

# Semiparametric Estimation in General Repeated Measures Problems

Xihong Lin

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.  
xlin@sph.umich.edu

Raymond J. Carroll

Department of Statistics, Texas A&M University, College Station TX 77843-3143, U.S.A.  
carroll@stat.tamu.edu

## Summary

This paper considers a wide class of semiparametric problems with a parametric part for some covariate effects and repeated evaluations of a nonparametric function. Special cases in our approach include marginal models for longitudinal/clustered data, conditional logistic regression for matched case-control studies, multivariate measurement error models, generalized linear mixed models with a semiparametric component, and many others. We propose profile-kernel and backfitting estimation methods for these problems, derive their asymptotic distributions, and show that in likelihood problems the methods are semiparametric efficient. While generally not true, with our methods profiling and backfitting are asymptotically equivalent. We also consider pseudolikelihood methods where some nuisance parameters are estimated from a different algorithm. The proposed methods are evaluated using simulation studies and applied to the Kenya hemoglobin data.

**Some key words:** Clustered/longitudinal data; Generalized estimating equations; Generalized linear mixed models; Kernel method; Marginal models; Measurement error; Nonparametric regression; Partially linear model; Profile method; Semiparametric information bound; Semiparametric efficient score; Time dependent covariate.

**Short title:** Semiparametric Repeated Measures Regression

# 1 Introduction

This paper considers a wide class of semiparametric problems with some covariates modelled parametrically and repeated evaluations of a nonparametric function of a covariate. We propose profile-kernel and backfitting estimation methods for these problems, derive their asymptotic distributions, and show that in likelihood problems the methods are semiparametric efficient.

To get some sense of the generality of our approach, consider the following examples. The first four are new, in the sense that neither the semiparametric efficient score function nor a constructive method of estimation and inference that achieve efficiency are known. In contrast, the fifth example has a large literature.

**Example 1:** One of the most common designs in epidemiology is the matched case-control study, a design that is attracting considerable interest in genetic epidemiology, see for example Schaid (1999). Matched case-control studies consist of groups that have discordant responses. Thus, in the 1-1 matched study, one considers matched pairs of subjects, with disease responses  $(Y_{i1}, Y_{i2})$  that are constrained to be discordant, so that  $Y_{i1} + Y_{i2} = 1$ . The underlying prospective semiparametric logistic regression model is that  $\text{pr}(Y_{ij} = 1 | X_{ij}, Z_{ij}) = H\{b_i + X_{ij}^T \beta_0 + \theta_0(Z_{ij})\}$ , where  $H(v) = \{1 + \exp(-v)\}^{-1}$  is the logistic distribution function,  $b_i$  is a random effect depending on the matched set,  $X_{ij}$  is a covariate vector whose effect is modelled parametrically and  $Z_{ij}$  is a scalar covariate whose effect is modelled using a nonparametric smooth function  $\theta_0(\bullet)$ . Let  $\tilde{X}_i = (X_{i1}, X_{i2})$  and  $\tilde{Z}_i = (Z_{i1}, Z_{i2})$ . Because the data are constrained to be discordant, and one does not want to model the stratum effects  $b_i$ , inference is based on the conditional likelihood function

$$\text{pr}(Y_{i1} = 1, Y_{i2} = 0 | \tilde{X}_i, \tilde{Z}_i, Y_{i1} + Y_{i2} = 1) = H[(X_{i1} - X_{i2})\beta_0 + \{\theta_0(Z_{i1}) - \theta_0(Z_{i2})\}]. \quad (1)$$

Note that in (1) the stratum effects have been eliminated, and that in the likelihood  $\theta_0(\bullet)$  is evaluated twice at different values of  $Z$ . In more complex matched studies,  $\theta_0(\bullet)$  is evaluated more than twice, e.g., the 1- $M$  matched design.

**Example 2:** Hafner (1998) and Carroll, et al. (2002) studied  $Y_i = \sum_{j=1}^m \beta_0^{j-1} \theta_0(Z_{ij}) + \epsilon_i$ , a model that arises in finance. The algorithm proposed by Carroll, et al. (2002) for this case is extremely unwieldy and difficult to implement, because it is based on an integration estimator (Linton and Nielson, 1995). Our methodology in this case is far easier to implement, and has the advantage of being semiparametric efficient in the Gaussian case.

**Example 3:** Generalized linear mixed models (Breslow and Clayton, 1993) have become popular as a means of quantifying and understanding variability. The simplest such model for binary data is the random intercept model  $\text{pr}(Y_{ij} = 1 | X_{ij}, Z_{ij}, b_i) = \mu\{X_{ij}^T \beta_0 + \theta_0(Z_{ij}) + b_i\}$ , where  $\mu(\bullet)$  is the inverse of a link function and  $b_i = \text{Normal}(0, \sigma_0^2)$ . Here the variance component  $\sigma_0^2$  may be of interest in itself, and may in some cases depend on components of  $X$  such as gender, see Heagerty and Kurland (2001) for an example.

**Example 4:** As discussed in a data example in Section 5.1.2, consider problems in which a family has  $m$  children, each of whom have a baseline measure  $Z_{ij}$  for  $j = 1, \dots, m$ , but for whom there are repeated measures  $Y_{ijk}$  over time for  $k = 1, \dots, K$  and a possible repeated time-varying covariate  $X_{ijk}$ . A reasonable marginal model for the  $Y_{ijk}$  is that their means are  $\mu\{X_{ijk}^T \beta_0 + \theta_0(Z_{ij})\}$  for a known inverse link function  $\mu(\bullet)$ , and a covariance matrix  $\Sigma$  reflecting the structure of the problem. In this case, note that the function  $\theta_0(\bullet)$  is evaluated  $m$  times for different children per family.

