $\sqrt{n}$-consistent and semiparametric efficient estimator of $\boldsymbol{\beta}$ (Carroll et al. 1997; Severini and Staniswalis 1994). We hence use kernel estimating equations similar to (6) and (7) to estimate $\theta(t)$, and use profile estimating equations to estimate $\boldsymbol{\beta}$ by modifying (4). We call the resulting estimating equations profile-kernel estimating equations. In the light of the discussion at the end of Section 3.1, we allow the working correlation matrices to be different in the two sets of estimating equations. In the same spirit of parametric GEEs, our primary goal is to investigate whether we can construct a $\sqrt{n}$-consistent and semiparametric efficient estimator of $\boldsymbol{\beta}$ by assuming the true correlation matrix. Our secondary goal is to investigate whether we could also construct a consistent and efficient estimator of $\theta(t)$ at the conventional nonparametric rate.

If $\boldsymbol{\beta}$ is known, then we estimate $\theta(t)$ using one of the following estimating equations:

$$\sum_{i=1}^{n} \boldsymbol{T}_i(t)^T \boldsymbol{\Delta}_i(\mathbf{X}_i, t)\mathbf{K}_{ih}^{1/2}(t)\mathbf{V}_{2i}^{-1}(\mathbf{X}_i, t)\mathbf{K}_{ih}^{1/2}(t) \times \{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{X}_i, t)\} = 0 \tag{8}$$

or

$$\sum_{i=1}^{n} \boldsymbol{T}_i(t)^T \boldsymbol{\Delta}_i(\mathbf{X}_i, t)\mathbf{V}_{2i}^{-1}(\mathbf{X}_i, t)\mathbf{K}_{ih}(t)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{X}_i, t)\} = 0, \quad (9)$$

where $\mathbf{K}_{ih}(t)$, $\boldsymbol{\mu}_i(\mathbf{X}_i, t)$, $\boldsymbol{\Delta}_i(\mathbf{X}_i, t)$, $\mathbf{V}_{2i}(\mathbf{X}_i, t) = \mathbf{S}_i^{1/2}(\mathbf{X}_i, t) \times \mathbf{R}_{2i}\mathbf{S}_i^{1/2}(\mathbf{X}_i, t)$ are the same as those in (6) and (7) except that they are evaluated at $\mu_{ij}(\mathbf{X}_{ij}, t; \boldsymbol{\beta}) = \mu\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \alpha_0 + \alpha_1(T_{ij} - t)/h\}$. Having estimated $\boldsymbol{\alpha}$ at $t$ as $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$, the kernel estimator of $\theta(t)$ is $\hat{\theta}(t; \boldsymbol{\beta}) = \hat{\alpha}_0(\boldsymbol{\beta})$. The working correlation matrix $\mathbf{R}_{2i}$ in $\mathbf{V}_{2i}(t)$ may again depend on a parameter vector $\boldsymbol{\tau}_2$, which can be estimated using the method of moments (Liang and Zeger 1986).

Estimation of $\boldsymbol{\beta}$ proceeds by solving the profile estimating equations obtained by modifying the parametric GEEs (4) and solving

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta} + \hat{\boldsymbol{\theta}}(\boldsymbol{T}_i; \boldsymbol{\beta})\}^T}{\partial \boldsymbol{\beta}} \mathbf{V}_{1i}^{-1}(\mathbf{X}_i, \boldsymbol{T}_i) \times [\mathbf{Y}_i - \boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta} + \hat{\boldsymbol{\theta}}(\boldsymbol{T}_i; \boldsymbol{\beta})\}] = 0, \quad (10)$$

where $\hat{\boldsymbol{\theta}}(\mathbf{T}_i; \boldsymbol{\beta}) = \{\hat{\theta}(T_{i1}; \boldsymbol{\beta}), \ldots, \hat{\theta}(T_{im_i}; \boldsymbol{\beta})\}^T$, $\mathbf{V}_{1i}(\mathbf{X}_i, \boldsymbol{T}_i) = \mathbf{S}_i^{1/2}(\mathbf{X}_i, \boldsymbol{T}_i)\mathbf{R}_{1i}\mathbf{S}_i^{1/2}(\mathbf{X}_i, \boldsymbol{T}_i)$, and $\mathbf{S}_i(\mathbf{X}_i, \boldsymbol{T}_i) = \operatorname{diag}\{\phi w_{ij}^{-1} V[\mu\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \hat{\theta}(T_{ij}; \boldsymbol{\beta})\}]\}$, where $\mathbf{R}_{1i}$ is a working correlation matrix depending on a parameter vector $\boldsymbol{\tau}_1$ that could be estimated using the method of moments (Liang and Zeger 1986). For example, in panel data $\mathbf{R}_{1i} \equiv \mathbf{R}$ can be estimated by $n^{-1}\sum_{i=1}^{n} \mathbf{S}_i^{-1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{S}_i^{-1/2}$, where $\mathbf{r}_i = \mathbf{Y}_i - \boldsymbol{\mu}\{\mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}}(\boldsymbol{T}_i; \hat{\boldsymbol{\beta}})\}$, where $\hat{\boldsymbol{\beta}}$ is computed from working independence. The estimators $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$ jointly solving (8) or (9), and (10) are termed profile-kernel estimators.

Our asymptotics assume that $(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$ are known, but in fact it can be shown that the results apply when they are estimated. Note that we allow the working correlation matrices $\mathbf{R}_{2i}$ in (8) or (9) and $\mathbf{R}_{1i}$ in (10) to be different. The estimator of Zeger and Diggle (1994) can be viewed as a special case of our profile-kernel estimators. They considered longitudinal Gaussian data and assumed working independence when estimating $\theta(t)$; that is, $\mathbf{R}_{2i} = \mathbf{I}$ and $\mathbf{R}_{1i}$ equal to the true correlation matrix when estimating $\boldsymbol{\beta}$. Severini and Staniswalis (1994) used (8) and (10) assuming the same working correlation matrices; that is, $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{R}_i$ or, equivalently, $\mathbf{V}_{1i} = \mathbf{V}_{2i} = \mathbf{V}_i$. Note that these authors considered local average kernel estimation instead of local linear kernel estimation as in (9). We study the asymptotic properties of the general profile-kernel estimators and these special cases in Section 4.

Our results are unexpected. Specifically, the key conclusions from our asymptotic analyses are as follows:

1. If standard smoothing is used, only when $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{I}$, i.e., assuming working independence, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent.
2. For other specifications of the working correlations $\{\mathbf{R}_{1i}, \mathbf{R}_{2i}\}$, including the case when $\mathbf{R}_{1i}$ is the true correlation matrix and any specification for $\mathbf{R}_{2i}$, except for special cases, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-inconsistent unless $\theta(t)$ is undersmoothed. When $\theta(t)$ is undersmoothed and the true correlation matrix is assumed, the resulting profile-kernel estimator $\hat{\boldsymbol{\beta}}$ is not semiparametric efficient.
3. Calculation of the semiparametric efficient estimator of $\boldsymbol{\beta}$ is complicated even in the multivariate Gaussian case: construction of the semiparametric efficient score requires solving a complicated Fredholm integral equation and estimating the multivariate joint distribution of $(\mathbf{X}, \boldsymbol{T})$.

## 4. ASYMPTOTIC RESULTS

In this section we study the asymptotic properties of the profile-kernel estimators $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$. We focus on the symmetric local linear kernel estimating equations (8) and the profile estimating equations (10). The reason that we focus on (8) instead of (9) in our asymptotic analysis is that the asymptotic properties of the estimators under (9) are difficult to study because of the asymmetric nature of (9) (Lin and Carroll 2000). However, we show that if one uses in (9) the local *average* kernel, which includes the existing estimators (Severini and Staniswalis 1994; Zeger and Diggle 1994) as special cases, then the resulting estimators have qualitatively similar asymptotic properties to those of $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$. In what follows, let $m_i = m < \infty$, (i.e., assuming finite cluster size) and let $T$ be a continuous observation-level covariate (e.g., a time-varying covariate in longitudinal studies).

We allow the $m$ components of $\mathbf{X}_i$ and $\boldsymbol{T}_i$ to be correlated unless stated otherwise and assume the density of $\boldsymbol{T}_i$ to be continuous. We further assume that the $(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{T}_i)$ $(i = 1, \ldots, n)$ are iid triplets and that both $\mathbf{V}_{1i}(\boldsymbol{\mu}_i, \boldsymbol{\tau}) = \mathbf{V}_1(\boldsymbol{\mu}_i, \boldsymbol{\tau})$ and $\mathbf{V}_{2i}(\boldsymbol{\mu}_i, \boldsymbol{\tau}) = \mathbf{V}_2(\boldsymbol{\mu}_i, \boldsymbol{\tau})$ are invertible. Let $d^{(r)}(\cdot)$ denote the $r$th derivative of any function $d(\cdot)$, let $v^{jk}$ denote the $(j, k)$th element of a matrix $\mathbf{V}^{-1}$, and let $f_j(t)$ denote the marginal density of $T_{ij}$. Suppose that the kernel density function $K(\cdot)$ has mean 0 and unit variance; that is, $\int sK(s)du = 0$ and $\int s^2K(s) = 1$.