**Example 5:** Consider a repeated measures Gaussian partially linear problem where for the  $i^{\text{th}}$  subject responses  $\tilde{Y}_i = (Y_{i1}, \dots, Y_{im})^T$  and predictors  $\tilde{X}_i = (X_{i1}, \dots, X_{im})^T$  and  $\tilde{Z}_i = (Z_{i1}, \dots, Z_{im})^T$  are observed, with  $Z_{ij}$  scalar. The basic model is that for a known function  $\mu(\bullet)$  and a true but unknown function  $\theta_0(z)$ ,

$$Y_{ij} = \mu\{X_{ij}^T \beta_0 + \theta_0(Z_{ij})\} + \epsilon_{ij}, \quad (2)$$

where given  $(\tilde{X}_i, \tilde{Z}_i)$ ,  $\tilde{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{im})^T$  has mean zero and covariance matrix  $\Sigma(\tau_0)$  for a parameter  $\tau_0$ . Note that the function  $\theta_0(\bullet)$  is evaluated repeatedly, and thus this problem is very much different from the standard partially linear model (Severini and Staniswalis, 1994). This problem has a large literature, with many kernel-based methods (Zeger and Diggle, 1994; Hoover, et al., 1998; Lin and Ying, 2001; Wu and Zhang, 2002; and many others), all of them estimating  $\theta_0(\bullet)$  while ignoring the correlation structure. Lin and Carroll (2000, 2001) and Fan and Li (2004) made an effort to incorporate the correlation structure in the estimation procedure within the traditional kernel framework. However, Lin and Carroll (2000) showed that the optimal estimator of  $\theta_0(\bullet)$  within the standard kernel framework requires ignoring the correlation. There is also an extensive spline-based literature (Wild and Yee, 1996; Zhang, et al, 1998; Wang, 1998; Rice and Wu, 2001). Fixing  $\Sigma(\tau_0)$  and pretending normality, Wang, et al. (2004) developed kernel-based consistent and asymptotically normal estimators for  $\beta_0$ : these are semiparametric efficient when  $\tilde{\epsilon}_i$  is actually Gaussian.

These examples can be placed into a common framework. There is a *criterion function*

$\mathcal{L}(\tilde{Y}, \tilde{Z}, \tilde{\eta}, \mathcal{B})$ , where  $\tilde{\eta}$  has  $m$ -components representing  $\theta(Z_1), \dots, \theta(Z_m)$ , and  $\mathcal{B}$  is a vector of parameters. For true values  $\tilde{\eta}_0$  and  $\mathcal{B}_0$ , the criterion function satisfies

$$0 = E\{\{\partial \mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}_0, \mathcal{B}_0) / \partial(\tilde{\eta}_0, \mathcal{B}_0)\} | \tilde{X}, \tilde{Z}\}. \quad (3)$$

For example, consider the model given in (2). Here,  $\mathcal{B}_0 = (\beta_0, \tau_0)$  and the criterion function is the Gaussian loglikelihood  $-1/2 \log[\det\{\Sigma(\tau_0)\}] - (1/2)(\tilde{Y} - \tilde{X}\beta_0 - \tilde{\eta}_0)^\top \Sigma^{-1}(\tau_0)(\tilde{Y} - \tilde{X}\beta_0 - \tilde{\eta}_0)$ . The criterion function in Example 1 is given in (1), while Examples 2-4 also have explicit forms.

In this paper, we show how to compute efficient estimators of the nonparametric component  $\theta_0(\bullet)$  for problems with and without the parametric component  $\mathcal{B}_0$ . The method is defined in Section 2, and is based on a likelihood-type generalization of the basic method of Wang (2003) using kernel methods. The methods are applicable to likelihood and non-likelihood problems, the only constraint being that (3) holds.

In Section 3 we take up estimation of the parameter  $\mathcal{B}$ . In this context, we derive two general methods, one incorporating profile-likelihood ideas and the other based on the often easier to compute backfitting algorithm. We show that in our case, using the smoother of Section 2, profiling and backfitting have identical limit distributions. The general folklore of course is that backfitting and profiling are in general asymptotically equivalent, independent of the method of smoothing, but in general this is not the case (Hu, et al., 2004). However, our use of an efficient smoother allows us to show that backfitting and profiling are asymptotically equivalent. It should be noted that undersmoothing of the nonparametric function is required by backfitting but not required by profiling. In this section, we also describe the semiparametric efficient score function when  $\mathcal{L}(\bullet)$  is a likelihood function, and show in our case that our method achieves the semiparametric information bound.

In many problems, there are nuisance parameters that can be estimated relatively conveniently by alternative means. In the example considered by Wang, et al. (2004), the covariance matrix  $\Sigma(\tau_0)$  depends on a parameter  $\tau_0$ . The parameter  $\tau_0$  is conveniently estimated by the simple device of ignoring the correlation of the data, forming residuals from the fit, and then using method of moments. This is a pseudolikelihood approach. In Section 4, we derive the limiting distribution of the pseudolikelihood estimator in the general case.

Section 5 first describes Example 4 in detail. We illustrate Example 4 using the Kenya hemoglobin data and a simulation study. The second case considered in 5 is a multivariate measurement error problem. The formulation of the measurement error model is new even in the parametric measurement error model literature. Sketches of the technical arguments are given in an appendix.

## 2 The Nonparametric Case

Before describing methods for the general semiparametric problem, we describe methods when there is no parametric component, a problem of interest in its own right. In the nonparametric case, the criterion function is  $\mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}) = \mathcal{L}\{\tilde{Y}, \tilde{X}, \theta(Z_1), \dots, \theta(Z_m)\}$ . Define  $\mathcal{L}_{j\theta}(\bullet) = \partial \mathcal{L}(\tilde{Y}, \tilde{X}, \eta_1, \dots, \eta_m) / \partial \eta_j$  and  $\mathcal{L}_{jk\theta}(\bullet) = \partial^2 \mathcal{L}(\tilde{Y}, \tilde{X}, \eta_1, \dots, \eta_m) / (\partial \eta_j \partial \eta_k)$  ( $j, k = 1, \dots, m$ ). We assume that  $0 = E[\mathcal{L}_{j\theta}\{\tilde{Y}, \tilde{X}, \theta(Z_1), \dots, \theta(Z_m)\} | \tilde{X}, \tilde{Z}]$ . Let  $K(\bullet)$  be a symmetric density function with variance 1.0, and define  $G_{ij}(z, h) = \{1, (Z_{ij} - z)/h\}$ . Let  $f_j(z)$  be the marginal density of  $Z_{ij}$ .

We propose to estimate  $\theta(\bullet)$  by solving the following kernel estimating equation

$$0 = \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) \times \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}), \dots, \hat{\theta}(z) + \hat{\theta}^{(1)}(z)(Z_{ij} - z), \dots, \hat{\theta}(Z_{im}) \right\}, \quad (4)$$

where  $\hat{\theta}^{(1)}(z)$  denotes the first derivative of  $\hat{\theta}(z)$ . Following Wang (2003), we propose to solve the kernel estimating equation (4) for  $\hat{\theta}(z)$  in the following iterative fashion. Suppose that the current estimate of  $\theta(\bullet)$  at the  $(\ell - 1)^{\text{st}}$  step is  $\hat{\theta}_{[\ell-1]}(\bullet)$ . Then  $\hat{\theta}_{[\ell]}(z) = \hat{\alpha}_0$ , where  $(\hat{\alpha}_0, \hat{\alpha}_1)$  solve

$$0 = \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) \times \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[\ell-1]}(Z_{i1}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \hat{\theta}_{[\ell-1]}(Z_{im}) \right\}. \quad (5)$$

At convergence,  $\hat{\theta}(z)$  solves the kernel estimating equation (4). In Gaussian cases such as in Examples 1-2, iteration is actually not needed, with explicit solutions being available, see Lin, et al. (2004) for Example 1, and see also Section 5.1 for another example. Define  $\mathcal{L}(\bullet) = \mathcal{L}\{\tilde{Y}, \tilde{X}, \theta(Z_1), \dots, \theta(Z_m)\}$ , and similarly for its derivatives. Make the definitions  $\Omega(z) = \sum_{j=1}^m f_j(z) E\{\mathcal{L}_{j\theta}(\bullet) | Z_j = z\}$  and

$$\begin{aligned} \mathcal{A}(B, z_1, z_2) &= \sum_{j=1}^m \sum_{k \neq j}^m f_j(z_1) E\{\mathcal{L}_{jk\theta}(\bullet) B(Z_k, z_2) / \Omega(Z_k) | Z_j = z_1\}; \\ \mathcal{Q}(z_1, z_2) &= \sum_{j=1}^m \sum_{k \neq j}^m f_{jk}(z_1, z_2) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z_1, Z_k = z_2\} / \Omega(z_2); \\ \Lambda(g, z) &= \sum_{j=1}^m \sum_{k \neq j}^m f_j(z) E\{\mathcal{L}_{jk\theta}(\bullet) g(Z_k) | Z_j = z\} / \Omega(z), \end{aligned}$$

where  $f_j(z)$  is the density of  $Z_j$  and  $f_{jk}(z_1, z_2)$  is the bivariate density of  $(Z_j, Z_k)$ . Let  $\mathcal{G}(z_1, z_2)$  and  $b(z)$  be the solutions to

$$\mathcal{G}(z_1, z_2) = \mathcal{Q}(z_1, z_2) - \mathcal{A}(\mathcal{G}, z_1, z_2); \quad (6)$$

$$b(z) = \theta^{(2)}(z) - \Lambda(b, z). \quad (7)$$

**Result #1: Expansion for the Nonparametric Part** Suppose that the  $Z_{ij}$  have support on a compact set and that their joint and marginal densities are bounded away from zero on that set. Assume that the algorithm converges to a unique solution and that equations (6) and (7) have unique solutions. Let the bandwidth sequence satisfy  $nh^2 \rightarrow \infty$  and  $nh^6 \rightarrow 0$ . Let  $\phi = \int z^2 K(z) dz$ . Denote by  $\theta_0(z)$  the true function. Then, at convergence,

$$\begin{aligned} \hat{\theta}(z) - \theta_0(z) &= (h^2/2)\phi b(z) - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \epsilon_{ij} / \Omega(z) \\ &\quad + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} \mathcal{G}(z, Z_{ij}) / \Omega(z) + o_p(n^{-1/2}), \end{aligned} \quad (8)$$

where  $\epsilon_{ij} = \mathcal{L}_{j\theta}\{\tilde{Y}_i, \tilde{X}_i, \theta_0(Z_{i1}), \dots, \theta_0(Z_{im})\}$ . Thus, the asymptotic bias and variance of  $\hat{\theta}(z)$  are

$$E\{\hat{\theta}(z)\} - \theta_0(z) = (h^2/2)\phi b(z) + o(h^2); \quad (9)$$

$$\text{var}\{\hat{\theta}(z)\} = \frac{1}{nh} \frac{\psi}{\Omega^2(z)} \sum_{j=1}^m E(D_{jj} | Z_j = z) f_j(z) + o\{(nh)^{-1}\}, \quad (10)$$

where  $\psi = \int K^2(s) ds$  and  $D_{jj}$  is the  $j^{\text{th}}$  diagonal element of  $\text{cov}(\tilde{\epsilon}_i | \tilde{X}_i, \tilde{Z}_i)$ , where  $\tilde{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{im})^T$ .

**Remark 1:** Equations (8)-(10) agree with the results of Wang (2003) in the special cases considered by her. In (8), since the first two terms are of order  $O_p\{h^2 + (nh)^{-1/2}\}$  while the third is of order  $O_p(n^{-1/2})$ , the first two terms dominate.

**Remark 2:** Note that (9) has design-density dependent bias. It is possible to remove this. Suppose the algorithm is run with an undersmoothing bandwidth  $h_1 = o(n^{-1/4})$ , thus obtaining  $\hat{\theta}(z, h_1)$  at convergence. Let  $\hat{\theta}_{os}(z, h)$  be the estimator defined by doing one step of the iteration from  $\hat{\theta}(z, h_1)$ , but now with bandwidth  $h$ , where  $h/h_1 \rightarrow 0$  as  $n \rightarrow \infty$ . Then (8) still holds except that the bias term  $(h^2/2)\phi b(z)$  is replaced by  $(h^2/2)\phi \theta^{(2)}(z)$ . The proof of this argument is a routine application of Lemma A.1 and equation (A.1) in the Appendix, starting from the expansion (8).

### 3 The Semiparametric Case: Methods and Results

In this section, we formulate the profile-kernel and backfitting estimation methods for  $\mathcal{B}_0$  in the semiparametric model  $\mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}_0, \mathcal{B}_0)$ , state their asymptotic distributions and show that when the criterion function  $\mathcal{L}(\bullet)$  is a loglikelihood function conditional on  $(\tilde{Z}, \tilde{X})$ , our method achieves the semiparametric information bound.