We first rewrite the profile estimating equations for $\boldsymbol{\beta}$ in (10) as

$$\sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^T \boldsymbol{\Delta}(\mathbf{X}_i, \boldsymbol{T}_i)\mathbf{V}_{1i}^{-1}(\mathbf{X}_i, \boldsymbol{T}_i) \times [\mathbf{Y}_i - \boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta} + \hat{\theta}(\boldsymbol{T}_i; \boldsymbol{\beta})\}] = 0, \quad (11)$$

where $\widetilde{\mathbf{X}}_i = \mathbf{X}_i + \partial\hat{\theta}(\boldsymbol{T}_i; \boldsymbol{\beta})/\partial\boldsymbol{\beta}^T$ and $\boldsymbol{\Delta}(\mathbf{X}_i, \boldsymbol{T}_i) = \operatorname{diag}[\mu^{(1)}\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \hat{\theta}(T_{ij}; \boldsymbol{\beta})\}]$. Calculations in Appendix A show that, asymptotically, $\partial\hat{\theta}(t; \boldsymbol{\beta})/\partial\boldsymbol{\beta} = -W_2^{-1}(t)\mathbf{W}_2^x(t) + op(1)$, where, suppressing the index $i$ denoting $\mu_l = \mu\{\mathbf{X}_l^T\boldsymbol{\beta} + \theta(t)\}$ $(l = 1, \ldots, m)$,

$$W_2(t) = \sum_{l=1}^{m} E\left[\left\{\mu_l^{(1)}\right\}^2 v_2^{ll} | T_l = t\right] f_l(t)$$

and

$$\mathbf{W}_2^x(t) = \sum_{l=1}^{m} E\left[\left\{\mu_l^{(1)}\right\}^2 v_2^{ll}\mathbf{X}_l | T_l = t\right] f_l(t).$$

It follows that $\widetilde{\mathbf{X}}_i = (\widetilde{\mathbf{X}}_{i1}, \ldots, \widetilde{\mathbf{X}}_{im})^T$, where $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij} - W_2^{-1}(T_{ij})\mathbf{W}_2^x(T_{ij})$. Using these results, in Result 1 we study the asymptotic distributions of $\{\hat{\theta}(t), \hat{\boldsymbol{\beta}}\}$. A sketch of its proof is given in Appendix A.

*Result 1.* Let $\{\hat{\theta}(t), \hat{\boldsymbol{\beta}}\}$ denote the solution of the profile-kernel estimating equations (8) and (10), where $\hat{\theta}(t) = \theta(t; \hat{\boldsymbol{\beta}})$. Suppose that $h \propto n^{-\alpha}$, $1/5 \le \alpha \le 1/3$ and $n \to \infty$. We then have the following:

a. If $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent, [i.e., $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$], then there is an asymptotically equivalent random variable such that

$$\text{bias}\{\hat{\theta}(t)\} \approx h^2\theta^{(2)}(t)/2 \tag{12}$$

and

$$\text{var}\{\hat{\theta}(t)\} \approx \frac{\gamma}{nh}\frac{\sum_{j=1}^m E\left[\left\{\mu_j^{(1)}\right\}^2\left\{v_2^{jj}\right\}^2\sigma_{jj}\middle|T_j = t\right]f_j(t)}{\left\{\sum_{j=1}^m E\left[\left\{\mu_j^{(1)}\right\}^2 v_2^{jj}\middle|T_j = t\right]f_j(t)\right\}^2}, \tag{13}$$

where $\sigma_{jj} = \text{var}(Y_j|\mathbf{X}_j, T_j) = \phi w_j^{-1}V(\mu_j)$. It follows that $\text{var}\{\hat{\theta}(t)\}$ is minimized when assuming working independence $\mathbf{R}_2 = \mathbf{I}$ and is

$$\text{var}\{\hat{\theta}(t)\} \approx \frac{\gamma}{nh}\left\{\sum_{j=1}^m E\left[\left\{\mu_j^{(1)}\right\}^2\sigma_{jj}^{-1}\middle|T_j = t\right]f_j(t)\right\}^{-1}. \tag{14}$$

b. The estimator $\hat{\boldsymbol{\beta}}$ converges in distribution: $\sqrt{n}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} - h^2 b(\boldsymbol{\beta}, \theta)/2\} \to N(0, \mathbf{V}_{\boldsymbol{\beta}})$, where, suppressing the subscript $i$ in each term inside the expectations,

$$\mathbf{b}(\boldsymbol{\beta}, \theta) = \{E(\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}})\}^{-1}E\{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\theta^{(2)}(\mathbf{T})\},$$

$$\mathbf{V}_{\boldsymbol{\beta}} = \{E(\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}})\}^{-1}E\{(\mathbf{Z}_1 - \mathbf{Z}_2)^T\boldsymbol{\Sigma}(\mathbf{Z}_1 - \mathbf{Z}_2)\} \times \{E(\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}})\}^{-1},$$

$\boldsymbol{\Sigma} = \text{cov}(\mathbf{Y}|\mathbf{X}, \boldsymbol{T})$ and $\mathbf{Z}_1 = \mathbf{V}_1^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}}$, and the $j$th row of $\mathbf{Z}_2$ is

$$\mathbf{Z}_{2j} = \mu_j^{(1)}v_2^{jj}\left\{\sum_{k=1}^m\sum_{l=1}^m E\left[\widetilde{\mathbf{X}}_k\mu_k^{(1)}v_1^{kl}\mu_l^{(1)}|T_l = T_j)\right]\right\} \times W_2^{-1}(T_j)f_j(T_j).$$

c. If these two conditions—working independence is assumed in both (8) and (10), (i.e., $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{I}$) and $(\mathbf{X}_{ij}, T_{ij})$ have the same marginal density, [i.e., $f_j(\mathbf{X}_{ij}, T_{ij}) = f(\mathbf{X}_{ij}, T_{ij})$]—are satisfied, then $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent; that is, the bias term $\mathbf{b}(\boldsymbol{\beta}, \boldsymbol{\theta}) = 0$ and $\sqrt{n}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\} \to N(0, \widetilde{\mathbf{V}}_{\boldsymbol{\beta}})$ in distribution, where, suppressing the subscript $i$ in each term inside the expectations,

$$\widetilde{\mathbf{V}}_{\boldsymbol{\beta}} = \{E(\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}})\}^{-1}E\{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}}\} \times \{E(\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}})\}^{-1},$$

and $\boldsymbol{\Sigma}_d$ is a diagonal matrix with the diagonal elements of $\boldsymbol{\Sigma}$, (i.e., $\sigma_{jj}$) on the diagonal.

d. For other specifications of the working correlation matrices $\mathbf{R}_{1i}$ and $\mathbf{R}_{2i}$, including the true correlation matrix, $\hat{\boldsymbol{\beta}}$ is often $\sqrt{n}$ inconsistent; that is, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \infty$ in distribution. However, if one assumes that $nh^4 \to 0$ [i.e., undersmooths $\theta(t)$], then for any specification of the working correlation matrices $\mathbf{R}_{1i}$ and $\mathbf{R}_{2i}$, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \mathbf{V}_{\boldsymbol{\beta}})$ in distribution.

In general, $\mathbf{V}_{\boldsymbol{\beta}}$ can be estimated by replacing terms in its expression by estimates of those terms. We conjecture that the bootstrap can also be used. The results in part a of Result 1 are similar to those of Lin and Carroll (2000) when the covariate $\mathbf{X}$ is absent in model (1), except that the variance of $\hat{\theta}(t)$ now involves conditional expectations of $\mathbf{X}_j$ given $T_j$. These results suggest that if the profile estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent, then $\hat{\theta}(t)$ is consistent and asymptotically normal at the regular nonparametric rate. The most efficient estimator of $\hat{\theta}(t)$ is obtained by completely ignoring the within-cluster correlation.

To see why the bias term $\boldsymbol{b}(\boldsymbol{\beta}, \theta) \ne 0$ for non-identity working correlation matrices, consider linear models for multivariate normal $\mathbf{Y}_i$. Suppose that the marginal density of $\{\mathbf{X}_{ij}, T_{ij}\}$ $(j = 1, \ldots, m)$ is the same. Then the $j$th component of $\widetilde{\mathbf{X}}$ is $\widetilde{\mathbf{X}}_j = \mathbf{X}_j - E(\mathbf{X}_j|T_j)$. It follows that the second term of $\boldsymbol{b}(\boldsymbol{\beta}, \theta)$ is $E\{\widetilde{\mathbf{X}}^T\mathbf{V}_1^{-1}\theta^{(2)}(\boldsymbol{T})\} = \sum_{j=1}^m\sum_{k=1}^m E\{C_{jk}(T_k)v_1^{jk}\theta(T_k)\}$, where $C_{jk}(T_k) = E(\mathbf{X}_j|T_k) - E\{E(\mathbf{X}_j|T_j)|T_k)\}$ is generally not equal to 0 except when $j = k$. This means that the bias term $\boldsymbol{b}(\boldsymbol{\beta}, \theta) \ne 0$ unless we assume working independence, (i.e., $\mathbf{R}_1 = \mathbf{I}$), or $E(\mathbf{X}_j|T_j, T_k) = E(\mathbf{X}_j|T_j)$ for any $j, k$ (e.g., when $\mathbf{X}$ and $\boldsymbol{T}$ are independent).

Simple calculations show that for multivariate normal $\mathbf{Y}$, if $\mathbf{X}$ and $\boldsymbol{T}$ are independent, then $\hat{\boldsymbol{\beta}}$ in fact is $\sqrt{n}$ consistent for any arbitrary working correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. Furthermore, as shown in Section 5, if one assumes $\mathbf{R}_{1i}$ equal to the true correlation matrix in (10) and working independence $\mathbf{R}_{2i} = \mathbf{I}$ in (8), then $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and semiparametric efficient, and $\hat{\theta}(t)$ is efficient as well. The foregoing independence assumption of $\mathbf{X}$ and $\boldsymbol{T}$ is strong and difficult to satisfy in practice if both covariates $\mathbf{X}$ and $T$ are time-varying covariates. But if $\mathbf{X}$ contains only one-time covariates and $T$ is time in longitudinal studies, then this condition is satisfied. Note that the outcome needs to be normally distributed for the foregoing results to hold. For non-Gaussian data, if the true correlation matrix is used, even when $\mathbf{X}$ and $T$ are independent, then $\hat{\boldsymbol{\beta}}$ is still $\sqrt{n}$ inconsistent.

Result 1 assumes that $\theta(t)$ is estimated using the symmetric local linear kernel estimating equation (8). Severini and Staniswalis (1994) and Zeger and Diggle (1994) proposed slightly different estimators. They estimated $\theta(t)$ by replacing the *symmetric local linear* kernel estimating equation (8) with the *asymmetric local average* kernel estimating equation, which is obtained by letting $\mu(\mathbf{X}_{ij}, t) = \mu(\mathbf{X}_{ij}^T\boldsymbol{\beta} + \alpha_0)$ and replacing $\boldsymbol{T}_i(t)$ by $\mathbf{1}_i$ in (9). We denote these estimators by $\{\hat{\boldsymbol{\beta}}_*, \hat{\theta}_*(t)\}$. Specifically, Severini and Staniswalis (1994) assumed the same working correlation matrix in both $\theta(t)$ and $\boldsymbol{\beta}$ estimating equations, that is, $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{R}_i$. Zeger and Diggle (1994) considered Gaussian data and assumed $\mathbf{R}_{1i}$ equal to the true correlation and $\mathbf{R}_{2i} = \mathbf{I}_i$ (working independence). It can be shown that the asymptotic properties of

$\{\hat{\boldsymbol{\beta}}_*, \hat{\theta}_*(t)\}$ are similar to those of $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$ in Result 1, and that the conclusions are the same.

*Computation.* A Fisher–Sivring algorithm for computation for the working indendence estimation is given in Appenix C.

## 5. SEMIPARAMETRIC EFFICIENT SCORE

It is of substantial interest to understand why the profile-kernel estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ inconsistent when the true correlation matrix is used unless $\theta(t)$ is undersmoothed. One way to address this question is to define a likelihood function for $\mathbf{Y}_i$ and compare how the profile-kernel estimating equation (10) differs from the semiparametric efficient score for $\boldsymbol{\beta}$ (Bickel et al., 1993).

The motivation of this investigation is as follows. For independent data, (i.e., the cluster size $m = 1$), suppose that the distribution of the outcome $Y$ belongs to the linear exponential family. If $\theta(t)$ is smoothed using standard kernel methods (e.g., cross-validation), then the profile-kernel estimating equation of $\boldsymbol{\beta}$ is asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$ (Carroll et al. 1997; Severini and Staniswalis 1994). The resulting profile estimator $\hat{\boldsymbol{\beta}}$ hence is $\sqrt{n}$ consistent and semiparametric efficient. If one uses an estimating equation for $\boldsymbol{\beta}$ asymptotically different from the semiparametric efficient score [e.g., by simply replacing $\widetilde{\mathbf{X}}_i$ in (11) (simplified for $m = 1$) by $\mathbf{X}_i$], then the resulting estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ inconsistent unless $\theta(t)$ is undersmoothed (Rice 1986).

Our key findings in this section are as follows. First, the semiparametric efficient score of $\boldsymbol{\beta}$ for multivariate Gaussian data is complicated and requires solving the Fredholm integral equation of the second kind and estimating the joint distribution of $\mathbf{X}_i$ and $\boldsymbol{T}_i$. Second, if regular smoothing is used for estimating $\theta(t)$, then the profile-kernel score of $\boldsymbol{\beta}$ estimates the semiparametric efficient score with a nonzero bias. This explains why the profile-kernel estimator $\hat{\boldsymbol{\beta}}$ is often $\sqrt{n}$ inconsistent. Finally, when $\hat{\theta}(t)$ is undersmoothed, the profile-kernel estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent but is still not semiparametric efficient, except for special cases.

We first derive the semiparametric efficient score of $\boldsymbol{\beta}$. We assume a constant cluster size $1 < m < \infty$ and suppress the index $i$. To understand the fundamental issues involved, we consider $\mathbf{Y}$ to be multivariate normal $N\{\mathbf{X}\boldsymbol{\beta} + \theta(\boldsymbol{T}), \mathbf{V}\}$, where $\theta(\boldsymbol{T}) = \{\theta(T_1), \ldots, \theta(T_m)\}^T$ and $\mathbf{V}$ is assumed known.

In Appendix B we show that the semiparametric efficient score of $\beta$ is

$$\{\mathbf{X} - \boldsymbol{\varphi}_*(\boldsymbol{T})\}^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \theta(\boldsymbol{T})\}, \tag{15}$$

where $\boldsymbol{\varphi}_*(\boldsymbol{T}) = \{\boldsymbol{\varphi}_*(T_1), \ldots, \boldsymbol{\varphi}_*(T_m)\}^T$, $\boldsymbol{\varphi}_*(T_j) = \{\varphi_{*1}(T_j), \ldots, \varphi_{*p}(T_j)\}^T$, and $p$ is the dimension of $\boldsymbol{\beta}$. The semiparametric efficiency bound of $\boldsymbol{\beta}$ is $E\{[\mathbf{X} - \boldsymbol{\varphi}_*(\boldsymbol{T})]^T \mathbf{V}^{-1}[\mathbf{X} - \boldsymbol{\varphi}_*(\boldsymbol{T})]\}$. The function $\boldsymbol{\varphi}_*(t)$ solves

$$\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E\{[\mathbf{X}_j - \boldsymbol{\varphi}_*(T_j)]|T_k = t\} f_k(t) = 0, \tag{16}$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_m)^T$, $v^{jk}$ is the $(j, k)$th element of $\mathbf{V}^{-1}$, and $f_k(t)$ is the density of $T_k$. Simple calculations show that (16) can be written as the Fredholm integral equation of the second kind (Bronshtein and Semendyayev 1985, sec. 8.4)

$$\boldsymbol{\varphi}_*(t) + \int H(t, s)\boldsymbol{\varphi}_*(s)ds = q(t), \tag{17}$$

where $H(t, s)$ and $\boldsymbol{q}(t)$ are defined as

$$H(t, s) = \frac{\sum\sum_{j\neq k} v^{jk} f(T_j = s, T_k = t)}{\sum_{j=1}^{m} v^{jj} f(T_j = t)}$$

and

$$q(t) = \frac{\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E(\mathbf{X}_j|T_k = t) f(T_k = t)}{\sum_{j=1}^{m} v^{jj} f(T_j = t)},$$

where $f(\cdot)$ denotes a density function.

If $H(t, s)$ is square-integrable, then (17) has only one solution, except when the eigenvalues of (17) contain $-1$ and its solution can be written as $\boldsymbol{\varphi}_*(t) = -\int \Gamma(t, s)\boldsymbol{q}(s)ds + \boldsymbol{q}(t)$, where $\Gamma(t, s)$ is called the resolvent kernel and can be written as the Fredholm series, $\Gamma(t, s) = \sum_{k=0}^{\infty} H_k(t, s)/\sum_{k=0}^{\infty} \delta_k$, with $\delta_0 = 0$, $H_0(t, s) = H(t, s)$, $\delta_k = k^{-1}\int H_{k-1}(t, t)dt$, and $H_k(t, s) = H_{k-1}(t, s)\delta_k - \int H(t, u)H_{k-1}(u, s)du$ (Bronshtein and Semendyayev, 1985, sec. 8.4.7). An alternative expression of $\Gamma(t, s)$ is given by the Neumann series (Bronshtein and Semendyayev 1985, sec. 8.4.6). The foregoing Fredholm series always converges but is of little use when numerically calculating $\boldsymbol{\varphi}_*(t)$, because in most cases the approximation is inadequate for small values of $k$. More useful is the Nyström method (Bronshtein and Semendyayev 1985, sec. 8.4.8).

The foregoing discussion suggests that construction of the semiparametric efficient score of $\boldsymbol{\beta}$ is complicated even in the multivariate normal case. One needs to solve the complicated integral equation (17), which requires estimating the pairwise joint densities of $(T_j, T_k)$ and the pairwise conditional expectations $E(\mathbf{X}_j|T_k)$ when calculating $H(t, s)$ and $\boldsymbol{q}(t)$. However, in the special case when the marginal density of $(\mathbf{X}_j, T_j)$ is the same and $E(\mathbf{X}_j|T_j, T_k) = E(\mathbf{X}_j|T_k)$ (e.g., when $\mathbf{X}$ and $T$ are independent), simple calculations show that the solution of (16) has the closed form $\boldsymbol{\varphi}_*(t) = E(\mathbf{X}_j|T_j = t)$.