### 3.1 Estimation: Profile-Kernel and Backfitting Methods

To estimate  $\mathcal{B}$ , we propose profile-kernel and backfitting methods. For any  $\mathcal{B}$ , we first obtain the modified kernel estimate of  $\hat{\theta}(z, \mathcal{B})$  and its first derivative  $\hat{\theta}^{(1)}(z, \mathcal{B})$  with respect to  $z$  by solving

$$0 = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) \times \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}(z, \mathcal{B}) + h\hat{\theta}^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h, \dots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B} \right\}. \quad (11)$$

We suggest to solve (11) by the following iterative algorithm. Suppose that the current estimate in the iteration is  $\hat{\theta}_{[\ell-1]}(z, \mathcal{B})$ . Then we update to  $\hat{\theta}_{[\ell]}(z, \mathcal{B})$  by solving  $(\alpha_0, \alpha_1)$  in the equation

$$0 = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) \times \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[\ell-1]}(Z_{i1}, \mathcal{B}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \hat{\theta}_{[\ell-1]}(Z_{im}, \mathcal{B}), \mathcal{B} \right\}.$$

Set  $\hat{\theta}_{[\ell]}(z, \mathcal{B}) = \alpha_0$ . At convergence, for any fixed  $\mathcal{B}$ , we have the kernel estimator  $\hat{\theta}(z, \mathcal{B})$ .

We now define two methods for estimating  $\mathcal{B}_0$ . The *profile-kernel estimator*  $\hat{\mathcal{B}}_p$  maximizes  $\sum_{i=1}^n \mathcal{L} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B} \right\}$ . Maximization of the profile likelihood requires calculating the derivative  $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B}) = \partial \hat{\theta}(z, \mathcal{B}) / \partial \mathcal{B}$ . This can be computed by numerical differentiation: in addition, in the Appendix (Section A.8), we show how to use an algorithm very similar to (5) to compute  $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$  by solving a kernel estimating equation.

In some cases, the profile-kernel method may be difficult to implement numerically due to the additional required computation of  $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$ . Instead, a *backfitting* algorithm can be used. In the iterative backfitting algorithm, suppose that the current estimate is  $\mathcal{B}_*$ . The updated backfitting estimate then maximizes in  $\mathcal{B}$  the function  $\sum_{i=1}^n \mathcal{L} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}_*), \dots, \hat{\theta}(Z_{im}, \mathcal{B}_*), \mathcal{B} \right\}$ . The fully-iterated solution to this algorithm is denoted by  $\hat{\mathcal{B}}_b$ . It is somewhat more general to write the updated backfitting estimate as the solution in  $\mathcal{B}$  to

$$0 = \sum_{i=1}^n \Psi_i(\mathcal{B}_*, \mathcal{B}) = \sum_{i=1}^n \mathcal{L}_{\mathcal{B}} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}_*), \dots, \hat{\theta}(Z_{im}, \mathcal{B}_*), \mathcal{B} \right\}, \quad (12)$$

where  $\mathcal{L}_{\mathcal{B}} \left\{ \tilde{Y}_i, \tilde{X}_i, \theta(Z_1), \dots, \theta(Z_m), \mathcal{B} \right\} = \partial \mathcal{L}(\tilde{Y}_i, \tilde{X}_i, \theta(Z_1), \dots, \theta(Z_m), \mathcal{B}) / \partial \mathcal{B}$ .

In general problems of this type, Hu, et al. (2004) have shown that backfitting and profiling lead to different asymptotic distributions. However, Hu, et al. (2004) also show that in Example 1 and equation (2), the use of the smoother defined in (5) leads to profiling and backfitting being asymptotically equivalent. Thus one would conjecture that the same equivalence holds in our general problem, a conjecture verified in Section 3.3. It should be noted that as shown in Section 3.3, to

obtain a  $\sqrt{n}$ -consistent estimator of  $\mathcal{B}$ , undersmoothing of the nonparametric function  $\theta(z)$  is required by the backfitting method: no such undersmoothing is needed when the profile-kernel method is used.

### 3.2 Optimal Semiparametric Score

To study the asymptotic properties of the profile-kernel and backfitting estimators of  $\mathcal{B}$ , we first derive the semiparametric efficiency bound and efficient semiparametric score function in the case that  $\mathcal{L}(\bullet)$  is a likelihood function.

**Result #2: Semiparametric Efficiency Bound** Assume that  $(\tilde{Y}_i, \tilde{X}_i, \tilde{Z}_i)$  are independent and identically distributed, and that  $\mathcal{L}(\bullet)$  is a likelihood function conditional on  $(\tilde{X}, \tilde{Z})$ . Then the optimal semiparametric score function is

$$\mathcal{L}_{\mathcal{B}}(\bullet) + \sum_{j=1}^m \mathcal{L}_{j\theta}(\bullet) \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0), \quad (13)$$

where the argument is  $\{\tilde{Y}, \tilde{X}, \theta_0(Z_1), \dots, \theta_0(Z_m), \mathcal{B}_0\}$ , and  $\theta_{\mathcal{B}}(Z_j, \mathcal{B}_0)$  is the asymptotic limit of  $\hat{\theta}_{\mathcal{B}}(Z_j, \mathcal{B}_0)$  and  $\mathcal{B}_0$  is the true value of  $\mathcal{B}$ . In addition, the asymptotic covariance matrix of the optimal semiparametric estimator is  $n^{-1}\mathcal{V}^{-1}$ , where

$$\mathcal{V} = \text{cov}\{\mathcal{L}_{\mathcal{B}}(\theta_0, \mathcal{B}_0) + \sum_{j=1}^m \mathcal{L}_{j\theta}(\theta_0, \mathcal{B}_0) \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0)\}. \quad (14)$$

The proof of (13) is given in Appendix A.3.