We now study for multivariate Gaussian data how the semiparametric efficient score (15) asymptotically differs from the profile-kernel estimating equation of $\boldsymbol{\beta}$ in (10) when the working correlation matrix $\mathbf{R}$ is the true correlation matrix. Using the results in Appendix A, we can easily show that the profile estimating equation for $\boldsymbol{\beta}$ in (11) is asymptotically equivalent to

$$(\widetilde{\mathbf{X}}^T\mathbf{V}^{-1} - \mathbf{Z}_2^T)\{\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \theta(\boldsymbol{T})\} + \widetilde{\mathbf{X}}^T\mathbf{V}^{-1}\boldsymbol{\theta}^{(2)}(T)h^2/2, \tag{18}$$

where the $j$th component of $\widetilde{\mathbf{X}}$ is $\widetilde{\mathbf{X}}_j = \mathbf{X}_j - E(\mathbf{X}_j|T_j)$ and $\mathbf{Z}_2$ is defined in Result 1. A comparison between (15) and (18) suggests that they are often different and that (18) is often subject to a nonzero bias. Even when $\theta(t)$ is undersmoothed [i.e., the second bias term in (18) is 0], some calculations show that the first term in (18) is still generally different from (15). In other words, the profile-kernel score (10) is often asymptotically different from the semiparametric efficient score (15). But when $\mathbf{X}$ and $T$ are independent, they are the same asymptotically,

and the profile-kernel estimator of $\boldsymbol{\beta}$ hence is $\sqrt{n}$ consistent and semiparametric efficient. Some calculations show that the same conclusion holds for the profile-kernel estimator $\hat{\boldsymbol{\beta}}_*$ when $\hat{\theta}_*(t)$ is the average kernel estimator obtained using the asymmetric kernel estimating equation (9); see Section 4.

It is difficult to construct the semiparametric efficient score directly using the complicated form of $\boldsymbol{\varphi}_*(t)$ in (15), because this involves theoretical density functions and expectations. This raises an open question on how to construct a practical semiparametric efficient estimator of $\boldsymbol{\beta}$. It is a reasonable conjecture that if such a construction is pushed through, then undersmoothing will not be required.

## 6. PRACTICAL IMPLICATIONS OF THE THEORETICAL RESULTS AND COMPUTATION OF THE ESTIMATES

*Cross-Validation.* Conventional bandwidth selection techniques, such as cross-validation by deleting one cluster data at a time, fail unless working independence is assumed. Because the bandwidth $h$ chosen by cross-validation satisfies $h = O(n^{-1/5})$, $\hat{\boldsymbol{\beta}}$ will be $\sqrt{n}$ inconsistent unless working independence is assumed (Result 1). Unfortunately, there is no generally accepted data-driven way to choose $h$ to undersmooth $\theta(t)$, although ad hoc methods have been proposed (Brockmann, Gasser, and Herrmann 1993). In our experience, we have found that multiplying the bandwidth by $n^{-2/15}$, which makes $h \propto n^{-1/3}$, often works quite well in practice. Presumably, other methods (e.g., higher–order kernels, twicing) can be used to eliminate the bias.

*Sandwich Method.* The sandwich method, which is commonly used in calculating the covariance estimator of $\hat{\boldsymbol{\beta}}$ in estimating equations (Liang and Zeger 1986), will give an inconsistent estimator of $\operatorname{cov}(\hat{\boldsymbol{\beta}})$ unless working independence is assumed. This is because it ignores the extra $\boldsymbol{Z}_2$ term in $\mathbf{V}_{\boldsymbol{\beta}}$ in part b of Result 1. This is true even when one undersmoothes $\theta(t)$. We conjecture that the bootstrap can be used.

*Hypothesis Testing.* One is often interested in testing $H_0$ : $\boldsymbol{\beta} = 0$ or part of $\boldsymbol{\beta}$ is 0. If conventional smoothing techniques such as cross-validation are used, then the Wald test and the score test for $H_0$ will be inconsistent unless working independence is assumed or $\theta(t)$ is undersmoothed. For example, when the Wald test is used, $\hat{\boldsymbol{\beta}}$ in fact estimates the true $\boldsymbol{\beta}$ plus the bias term $b(\boldsymbol{\beta}, \theta)h^2/2$.

*Functional Data Analysis.* The simplest functional regression model (Ramsay and Dalzell 1991) is $Y_i(t) = \theta(t) + \epsilon_i(t)$, where $i$ indexes the $i$th subject, $t$ indexes time $t$, and $\epsilon_i(t)$ is an error whose distribution is a Gaussian process with mean 0 and $\operatorname{cov}\{e(t), e(s)\} = \sigma(t, s)$. Rice and Silverman (1991) considered estimating $\theta(t)$ using a smoothing spline. The results of Lin and Carroll (2000) suggest that the most efficient estimator of $\theta(t)$ when the kernel method is used is obtained by entirely ignoring the correlation of the repeated measures of $Y_i(t)$ over time. In the presence of covariates $\mathbf{X}_i(t) = \{X_{i1}(t), \ldots, X_{ip}(t)\}^T$, a semiparametric functional regression model could be considered,

$$Y_i(t) = \mathbf{X}_i(t)^T\boldsymbol{\beta} + \theta(t) + \epsilon_i(t). \tag{19}$$

The semiparametric model (1) is a discrete version of (19).

Suppose that the profile-kernel method is used to estimate $\{\boldsymbol{\beta}, \theta(t)\}$. Our results suggest that (a) if $\mathbf{X}_i(t)$ is a vector of one-time subject-level covariates (i.e., $\mathbf{X}_i(t) = \mathbf{X}_i$ free of $t$), by specifying $\mathbf{R}_1$ as the true correlation matrix and $\mathbf{R}_2 = \mathbf{I}$, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and semiparametric efficient and $\hat{\theta}(t)$ is asymptotically efficient as well, and (b) if $\mathbf{X}_i(t)$ contains time-varying covariates (i.e., $\mathbf{X}$ and $T$ are not independent), then one must assume working independence ($\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$) or undersmooth $\theta(t)$ to obtain a $\sqrt{n}$ consistent (but inefficient) estimator of $\hat{\boldsymbol{\beta}}$.

It is important to emphasize that our results assume that the number of observations per subject $m$ is finite, as is common in longitudinal studies. With $T$ being time, our asymptotic analysis thus assumes that observations from different subjects may be observed at different time points asymptotically, but the number of observations per subject remains bounded.

*Computation.* A Fisher-Sivring algorithm for computation for the working indendence estimator is given in Appendix C.

## 7. SIMULATION STUDY

We conducted a simulation study to evaluate the finite-sample performance of the profile-kernel method. Each dataset comprised $n = 100$ subjects and $m_i = 3$ observations per subject over time. The covariate vector $\mathbf{X}_{ij}$ was set at $\mathbf{X}_{ij} = (X_{1ij}, X_{2i})^T$, where $X_{1ij}$ a time-varying covariate and $X_{2i}$ is a subject level covariate that takes value 1 for half of the subjects and 0 for the other half and mimics a binary treatment indicator. We generated $X_{1ij}$ and $T_{ij}$ according to the model $X_{1ij} = b_i + e_{ij}$ and $T_{ij} = b_i + e'_{ij}$, where $b_i \sim$ uniform$(-1, 1)$ and $e_{ij}$ and $e'_{ij}$ are independent and follow uniform$(-1, 1)$. This setup allows the $X_{1ij}$ and the $T_{ij}$ to be correlated with each other and over time between their repeated measures with exchangeable correlation .5. Conditional on $\mathbf{X}_{ij}$ and $T_{ij}$, we generated the outcome $Y_{ij}$ from multivariate normal with mean $\mu_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2i} + \theta(T_{ij})$, where $\beta_1 = \beta_2 = 1.0$ and $\theta(t) = \sin(2t)$, and $Y_{ij}$ has variance 1 and exchangeable correlation .5.

We generated 200 datasets with $N = 300$ observations each and analyzed them using the profile-kernel methods. For each simulated dataset, we first assumed working independence when we calculated the profile-kernel estimate of $\boldsymbol{\beta}$ and $\theta(t)$ and estimated the bandwidth parameter $h$ needed for the kernel estimate of $\theta(t)$ using cross-validation by deleting one subject data at a time. We next calculated the profile-kernel estimate of $\boldsymbol{\beta}$ and $\theta(t)$ by accounting for the within-subject correlation. Specifically, we estimated the true covariance of $\mathbf{Y}_i$ using the method of moments and calculated the bandwidth parameter $h$ by multiplying the cross-validation bandwidth estimate by $n^{-2/15}$. This undersmooths $\theta(t)$ and eliminates the bias term (Sec. 6), at least theoretically.

Table 1 gives the averaged estimated regression coefficients of $\beta_1$ and $\beta_2$, along with their empirical and estimated standard errors (SEs) when working independence is assumed and when the true covariance of $\mathbf{Y}_i$ is estimated. When assuming working independence, we estimated the SEs of $\hat{\boldsymbol{\beta}}$ using the sandwich estimate given in Appendix C. When assuming that the true covariance is estimated, we estimated the SEs of $\hat{\boldsymbol{\beta}}$ using a finite-sample estimate of $\mathbf{V}_{\boldsymbol{\beta}}$ given in part b of

*Table 1. Means and Standard Errors of Regression Coefficient Estimates Over 200 Replications*

| | Working independence | | | True covariance | | |
|---|---|---|---|---|---|---|
| *Parameter* | *Mean* | *Empirical SE* | *Estimated SE* | *Mean* | *Empirical SE* | *Estimated SE* |
| $\beta_1$ | 1.005 | .088 | .084 | 1.002 | .075 | .070 |
| $\beta_2$ | 1.020 | .160 | .160 | 1.022 | .161 | .158 |

NOTE: True values are $\beta_1 = 1.0$ and $\beta_1 = 1.0$.

Result 1. Table 1 reports the averages of the estimated standard errors over 200 replications. The results in the table show that the profile-kernel method performs well in finite samples and that the biases in the profile-kernel estimates of $\boldsymbol{\beta}$ are minimal under both covariance assumptions. The estimate of $\boldsymbol{\beta}_1$, the coefficient of the time-varying covariate $X_1$, is more efficient when the true covariance is estimated than when working independence is assumed. However, no gain in efficiency is realized in $\hat{\boldsymbol{\beta}}_2$ by estimating the true covariance of $\mathbf{Y}_i$. This is because $X_2$ is a subject-level covariate and is independent of $T_{ij}$ and the design is balanced with respect to $X_2$. The simulation results are consistent with the theory. The estimated SEs of $\hat{\boldsymbol{\beta}}$ also agree well with the simulated SEs.

Figure 1 compares the true nonparametric function $\theta(t)$ to the kernel estimates of $\theta(t)$ when assuming working independence and when the true covariance is estimated. Both kernel estimates of $\theta(t)$ are close to the true $\theta(t)$. Figure 2 compares the SEs of these two kernel estimates. It suggests that assuming working independence gives a more efficient kernel estimate of $\theta(t)$ than that achieved when assuming the true covariance. These results agree well with the theory.

## 8. APPLICATION TO THE INFECTIOUS DISEASE DATA

In this section we apply the semiparametric model (1) to analyzing the longitudinal infectious disease data introduced in Section 1. A total of 1,200 binary indicators for the presence of respiratory infection (0 = no, 1 = yes) were collected on 275 preschool-age children examined every quarter for up to six consecutive quarters. The primary interest was to study the association between respiratory infection and the exposure variable vitamin A deficiency, which was manifested by xerophthalmia status (0 = no; 1 = yes), while adjusting for several key confounders. These confounders include age in years, sex (0 = male, 1 = female), height for age, and stunting status (0 = no, 1 = yes). (For a detailed description of the covariates, see Zeger and Karim 1991.)

Examination of the distribution of the vertical strokes in Figure 3 suggests that the age effect departs dramatically from linearity. To avoid possible confounding of misspecification of the age effects on estimation of the effect of the key exposure xerophthalmia, we consider a semiparametric logistic model for the $j$th observation of the $i$th subject as

$$\text{logit}\{\Pr(Y_{ij} = 1)\} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \theta(\text{age}_{ij}), \tag{20}$$

where $\mathbf{X}_{ij}$ comprises xerophthalmia status, seasonal cosine and sine, sex, height for age, and stunting, and $\theta(\text{age}_{ij})$ is a smooth function of age. Examination of the data suggested that the height for age effect was linear, and hence we included it in $\mathbf{X}_{ij}$.

We used the profile-kernel method assuming working independence using the algorithm in Appendix C and calculated the SEs using the sandwich method. We chose the bandwidth parameter $h$ using the empirical bias bandwidth selection (EBBS) method (Ruppert 1997). Figure 3 shows the estimated nonparametric function of age and its 95% confidence interval. The risk of respiratory infection increased slightly during the first 2 years of life and decreased thereafter. Table 2 gives the estimated regression coefficients $\boldsymbol{\beta}$. The data provide no evidence for vitamin A deficiency on respiratory infection, but strong evidence for the association between respiratory infection and sex and season.

To examine whether a simple parametric model can fit the data equally well as the semiparametric model, we fit a parametric GEE model with $\theta(\text{age})$ to be quadratic assuming

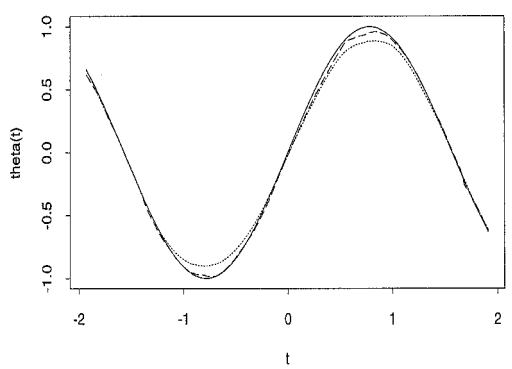


*Figure 1. True and Estimated Nonparametric Functions $\hat{\theta}(t)$ Based on 200 Replications: (—— True; - - - - assuming working independence; – – – assuming that the true covariance is estimated).*

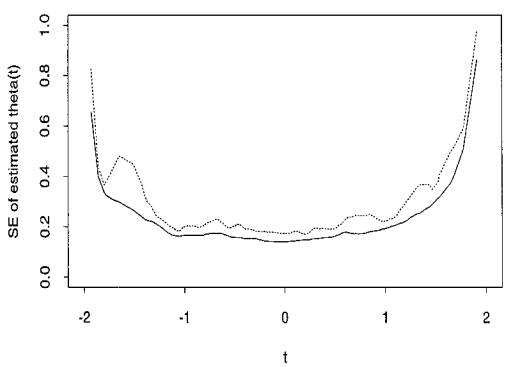


*Figure 2. Empirical Pointwise SEs of the Estimated Nonparametric Functions $\hat{\theta}(t)$ Based on 200 Replications: (—— assuming working independence; - - - - assuming that the true covariance is estimated).*

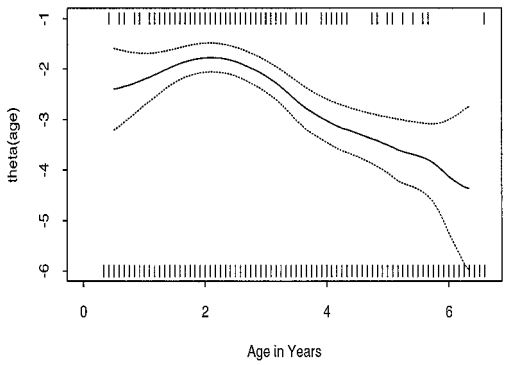


*Figure 3. Estimated Kernel Estimate $\hat{\theta}$(age) When Fitting the Semiparametric Model (20) to the Infectious Disease Data Assuming Working Independence and Its 95% Pointwise Confidence Intervals (—— $\theta$(age); - - - - 95% confidence interval). The vertical strokes at 0 and − 6 indicate the occurrence of 1 and 0 in the response.*

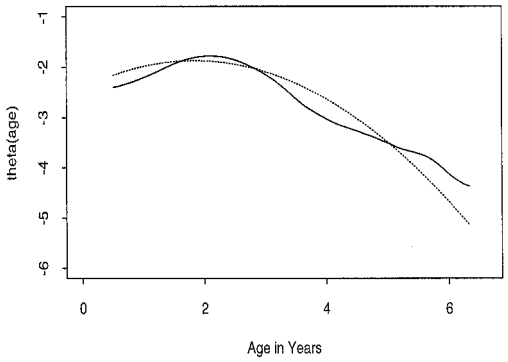


*Figure 4. Comparison of the Kernel Estimate $\hat{\theta}$(age) (——) and the Quadratic Estimate of Age (- - - -).*

working independence. Figure 4 compares the semiparametric kernel estimate of $\theta(t)$ to its quadratic counterpart (Diggle et al. 1994, p. 161). The semiparametric kernel estimate suggests that some excess nonlinearity may be undetected by the quadratic age model, a conjecture confirmed by the fact that a cubic age model fit using GEE had a statistically significant cubic age term ($p$ value .02). Table 2 compares the regression coefficients $\boldsymbol{\beta}$ estimated using the semiparametric model and the parametric quadratic age model. The coefficient estimates of stunting were considerably different using the two methods, although the other coefficient estimates are similar. This difference was due mainly to misspecification of the quadratic age effect.

## 9. DISCUSSION

We have considered a marginal semiparametric partially linear generalized linear model for clustered data, where the effects of some covariates $\mathbf{X}$ are modeled parametrically as $\mathbf{X}\boldsymbol{\beta}$ and the effect of some other covariate $T$ is modeled nonparametrically as $\theta(t)$. Our results apply to the case where the number of observations per cluster is finite and the number of clusters is large. The profile-kernel estimating equations in the literature are used for estimation. The results are unexpected.

We show that for clustered data, this conventional profile-kernel method fails to yield a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ unless working independence is assumed or $\theta(t)$ is artificially undersmoothed. Under working independence, one may need to greatly sacrifice efficiency to achieve $\sqrt{n}$ consistency of $\boldsymbol{\beta}$.