### 3.3 Asymptotic Distribution Theory

We study in this section the asymptotic properties of the profile-kernel estimator  $\hat{\mathcal{B}}_p$  and the backfitting estimator  $\hat{\mathcal{B}}_b$  under a general criterion function  $\mathcal{L}(\bullet)$ . To study the asymptotic properties of the profile-kernel estimator  $\hat{\mathcal{B}}_p$ , we first provide the asymptotic properties of the kernel estimator of the derivative  $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$ . Define  $\mathcal{L}_{j\theta_{\mathcal{B}}}(\bullet) = \partial \mathcal{L}_{j\theta}(\tilde{Y}, \tilde{X}, \eta_1, \dots, \eta_m, \mathcal{B}) / \partial \mathcal{B}$ , and

$$\begin{aligned} \epsilon_{ij}^{\#}(\theta, \mathcal{B}) &= \mathcal{L}_{j\theta_{\mathcal{B}}}\{\tilde{Y}_i, \tilde{X}_i, \theta(Z_{i1}), \dots, \theta(Z_{im}), \mathcal{B}\} \\ &\quad + \sum_{k=1}^m \mathcal{L}_{jk\theta}\{\tilde{Y}_i, \tilde{X}_i, \theta(Z_{i1}), \dots, \theta(Z_{im}), \mathcal{B}\} \theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}). \end{aligned}$$

As we show in the Appendix A.5,  $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B}_0) = \theta_{\mathcal{B}}(z, \mathcal{B}_0) + o_p(1)$ , where  $\theta_{\mathcal{B}}(z, \mathcal{B}_0)$  satisfies

$$0 = \sum_{j=1}^m f_j(z) E\{\epsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0) | Z_j = z\}. \quad (15)$$

Define

$$\mathcal{F} = E\{\mathcal{L}_{\mathcal{B}\mathcal{B}} + \sum_{j=1}^m \mathcal{L}_{j\theta\mathcal{B}}(\bullet)\theta_{\mathcal{B}}^T(Z_j, \mathcal{B}_0)\},$$

where  $\mathcal{L}_{\mathcal{B}\mathcal{B}}(\bullet) = \partial^2 \mathcal{L}(\bullet) / \partial \mathcal{B}^2$ .

**Result #3: Profile-Kernel Method** Assume that  $(\tilde{Y}_i, \tilde{X}_i, \tilde{Z}_i)$  are independent and identically distributed, and that  $0 = E\{\mathcal{L}_{\mathcal{B}}(\bullet) | \tilde{Z}\} = E\{\mathcal{L}_{j\theta}(\bullet) | \tilde{Z}\}$ . Suppose further that the bandwidth  $h \propto n^{-c}$  with  $1/5 \leq c \leq 1/3$ . Then

$$\begin{aligned} n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) &= -\mathcal{F}^{-1} n^{-1/2} \sum_{i=1}^n \{\mathcal{L}_{i\mathcal{B}} + \sum_{j=1}^m \epsilon_{ij} \theta_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0)\} + o_p(1) \\ &\rightarrow \text{Normal}(0, \mathcal{F}^{-1} \mathcal{V} \mathcal{F}^{-1}), \end{aligned} \quad (16)$$

where  $\epsilon_{ij} = \mathcal{L}_{ij\theta}(\bullet)$  and  $\mathcal{V}$  is defined in equation (14). In the case that  $\mathcal{L}(\bullet)$  is a loglikelihood conditioned on  $(\tilde{X}, \tilde{Z})$ ,  $\mathcal{F} = -\mathcal{V}$ , the resulting asymptotic variance is  $\mathcal{V}^{-1}$ , and the profile estimator is semiparametric efficient. The proof of (16) is given in Appendix A.5.

**Result #4: Backfitting Method** Make the same assumptions as in Result #3, except that  $nh^4 \rightarrow 0$  is required, i.e., undersmoothing is required. Then the backfitting estimator  $\hat{\mathcal{B}}_b$  has the same asymptotic distribution as does the profile estimator  $\hat{\mathcal{B}}_p$ . The proof is given in Appendix A.6.

**Result #5: Covariance Matrix Estimation** Consistent estimates of  $\mathcal{F}$  and  $\mathcal{V}$  can be constructed as follows. Let  $\hat{\mathcal{L}}_{i\mathcal{B}}$ ,  $\hat{\mathcal{L}}_{ij\theta}$ ,  $\hat{\mathcal{L}}_{i\mathcal{B}\mathcal{B}}$  and  $\hat{\mathcal{L}}_{ij\theta\mathcal{B}}$  be the estimated versions of the indicated quantities. Let  $\hat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B})$  be the solution of the kernel estimating equation (A.15). Then a consistent estimator of  $\mathcal{V}$  is the sample covariance matrix of the terms  $\hat{\mathcal{L}}_{i\mathcal{B}} + \sum_{j=1}^m \hat{\mathcal{L}}_{ij\theta} \hat{\theta}_{\mathcal{B}}(Z_{ij}, \hat{\mathcal{B}})$ . Further, a consistent estimator of  $\mathcal{F}$  is

$$\hat{\mathcal{F}} = n^{-1} \sum_{i=1}^n [\hat{\mathcal{L}}_{i\mathcal{B}\mathcal{B}} + \hat{\mathcal{L}}_{ij\theta\mathcal{B}} \hat{\theta}_{\mathcal{B}}^T(Z_{ij}, \hat{\mathcal{B}})].$$

## 4 Pseudolikelihood With Nuisance Parameters

In many problems, it is convenient to estimate a subset of parameters by alternative algorithms. For example, in the partially linear model problem of Wang, et al. (2004), the mean functions are  $X_{ij}^T \beta_0 + \theta_0(Z_{ij})$  and the covariance matrix is  $\Sigma_{\epsilon_0}$ . In our notation,  $\mathcal{B}_0 = \{\beta_0^T, \text{vec}^T(\Sigma_{\epsilon_0})\}^T$ . Wang, et al. (2004) provided an initial estimate  $\hat{\Sigma}_{\epsilon p}$  of  $\Sigma_{\epsilon_0}$ , and then applied our algorithm only to  $\beta$  while pretending that  $\Sigma_{\epsilon_0}$  is known and equal to  $\hat{\Sigma}_{\epsilon p}$ .

Problems such as this are easily handled in our context as follows. Suppose that  $\mathcal{B}^T = (\kappa^T, \gamma^T)$  and that we have a preliminary estimate  $\hat{\gamma}_{\text{prelim}}$  with the property that it has the asymptotic expansion  $n^{1/2}(\hat{\gamma}_{\text{prelim}} - \gamma_0) = n^{-1/2} \sum_{i=1}^n \mathcal{U}_i + o_p(1)$ , where  $E(\mathcal{U}) = 0$ . Let  $e_1 = (I, 0)$  so that  $\kappa = e_1 \mathcal{B}$  and

write  $(\mathcal{F}_{11}, \mathcal{F}_{12}) = e_1 \mathcal{F}$ . Then in the Appendix at equation (A.10), we show that for either profiling or backfitting,