*Table 2. Regression Coefficient Estimates in Analysis of the Infectious Disease Data Using the Semiparametric Model and the Quadratic Age Model*

| | *Semiparametric model* | | *Quadratic age model* | |
|---|---|---|---|---|
| | *Estimate* | *SE* | *Estimate* | *SE* |
| Vitamin A | .611 | .529 | .629 | .413 |
| Seasonal cosine | −.587 | .210 | −.590 | .172 |
| Seasonal sine | −.161 | .183 | −.170 | .148 |
| Sex | −.508 | .295 | −.485 | .240 |
| Height | −.026 | .035 | −.030 | .029 |
| Stunting | .463 | .525 | .272 | .417 |

When $\theta(t)$ is artificially undersmoothed, the profile-kernel estimator of $\boldsymbol{\beta}$ is still not semiparametric efficient, except for special cases.

To explain why the profile-kernel method fails in clustered data, we have derived the semiparametric efficient score of $\boldsymbol{\beta}$ for multivariate normal semiparametric models. We show that unlike in the independent data case, the profile-kernel method fails to provide an estimated score equation that is asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$. Even in this simple multivariate normal case, the semiparametric efficient score of $\boldsymbol{\beta}$ is complicated and requires solving the Fredholm integral equation and estimating the pairwise joint distributions of all observations $(\mathbf{X}_j, \mathbf{X}_k, T_j, T_k)$ in the same cluster. Direct estimation of such densities is complicated and could well be infeasible or cumbersome, especially when cluster sizes vary from one cluster to another. For example, in longitudinal data, different subjects could have different numbers of observations, and these different observations might be observed at different time points. Estimation of the joint distribution of $\mathbf{X}$ and $\boldsymbol{T}$ is hence difficult. One strategy is to assume a parametric model for $\mathbf{X}$ and $\boldsymbol{T}$ to estimate the joint distribution of $\mathbf{X}$ and $\boldsymbol{T}$. But this could lead to an inconsistent estimator of $\boldsymbol{\beta}$ if such a parametric model for $\mathbf{X}$ and $\boldsymbol{T}$ is misspecified. This leaves an open question on how to construct a semiparametric efficient estimator of $\boldsymbol{\beta}$ in practice for clustered data. Further research is needed.

We should note that the results in this article assume that $T_{ij}$ varies within each cluster. If $T_{ij}$ is a cluster-level covariate (i.e., $T_{ij} = T_i$), then, in contrast to the results reported in this paper, Lin and Carroll (2001) showed that the profile-kernel method works as usual and yields a $\sqrt{n}$ consistent and semiparametric efficient estimate of $\boldsymbol{\beta}$ if the true covariance is assumed and regular smoothing is used.

## APPENDIX A: PROOF OF RESULT 1

### A Note on Technical Conditions

It is possible to write down detailed technical conditions that would allow rigorous proofs of the results that follow for panel data. We have chosen not to do so, both in the interest of space and also because similar details have been written down by other authors in similar situations, without any real impact on statistical practice. These authors include Carroll et al. (1997), Carroll, Knickerbocker, and Wang (1995), Carroll and Wand (1991), Severini and Staniswalis (1994), and Severini and Wong (1992).

However, there is one situation for which it is easy to write down technical conditions leading to precise proofs—namely, the Gaussian linear case with constant true and working covariance matrices independent of $\boldsymbol{\beta}$. Happily, this is the problem of most interest, because all of our global conclusions have been made using this problem as an illustration.

To do this, one must first assume that, as in Carroll et al. (1995) and Severini and Staniswalis (1994), the $(T_{ij})_i$ have common compact support over $j$ and their marginal and joint densities are bounded away from 0 on this support. We assume that $h \propto n^{-\alpha}$, where $1/5 \leq \alpha \leq 1/3$. Then, using the techniques of Mack and Silverman (1982) or Marron and Härdle (1986), one can show that (A.2) holds *uniformly* in $t$. In some cases, (as in Carroll et al. 1995), it is easier to prove this by restricting attention to $(T_{ij})_j$ that fall within a proper compact subset of the common support, in which case statements of results must be modified appropriately. In either case, the Gaussian linear problem means that nonparametric regressions are standard ones and do not involve solving nonlinear equations.

We now note the other key features of the Gaussian case. For the Gaussian case, (A.3)–(A.4) are *exact*, with $\widetilde{\mathbf{X}}_i$ defined just after (A.2) being *independent* of $\boldsymbol{\beta}$. In particular, the term $o_p(1)$ in (A.3) equals 0. With the uniformity of (A.2), the calculations following (A.3)–(A.4) are then routine.

## Sketch of the Proof

To prove part a, we first assume that $\boldsymbol{\beta}$ is known and show that the asymptotic bias and variance of $\hat{\theta}(t;\boldsymbol{\beta})$ are given in (12) and (13). The proof is similar to appendix A.4 of Lin and Carroll (2000) and is hence omitted. Following that work, simple application of the Cauchy–Schwartz inequality shows that $\operatorname{var}\{\hat{\theta}(t;\boldsymbol{\beta})\}$ is minimized when $\mathbf{R}_2 = \mathbf{I}$ and is given in (14). We next study the distribution of $\hat{\theta}(t;\hat{\boldsymbol{\beta}})$ when $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent; that is, $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = O_p(1)$. We write

$$\begin{aligned}
&\sqrt{nh}\{\hat{\theta}(t;\hat{\boldsymbol{\beta}})-\theta(t)\}\\
&\quad=\sqrt{nh}\{\hat{\theta}(t;\hat{\boldsymbol{\beta}})-\hat{\theta}(t;\boldsymbol{\beta})\}+\sqrt{nh}\{\hat{\theta}(t;\boldsymbol{\beta})-\theta(t)\}\\
&\quad=\sqrt{h}\left\{\frac{\hat{\theta}(t;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right\}\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}\\
&\qquad+\sqrt{nh}\{\hat{\theta}(t;\boldsymbol{\beta})-\theta(t)\}+o_p(1), \qquad \text{(A.1)}
\end{aligned}$$

where $\hat{\theta}(t;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T = -W_2^{-1}(t)\mathbf{W}_2^x(t)+o_p(1) = O_p(1)$, where $W_2(t)$ and $\mathbf{W}_2^x(t)$ are defined in Section 4. Because $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = O_p(1)$, the first term in (A.1) is $o_p(1)$. Hence the asymptotic distribution of $\hat{\theta}(t;\hat{\boldsymbol{\beta}})$ is the same as that of $\hat{\theta}(t;\boldsymbol{\beta})$.

We now study the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. First, using part a of Result 1 and following Lin and Carroll (2000), we have

$$\begin{aligned}
\hat{\theta}(t;\boldsymbol{\beta})-\theta(t) = W_2^{-1}(t)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}^{(1)}v_{2i}^{jj}K_h(T_{ij}-t)(Y_{ij}-\mu_{ij})\\
+\frac{\theta^{(2)}(t)h^2}{2}+o_p\{n^{-1/2}\}. \qquad \text{(A.2)}
\end{aligned}$$

Define $\widetilde{\mathbf{X}}_i$ as $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij}+\partial\hat{\theta}(T_{ij};\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T = \mathbf{X}_{ij}-W_2^{-1}(T_{ij})\mathbf{W}_2^x(T_{ij})$. A linear Taylor expansion of (10) gives

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\} = \boldsymbol{D}_n^{-1}\{\sqrt{n}\mathbf{C}_n\}+o_p(1), \qquad \text{(A.3)}$$

where

$$\boldsymbol{D}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}\boldsymbol{\Delta}_i\widetilde{\mathbf{X}}_i$$

and

$$\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}[\mathbf{Y}_i-\boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta}+\hat{\boldsymbol{\theta}}(\mathbf{T}_i;\boldsymbol{\beta})\}]. \qquad \text{(A.4)}$$

Denote $\boldsymbol{D} = \lim_{n\to\infty}\boldsymbol{D}_n = E(\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}})$. Simple calculations show that $\boldsymbol{C}_n$ can be expanded as $\boldsymbol{C}_n = \boldsymbol{C}_{1n}-\boldsymbol{C}_{2n}+o_p(1)$, where, denoting $\boldsymbol{\mu}_i = \boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta}+\theta(\boldsymbol{T}_i)\}$ and $\mathbf{Z}_{1i}^T = \widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}$,

$$\boldsymbol{C}_{1n} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{1i}^T(\mathbf{Y}_i-\boldsymbol{\mu}_i)$$

and

$$\boldsymbol{C}_{2n} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}\boldsymbol{\Delta}_i\{\hat{\boldsymbol{\theta}}(\boldsymbol{T}_i;\boldsymbol{\beta})-\boldsymbol{\theta}(\boldsymbol{T}_i)\}.$$

Obtaining asymptotic distribution of $\sqrt{n}\boldsymbol{C}_{1n}$ is simple. Now examine the distribution of $\sqrt{n}\boldsymbol{C}_{2n}$. Using the Taylor expansion (A.2), we have

$$\begin{aligned}
\boldsymbol{C}_{2n} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}\{\hat{\theta}(T_{ik};\boldsymbol{\beta})-\theta(T_{ik})\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}\left\{\left[W_2^{-1}(T_{ik})\frac{1}{n}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}v_{2i'}^{j'j'}\right.\right.\\
&\qquad\left.\left.\times K_h(T_{i'j'}-T_{ik})(Y_{i'j'}-\mu_{i'j'})\right]+\frac{h^2}{2}\theta^{(2)}(T_{ik})\right\}+o_p(1)\\
&= \frac{1}{n}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}v_{2i'}^{j'j'}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}W_2^{-1}(T_{ik})\right.\\
&\qquad\left.\times K_h(T_{ik}-T_{i'j'})\right\}(Y_{i'j'}-\mu_{i'j'})\\
&\quad+\frac{h^2}{2}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}\theta^{(2)}(T_{ik})\right\}+o_p(1)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{Z}_{2ij}(Y_{ij}-\mu_{ij})+\frac{h^2}{2}\sum_{j=1}^{m}\sum_{k=1}^{m}\\
&\quad\times E\left\{\widetilde{\mathbf{X}}_j\mu_j^{(1)}v_1^{jk}\mu_k^{(1)}\theta^{(2)}(T_k)\right\}+o_p(1)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{2i}^T(\mathbf{Y}_i-\boldsymbol{\mu}_i)+\frac{h^2}{2}E\{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\theta^{(2)}(\boldsymbol{T})\}+o_p(1),
\end{aligned}$$

where $\mathbf{Z}_{2i} = \{\mathbf{Z}_{2i1},\ldots,\mathbf{Z}_{2im}\}^T$ and

$$\mathbf{Z}_{2ij} = \mu_{ij}^{(1)}v_{2i}^{jj}\left\{\sum_{k=1}^{m}\sum_{l=1}^{m}E\left(\widetilde{\mathbf{X}}_k\mu_k^{(1)}v_1^{kl}\mu_l^{(1)}\middle|T_l=T_{ij}\right)\right\}W_2^{-1}(T_{ij})f_j(T_{ij}).$$

It follows that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \boldsymbol{D}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{Z}_{1i}-\mathbf{Z}_{2i})(\mathbf{Y}_i-\boldsymbol{\mu}_i)\\
+\sqrt{nh^4}\boldsymbol{b}(\boldsymbol{\beta},\theta)/2+o_p(1), \qquad \text{(A.5)}
\end{aligned}$$

where the bias term $\boldsymbol{b}(\boldsymbol{\beta},\theta) = \boldsymbol{D}^{-1}E\{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\theta^{(2)}(\boldsymbol{T})\}$. Equivalently,

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}-h^2b(\boldsymbol{\beta},\theta)/2\} \to N(0,\mathbf{V}_{\boldsymbol{\beta}}),$$

where $\mathbf{V}_{\boldsymbol{\beta}} = \boldsymbol{D}^{-1}E\{(\mathbf{Z}_1-\mathbf{Z}_2)^T\boldsymbol{\Sigma}(\mathbf{Z}_1-\mathbf{Z}_2)\}\boldsymbol{D}^{-1}$ with $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{Y}|\mathbf{X},\boldsymbol{T})$.

One can see easily that the bias term $\boldsymbol{b}(\boldsymbol{\beta},\theta)$ in (A.5) is generally nonzero. Under conventional asymptotics, $n\to\infty$, $h\to 0$, and $nh\to\infty$, to obtain a $\sqrt{n}$ consistent estimate of $\boldsymbol{\beta}$, one must identify working correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ to make the bias term

$b(\boldsymbol{\beta}, \theta) = 0$. Simple calculations show that this requires assuming working independence $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$, and the same marginal joint density of $(\mathbf{X}_j, T_j)$; that is, $f_j(\mathbf{X}_j, T_j) = f(\mathbf{X}_j, T_j)$. Under these two assumptions,

$$\boldsymbol{b}(\boldsymbol{\beta}, \theta) = \boldsymbol{D}^{-1} E\left\{\sum_{j=1}^{m} E\left[\widetilde{\mathbf{X}}_j \left\{\mu_j^{(1)}\right\}^2 \sigma_{jj}^{-1} | T_j\right] \theta^{(2)}(T_j)\right\},$$

where $\widetilde{\mathbf{X}}_j = \mathbf{X}_j - E[\{\mu_j^{(1)}\}^2 \sigma_{jj}^{-1} \mathbf{X}_j | T_j])^{-1} E[\{\mu_j^{(1)}\}^2 \sigma_{jj}^{-1} | T_j])^{-1}$. One can see easily that $E[\widetilde{\mathbf{X}}_j \{\mu_j^{(1)}\}^2 \sigma_{jj}^{-1} | T_j] = 0$. It follows that $\boldsymbol{b}(\boldsymbol{\beta}, \theta) = 0$. Similar calculations show that $\boldsymbol{Z}_{2i} = 0$. This implies $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \widetilde{\mathbf{V}}_{\boldsymbol{\beta}})$, where $\widetilde{\mathbf{V}}_{\boldsymbol{\beta}}$ is given in part c of Result 1 and can be estimated using a sandwich estimator.

For any nonidentity working correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, even when $\mathbf{R}_1$ and $\mathbf{R}_2$ are the true correlation matrices, under the foregoing conventional asymptotics [e.g., with $h$ chosen using cross-validation; i.e., $h = O(n^{-1/5})$], the bias term $\sqrt{nh^4}\boldsymbol{b}(\boldsymbol{\beta}, \theta) \to \infty$. This means that $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ inconsistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \infty$. Furthermore, $\boldsymbol{Z}_{2i} \neq 0$, implying that the standard sandwich estimator will be an inconsistent estimator of $\mathbf{V}_{\boldsymbol{\beta}}$, because it estimates $\boldsymbol{D}^{-1} E\{\boldsymbol{Z}_1 \boldsymbol{\Sigma} \boldsymbol{Z}_1^T\} \boldsymbol{D}^{-1}$ and ignores the nonzero term $\boldsymbol{Z}_2$. If one undersmooths $\theta(t)$ by letting $nh^4 \to 0$, then the bias term $\sqrt{nh^4}\boldsymbol{b}(\boldsymbol{\beta}, \theta) \to 0$, and $\hat{\boldsymbol{\beta}}$ will be $\sqrt{n}$ consistent for arbitrary working correlation matrices $(\mathbf{R}_1, \mathbf{R}_2)$.

## APPENDIX B: SEMIPARAMETRIC EFFICIENT SCORE

We focus on the case where $\mathbf{X}_j$ and $\boldsymbol{\beta}$ are scalars (i.e., $p = 1$) and briefly discuss how to extend this result to the case where $\mathbf{X}_j$ and $\boldsymbol{\beta}$ are vectors. Let $f(\beta, \theta)$ denote the multivariate normal density of $\mathbf{Y} \sim N\{\mathbf{X}\beta + \theta(\boldsymbol{T}), \mathbf{V}\}$ , where $\theta(\boldsymbol{T}) = \{\theta(T_1), \ldots, \theta(T_m)\}^T$. Following Begun, Hall, Huang, and Wellner (1983), we first calculate the Hellinger derivative with respect to $\theta(\cdot)$. Suppose that the sequence $\{\theta_n(t)\}$ satisfies $\sqrt{n}\{\theta_n(t) - \theta(t)\} - \varphi(t) \to 0$ as $n \to \infty$ for any given continuous function $\varphi(t)$. The Hellinger derivative $A\varphi(\cdot)$ with respect to $\theta(\cdot)$ is defined as

$$2n^{1/2}\left\{\frac{f^{1/2}(\beta, \theta_n) - f^{1/2}(\beta, \theta)}{f^{1/2}(\beta, \theta)}\right\} - \frac{2A\varphi}{f^{1/2}(\beta, \theta)} \to 0, \quad \text{as} \quad n \to \infty,$$

where $A$ denotes a linear operator. Denote $f_n = f\{\beta, \boldsymbol{\theta}_n\}$ and $f = f\{\beta, \boldsymbol{\theta}\}$, where $\boldsymbol{\theta}_n = \{\theta_n(T_1), \ldots, \theta_n(T_m)\}^T$ and $\boldsymbol{\theta} = \{\theta(T_1), \ldots, \theta(T_m)\}^T$. Let $\ell_n = \log f_n$ and $\ell = \log f$. A simple Taylor expansion shows that

$$\begin{aligned} 2\sqrt{n}\left\{\frac{\sqrt{f_n} - \sqrt{f}}{\sqrt{f}}\right\} &= \sqrt{n}\left\{\frac{f_n - f}{f}\right\} + o_p(1) \\ &= \sqrt{n}\{\ell_n - \ell\} + o_p(1) \\ &= \frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\left\{\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})\right\} + o_p(1) \\ &= \varphi(T)^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\} + o_p(1). \end{aligned}$$

It follows that $2A\varphi(\boldsymbol{T})/f^{1/2} = \varphi(\boldsymbol{T})^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\}$. Let $\dot{\ell}_\beta = \partial \ell / \partial \beta = \mathbf{X}^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\}$. Then the semiparametric efficient score $\dot{\ell}_\beta^{\natural}$ of $\beta$ is

$$\begin{aligned} \dot{\ell}_\beta^{\natural} &= \dot{\ell}_\beta - 2A\varphi_*(\boldsymbol{T})/f^{1/2}(\beta, \theta) \\ &= \{\mathbf{X} - \varphi_*(\boldsymbol{T})\}^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\}, \end{aligned}$$

which is (15), where $\varphi_*(t)$ satisfies

$$E\left\{\dot{\ell}_\beta^{\natural}[A\varphi(\boldsymbol{T})/f^{1/2}]\right\} = E\{[\mathbf{X} - \varphi_*(\boldsymbol{T})]^T \mathbf{V}^{-1} \varphi(\boldsymbol{T})\} = 0 \qquad \text{(B.1)}$$

for all functions $\varphi(\boldsymbol{T}) = \{\varphi(T_1), \ldots, \varphi(T_m)\}^T$, where $\varphi_*(\boldsymbol{T}) = \{\varphi_*(T_1), \ldots, \varphi_*(T_m)\}^T$. The semiparametric efficiency bound of $\boldsymbol{\beta}$ is $E\{[\dot{\ell}_\beta^{\natural}]^2\}$. Equation (B.1) can be written as

$$\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E\{[X_j - \varphi_*(T_j)]\varphi(T_k)\} = \sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E[E\{E[X_j - \varphi_*(T_j)|T_k]\}\varphi(T_k)] = 0.$$

Simple calculations show that this equation can be written as

$$\int\left[\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk}\{E[X_j - \varphi_*(T_j)|T_k = t]\} f_k(t)\right]\varphi(t)dt = 0$$

for any $\varphi(t)$. It follows that $\varphi_*(t)$ must solve $\sum_{j=1}^{m}\sum_{k=1}^{m} \times v^{jk}\{E[X_j - \varphi_*(T_j)|T_k = t]\} f_k(t) = 0$, which is (16).