$$\begin{aligned} n^{1/2}(\hat{\kappa} - \kappa_0) &= -\mathcal{F}_{11}^{-1} [n^{-1/2} \sum_{i=1}^n \{\mathcal{L}_{i\kappa} + \sum_{j=1}^m \mathcal{L}_{ij\theta} \theta_\kappa(Z_{ij}, \mathcal{B}_0)\} + \mathcal{F}_{12} n^{1/2} (\hat{\gamma}_{\text{prelim}} - \gamma_0)] + o_p(1) \\ &= -\mathcal{F}_{11}^{-1} n^{-1/2} \sum_{i=1}^n \{\mathcal{L}_{i\kappa} + \sum_{j=1}^m \mathcal{L}_{ij\theta} \theta_\kappa(Z_{ij}, \mathcal{B}_0) + \mathcal{F}_{12} \mathcal{U}_i\} + o_p(1), \end{aligned}$$

from which the covariance of the asymptotic distribution of  $n^{1/2}(\hat{\kappa} - \kappa_0)$  follows. In some cases, such as that investigated by Wang, et al. (2004),  $\mathcal{F}_{12} = 0$ , in which case the asymptotic covariance matrix becomes  $\mathcal{F}_{11}^{-1} \mathcal{V}_{11} \mathcal{F}_{11}^{-1}$ . In either case, a consistent estimator of the asymptotic covariance matrix is easily constructed.

## 5 Examples

### 5.1 Data With Common $Z$ Values

In some situations, the  $Z_{ij}$  have sets of common values in a way that the first  $m_1$  observations have common value  $Z_{i1}^*$ , the next  $m_2$  have common value  $Z_{i2}^*$ , etc. For example, consider problems in which there are  $n$  families, family  $i$  ( $i = 1, \dots, n$ ) has  $L_i$  children, the  $j$ th child ( $j = 1, \dots, L_i$ ) has a baseline measure  $Z_{ij}^*$  and repeated measures  $Y_{ijk}$  over time for  $k = 1, \dots, m_{ij}$  and a possible repeated time-varying covariate  $X_{ijk}$ . Consider a three-level hierarchical model

$$Y_{ijk} = X_{ijk}^T \beta_0 + \theta_0(Z_{ij}^*) + \epsilon_{ijk}, \quad (17)$$

where  $i = 1, \dots, n$  (e.g.,  $i^{\text{th}}$  family),  $j = 1, \dots, L_i$  (e.g.,  $j^{\text{th}}$  member in the  $i^{\text{th}}$  family),  $k = 1, \dots, m_{ij}$  (e.g.,  $k^{\text{th}}$  time point). Equation (17) models the effect of the baseline subject-level covariate  $Z_{ij}^*$  nonparametrically and other covariates  $X_{ijk}$  parametrically. Denote the covariance matrix of  $\epsilon_i$  by  $\Sigma_i$ , which is a  $\sum_{j=1}^{L_i} m_{ij} \times \sum_{j=1}^{L_i} m_{ij}$  matrix. Assuming  $\Sigma_i$  is known, the criterion function is

$$[\tilde{Y}_i - \tilde{X}_i \beta - \{\theta(Z_{i1}^*) e_{i1}^T, \dots, \theta(Z_{iL_i}^*) e_{iL_i}^T\}^T]^T \Sigma_i^{-1} [\tilde{Y}_i - \tilde{X}_i \beta - \{\theta(Z_{i1}^*) e_{i1}^T, \dots, \theta(Z_{iL_i}^*) e_{iL_i}^T\}^T]. \quad (18)$$

where  $e_{ij}$  be a  $m_{ij} \times 1$  vector of ones. Let  $\epsilon_{ij} = (\epsilon_{ij1}, \dots, \epsilon_{ijm_{ij}})^T$ ,  $\epsilon_i = (\epsilon_{i1}^T, \dots, \epsilon_{iL_i}^T)^T$  and  $\tilde{\epsilon} = (\epsilon_1^T, \dots, \epsilon_n^T)^T$ . Now partition  $\Sigma_i$  as follows: the  $(jk)^{\text{th}}$  block  $\Sigma_{i,jk} = \text{covariance}(\epsilon_{ij}, \epsilon_{ik})$  and the dimension of  $\Sigma_{i,jk}$  is  $m_{ij} \times m_{ik}$ . Denote  $\Sigma_i^{-1} = \{\Sigma_i^{jk}\}$ , where the partition of  $\Sigma_i^{-1}$  is the same as  $\Sigma_i$ . Chen and Jin (2001) considered a problem similar to our setting without the parametric component, and proposed to apply Wang's (2003) smoothing algorithm pretending that the repeated baseline values of  $Z_{ij}^*$  from the same subject were distinct overtime. Estimation based on our criterion function

(18) effectively accounts for the nature that the data have common  $Z$  values, and would yield a more efficient estimator.

Specifically, for any given  $\beta$ , define  $\mathcal{Y}_{ijk} = \mathcal{Y}_{ijk}(\beta) = Y_{ijk} - X_{ijk}^T \beta$ , and define  $\mathcal{Y}_{ij}$ ,  $\mathcal{Y}_i$  and  $\tilde{\mathcal{Y}}$  in the same fashion as  $\epsilon_{ij}$ ,  $\epsilon_i$  and  $\tilde{\epsilon}$ . Define  $Z_i^* = (Z_{i1}^*, \dots, Z_{iL_i}^*)^T$  and  $\tilde{Z}^* = (Z_{11}^*, \dots, Z_{m,L_n}^*)^T$  and define  $\tilde{X} = (X_{111}, \dots)^T$ . Then, the linear kernel estimating equation at the  $\ell^{\text{th}}$  iteration is

$$\sum_{i=1}^n \sum_{j=1}^{L_i} K_h(Z_{ij}^* - z) G_{ij}(z) (0, \dots, 0, e_{ij}^T, 0, \dots, 0) \Sigma_i^{-1} \{\mathcal{Y}_i - \mu_i(Z_i^*, z_0)\} = 0, \quad (19)$$

where  $G_{ij}(z)$  defined in Section 2, and

$$\mu_i(Z_i^*, z_0) = \{\hat{\theta}_{[\ell-1]}(Z_{i1}^*) e_{i1}^T, \dots, \{\hat{\alpha}_0 + \hat{\alpha}_1(Z_{ij}^* - z)\} e_{ij}^T, \dots, \hat{\theta}_{[\ell-1]}(Z_{iL_i}^*) e_{iL_i}^T\}^T.$$

In Section A.9, we give an explicit, closed form solution to (19): no iteration is necessary, and (19) is only a descriptive device. Indeed, we derive an explicit form of a smoother matrix  $\mathcal{S}$  such that  $\hat{\theta}(\tilde{Z}^*, \beta) = \mathcal{S}\tilde{\mathcal{Y}}(\beta) = \mathcal{S}\tilde{Y} - \mathcal{S}\tilde{X}\beta$ , where  $\mathcal{S}$  is given in equation (A.21). This means that the profile-kernel estimator of  $\beta$  is also explicit, i.e., non-iterative, since it is the generalized least squares estimator in the model with responses  $(I - \mathcal{S}_*)\tilde{Y}$  and predictors  $(I - \mathcal{S}_*)\tilde{X}$ , where  $\mathcal{S}_*$  is the expanded version of  $\mathcal{S}$  appropriate for the smoothing of all the responses by accounting for the common  $Z_{ij}$  within the same subject, i.e.,  $\mathcal{S}_* = E\mathcal{S}$ , where  $E = \text{diag}(e_{11}, \dots, e_{nL_n})$  is an  $N \times \sum_{i=1}^n L_i$  matrix and  $N = \sum_{i=1}^n \sum_{j=1}^{L_i} m_{ij}$  is the total sample size. The profile-kernel estimator is

$$\hat{\beta} = \{\tilde{X}^T (I - \mathcal{S}_*)^T \tilde{\Sigma}^{-1} (I - \mathcal{S}_*) \tilde{X}\}^{-1} \tilde{X}^T (I - \mathcal{S}_*)^T \tilde{\Sigma}^{-1} (I - \mathcal{S}_*) \tilde{Y}, \quad (20)$$

where  $\tilde{\Sigma} = \text{diag}(\Sigma_1, \dots, \Sigma_n)$ .

### 5.1.1 Simulation Study

We applied our method to the case of  $n = 100$  clusters with 6 observations per cluster, with  $Z_{i1} = Z_{i2} = Z_{i3}$ ,  $Z_{i4} = Z_{i5} = Z_{i6}$ , i.e., we fit the hierarchical model (17) with  $n = 100$  families,  $L = 2$  subjects per family and  $m = 3$  repeated measures over time per subject. We assume the correlation structure as autoregressive with correlation 0.60 among repeated measures over time and common between-subject (within-family) correlation 0.20: let  $\Sigma$  denote the resulting covariance matrix. The true function was  $\theta_0(z) = \sin(8z - 2)$ . The  $Z$ -values were generated as independent uniforms, while the  $X$ -values were bivariate independent uniforms minus the corresponding value of  $Z$ . The true value was  $\beta_0 = (1, 1)^T$ .

The Epanechnikov kernel was used. Working independence was based on bandwidths selected using the method of Ruppert, et al. (1995). The covariance matrix  $\hat{\Sigma}$  of the  $\epsilon_{ij}$  was estimated as the

sample covariance matrix of the residuals formed by a preliminary working independence regression spline fit. We used pseudolikelihood, with the estimated covariance matrix fixed as above. Both the method that ignored the fact that there were common values of  $Z$  and our method were applied with bandwidth selected via the following simple device. For a given  $\beta$  we formed  $Y_{ij} - X_{ij}^T \beta$  and then calculated  $\hat{\theta}(\cdot)$  using the closed form expression (A.21). With  $\mathcal{S}$  as the smoother matrix,  $\text{cov}\{\hat{\theta}(\bullet)\}$  is estimated as  $\mathcal{S} \text{diag}(\hat{\Sigma}) \mathcal{S}^T$ , and the estimated average variance of the fit follows directly. Bias was estimated as in Wang (2003). We then minimized the estimated mean squared error as a function of the bandwidth. The estimator of the profile-kernel estimator of  $\beta$  was calculated using the closed form formula (20).

In 1000 simulated data sets, both weighted methods achieved over 70% greater mean squared error efficiency for estimating  $\beta_0$  than the working independence estimator. For estimating  $\theta_0(z)$ , the method that ignored the common  $Z$ -values was 35% more efficient in mean squared error than working independence, but our method was 65% more efficient.

### 5.1.2 Analysis of The Kenya Hemoglobin Data

We applied our method to analyze a subset of the Kenya hemoglobin data to study the changes of hemoglobin over time in the first year since birth and the risk factors of hemoglobin among Kenya children. This subset contained  $n = 68$  families with  $L = 2$  children per family and  $m = 4$  repeated measures per child over time in the first year since birth. Hemoglobin was measured at each visit and visit times varied from child to child. The risk factors of interest include mother's age at child birth, child sex, and placental parasitemia density (PDEN), a marker for malaria, which could affect hemoglobin. Log transformation was applied to PDEN to make the normality assumption plausible. Preliminary analysis showed that the effect of mother's age was nonlinear. We considered the semiparametric model (17) and modelled the mother's age effect nonparametrically, and sex, PDEN and time effects parametrically. Specifically, we set  $Z_{ij}$  = mother's age at birth,  $X_{ijk} = \{\text{sex}, \text{logdpden}, \text{month}, (\text{month} - 4)_+\}$ , where  $\text{sex}=1$  if female and 0 if male,  $\text{logdpden}=\log(\text{PDEN}+1)$ , the function  $f_+ = f$  if  $f > 0$  and 0 if  $f \leq 0$ . Note that the terms  $\{\text{months}, (\text{month} - 4)_+\}$  model the time effect as a piece-wise linear function with a knot at 4 months. This trend is observed by preliminary analysis of the data.

In our analysis, we used pseudolikelihood, with the following modifications from the simulation. We started with an estimate of  $\Sigma$  as obtained from a preliminary regression spline fit, then estimated the bandwidth using leaving-one-mother-out cross validation, and thus obtained estimates of  $\theta_0(\bullet)$

and  $\beta_0$ . From this, we formed residuals  $Y_{ij} - X_{ij}^T \hat{\beta} - \hat{\theta}(Z_{ij}, \hat{\beta})$ , re-estimated the covariance matrix, re-estimated the bandwidth, etc., repeating this process 10 times.