To extend the results to the case where $\mathbf{X}_j$ and $\boldsymbol{\beta}$ are vectors, we need to find $\varphi_*(t)$ for each component of $\mathbf{X}_j$ using (16) (Begun et al. 1983). Specifically, we calculate $\boldsymbol{\varphi}_*(t) = \{\varphi_{*1}(t), \ldots, \varphi_{*p}(t)\}^T$, where, letting $X_{jr}$ denote the $r$th component of $\mathbf{X}_j$, $\varphi_{*r}(t)$ solves

$$\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E[\{X_{jr} - \varphi_{*r}(T_j)\}|T_k = t] f_k(t) = 0.$$

Hence semiparametric efficient score of $\boldsymbol{\beta}$ is given by (15) and (16).

## APPENDIX C: COMPUTATION ASSUMING WORKING INDEPENDENCE

In this section we assume working independence ($\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$) in the profile-kernel estimating equations (8) and (10), and discuss the use of the Fisher scoring algorithm to solve for $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}(t)$, where $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent. Specifically, under working independence ($\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$), (8) and (10) are solved in the following steps:

1. Assume a parametric function for $\theta(t)$, [e.g., $\theta(t) = \alpha_0 + \alpha_1 t$], and fit a parametric generalized linear model, $g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \alpha_0 + \alpha_1 T_{ij}$, to obtain an initial value of $\boldsymbol{\beta}$.
2. Given the value of $\boldsymbol{\beta}$, use the Fisher scoring algorithm to solve (8) (with $\mathbf{R}_2 = \mathbf{I}$) for $t = T_{11}, \ldots, T_{nm_n}$. This gives $\{\hat{\theta}(T_{11}; \boldsymbol{\beta}), \ldots, \hat{\theta}(T_{nm_n}; \boldsymbol{\beta})\}$.
3. Update $\boldsymbol{\beta}$ using the one-step Fisher scoring algorithm to solve (10) (with $\mathbf{R}_1 = \mathbf{I}$) given the $\hat{\theta}(T_{ij}; \boldsymbol{\beta})$ in step 2.
4. Iterate between steps 2 and 3 until convergence.

In step 2, it can be easily shown that the Fisher scoring algorithm updates $\hat{\boldsymbol{\alpha}}$ by

$$\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{T}_{ij}(t) W_{ij}(t) K_h(T_{ij} - t) \boldsymbol{T}_{ij}(t)^T\right\}\hat{\boldsymbol{\alpha}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{T}_{ij}(t) W_{ij}(t) K_h(T_{ij} - t) y_{ij}(t),$$

where $W_{ij}(t) = \{\mu_{ij}^{(1)}(t)\}^2 V_{ij}^{-1}(t)$ is the generalized linear model working weight and $y_{ij}(t) = \boldsymbol{T}_{ij}(t)^T \boldsymbol{\alpha} + \mu_{ij}^{(1)}(t)\{Y_{ij} - \mu_{ij}(t)\}$ is the generalized linear model working vector.

In step 3, the one-step Fisher scoring algorithm updates $\boldsymbol{\beta}$ using the weighted least squares,

$$\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \widetilde{\mathbf{X}}_{ij} W_{ij} \widetilde{\mathbf{X}}_{ij}^T\right\}\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \widetilde{\mathbf{X}}_{ij} W_{ij} y_{ij},$$

where $W_{ij} = \{\mu_{ij}^{(1)}\}^2 V^{-1}(\mu_{ij})$ is the working weight, $y_{ij} = \widetilde{\mathbf{X}}_{ij}^T\boldsymbol{\beta} + \mu_{ij}^{(1)}(Y_{ij} - \mu_{ij})$ is the working vector, $\mu_{ij} = \mu\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \hat{\theta}(T_{ij};\boldsymbol{\beta})\}$, and $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij} + \partial\hat{\theta}(T_{ij};\boldsymbol{\beta}))\partial\boldsymbol{\beta}$. To calculate $\widetilde{\mathbf{X}}_{ij}$, we need to construct a consistent estimate of $\partial\hat{\theta}(t;\boldsymbol{\beta}))\partial\boldsymbol{\beta}$. Using the results in Section 4, we can easily see that a consistent estimator of $\partial\hat{\theta}(t;\boldsymbol{\beta}))\partial\boldsymbol{\beta}$ is

$$-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} W_{ij}(t)K_h(T_{ij}-t)\mathbf{X}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} W_{ij}(t)K_h(T_{ij}-t)}.$$

The covariance estimators of $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}(t)$ at convergence are sandwich estimators given by

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \left\{\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\mathbf{W}_i\widetilde{\mathbf{X}}_i\right\}^{-1}\left\{\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\boldsymbol{\Sigma}_{id}^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}_i) \times (\mathbf{Y}_i-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_{id}^{-1}\boldsymbol{\Delta}_i\widetilde{\mathbf{X}}_i\right\}\left\{\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\mathbf{W}_i\widetilde{\mathbf{X}}_i\right\}^{-1}$$

and

$$\operatorname{cov}\{\hat{\theta}(t)\} = \boldsymbol{e}_1^T\boldsymbol{\Omega}_1^{-1}(t)\boldsymbol{\Omega}_2(t)\boldsymbol{\Omega}_1^{-1}(t)\boldsymbol{e}_1,$$

where $\boldsymbol{e}_1 = (1,0)^T$, $\boldsymbol{\Delta}_i = \operatorname{diag}\{\mu_{ij}^{(1)}\}$, and $\boldsymbol{\Sigma}_{id} = \operatorname{diag}\{V_{ij}\}$ and all are evaluated at $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$ and

$$\boldsymbol{\Omega}_1(t) = \sum_{i=1}^{n}\mathbf{T}_i^T(t)\mathbf{W}_i(t)\mathbf{K}_{ih}(t)T_i(t)$$

and

$$\boldsymbol{\Omega}_2(t) = \sum_{i=1}^{n}\boldsymbol{T}_i^T(t)\boldsymbol{\Delta}_i(t)\boldsymbol{\Sigma}_{id}^{-1}\mathbf{K}_{ih}(t)[\mathbf{Y}_i-\boldsymbol{\mu}_i(t)] \times [\mathbf{Y}_i-\boldsymbol{\mu}_i(t)]^T\mathbf{K}_{ih}(t)\boldsymbol{\Sigma}_{id}^{-1}\boldsymbol{\Delta}_i(t)\boldsymbol{T}_i(t).$$

Estimation of $\theta(t)$ requires choosing the bandwidth parameter $h$. One approach is to use cross-validation by deleting one cluster at a time. Another approach is to extend Ruppert's (1997) empirical bias bandwidth selection (EBBS) method to clustered data. We use the EBBS method to choose $h$ for given $\boldsymbol{\beta}$. (For details, see Lin and Carroll 2000.)

*[Received February 2000. Revised January 2001.]*

## REFERENCES


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

Bickel, P. J., Klaassen, C. J., Ritov, Y., and Wellner, J. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Brockmann, M., Gasser, T., and Herrmann, E. (1993), "Locally Adaptive Bandwidth Choice for Kernel Regression Estimators," *Journal of the American Statistical Association*, 88, 1302–1309.

Bronshtein, J. N., and Semendyayev, K. A. (1985), *Handbook of Mathematics*, New York: Van Nostrand Reinhold.

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

Carroll, R. J., Knickerbocker, R. K., and Wang, C. Y. (1995), "Dimension Reduction in Semiparametric Measurement Error Models," *The Annals of Statistics*, 23, 161–181.

Carroll, R. J., and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error," *Journal of the Royal Statistical Society*, Ser. B, 53, 573–585.

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

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

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.

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. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error," *Journal of the American Statistical Association*, 95, 520–534.

——— (2001), "Semiparametric Regression For Clustered Data," *Biometrika*, in press.

Mack, Y., and Silverman, B. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 60, 405–415.

Marron, J. S., and Härdle, W. (1986), "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91–113.

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

Ramsay, J. O., and Dalzell, C. J. (1991), "Some Tools for Functional Data Analysis" (with discussions), *Journal of the Royal Statistical Society*, Ser. B, 53, 539–572.

Rice, J. (1986), "Convergence Rates for Partially Splined Models," *Statistical and Probability Letters*, 4, 203–208.

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

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

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

Severini, T. A., and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

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.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semi-Parametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.