For numerical stability, we standardized hemoglobin. We obtained an estimated residual variance of 0.66, an estimated autocorrelation of 0.20 and an estimated between-child (within-mother) correlation of 0.13. The estimated CV bandwidth was 0.23. The correlation was low/moderate in this example. In Figure 1, we compared the estimated nonparametric curve estimates of the effects of mother’s age at birth using the working independence kernel estimator and our proposed likelihood-based kernel estimator (with/without accounting for ties in mother age). The estimated curves were similar. Children’s hemoglobin increased with mother’s age at birth for mothers younger than 22 years old then decreased slightly with mother’s age until early 30 then started decreasing quickly with mother’s age, indicating children are likely to have much lower hemoglobin if mothers give birth after early 30 years of age, i.e., giving birth after early 30 is likely to considerably increase children’s risk of anemia (low hemoglobin).

As expected, since the correlation was not high, the estimates of the regression coefficients  $\beta$  were roughly the same for the working independence kernel fit with bandwidths selected using the method of Ruppert, et al. (1995), the method of Wang, et al. (2004) ignoring the common  $Z$ -values, and our method. Estimated standard errors were computed ignoring the correlation for the working independence methods, and using the sandwich method for our likelihood-based methods. These standard errors were roughly the same in all cases. The results are given in Table 1. Hemoglobin drops quickly after birth and decreases at a slower rate after month 4. Both sex and placental parasitemia density do not affect hemoglobin significantly.

## 5.2 Measurement Error Models

Here we consider the multivariate partially linear measurement error model, where

$$Y_{ij} = C_{ij}^T \beta_0 + \theta_0(Z_{ij}) + \epsilon_{ij}, \tag{21}$$

where  $\tilde{\epsilon}_i$  has covariance matrix  $\Sigma_{\epsilon_0}$ . Instead of observing  $C_{ij}$  we observe  $W_{ij} = C_{ij} + U_{ij}$ . Define  $\tilde{U}_i = (U_{i1}, \dots, U_{im})^T$ . These measurement errors have mean zero and the property that  $\text{cov}\{\text{vec}(\tilde{U}_i)\} = \Sigma_{u_0}$ , assumed here to be known. There is to date no literature on this problem other than Lin and Carroll (2000), which came to unsatisfactory conclusions such as that in panel data it was better to ignore the correlation structure in the responses.

Define  $G(\Sigma_{\epsilon}, \Sigma_{u_0}) = E(\tilde{U}^T \Sigma_{\epsilon}^{-1} \tilde{U})$  and define  $\mathcal{K}(\Sigma_{u_0}, \beta) = E(\tilde{U} \beta \beta^T \tilde{U}^T)$ . Note that  $\beta^T G(\Sigma_{\epsilon}, \Sigma_{u_0}) \beta =$

$\text{trace}\{\Sigma_\epsilon^{-1}E(\tilde{U}\beta\beta^T\tilde{U}^T)\} = \text{trace}\{\Sigma_\epsilon^{-1}\mathcal{K}(\Sigma_{u0},\beta)\}$ . In (21),  $\mathcal{B} = (\beta, \tau, \Sigma_\epsilon)$  and the criterion function is

$$(1/2)\log\{\det(\Sigma_\epsilon^{-1})\} + (1/2)\beta^T G(\Sigma_\epsilon, \Sigma_{u0})\beta - (1/2)\{\tilde{Y} - \tilde{W}\beta - \theta(\tilde{Z})\}^T \Sigma_\epsilon^{-1} \{\tilde{Y} - \tilde{W}\beta - \theta(\tilde{Z})\}. \quad (22)$$

Equation (22) is new even in the parametric measurement error literature.

For symmetric matrices  $\Sigma$ ,  $\partial \log(|\Sigma|)/\partial \Sigma = 2\Sigma^{-1} - \text{diag}(\Sigma^{-1})$  and  $\partial \text{trace}(\Sigma A)/\partial \Sigma = 2A - \text{diag}(A)$ . It is readily see that the derivative of (22) with respect to  $\beta$ ,  $\Sigma_\epsilon$  and  $\theta$  evaluated at the true parameters has expectation zero, and thus (22) satisfies the essential condition (3).

In this problem, the backfitting algorithm is computationally convenient. Of course, for given  $\mathcal{B} = (\beta, \Sigma_\epsilon)$ , forming the estimate  $\hat{\theta}(z, \mathcal{B})$  is easy since it is simply the estimate of Wang (2002) applied to the terms  $Y_{ij} - W_{ij}^T \beta$ . Indeed, define  $\mathcal{Y} = (Y_{11}, \dots, Y_{nm})^T$ ,  $\mathcal{Z} = (Z_{11}, \dots, Z_{nm})^T$  and  $\mathcal{W} = (W_{11}, \dots, W_{nm})^T$ . Then as Lin, et al. (2004) show, there is a smoother matrix  $\mathcal{S} = \mathcal{S}(\Sigma_\epsilon)$  such that  $\hat{\theta}(\mathcal{Z}, \mathcal{B}) = \mathcal{S}(\mathcal{Y} - \mathcal{W}\beta)$ . If  $\hat{\beta}_c$ ,  $\hat{\mathcal{B}}_c$  and  $\hat{\Sigma}_{\epsilon,c}$  are the current estimates, the updated estimates are

$$\begin{aligned} \hat{\beta}_{\text{new}} &= \{n^{-1} \sum_{i=1}^n \tilde{W}_i^T \hat{\Sigma}_{\epsilon,c}^{-1} \tilde{W}_i - G(\hat{\Sigma}_{\epsilon,c}, \Sigma_{u0})\}^{-1} n^{-1} \sum_{i=1}^n \tilde{W}_i^T \hat{\Sigma}_{\epsilon,c}^{-1} \{\tilde{Y}_i - \hat{\theta}(\tilde{Z}_i, \hat{\mathcal{B}}_c)\}; \\ \hat{\Sigma}_{\epsilon,\text{new}} &= n^{-1} \sum_{i=1}^n \{\tilde{Y}_i - \tilde{W}_i \hat{\beta}_c - \hat{\theta}(\tilde{Z}_i, \hat{\mathcal{B}}_c)\} \{\tilde{Y}_i - \tilde{W}_i \hat{\beta}_c - \hat{\theta}(\tilde{Z}_i, \hat{\mathcal{B}}_c)\}^T - \mathcal{K}(\Sigma_{u0}, \hat{\beta}_c). \end{aligned} \quad (23)$$

Profile-Pseudolikelihood estimates are also easily constructed. Let  $\tilde{\Sigma}_\epsilon = I_n \otimes \Sigma_\epsilon$ . Let  $\mathcal{W}_* = (I - \mathcal{S})\mathcal{W}$  and  $\mathcal{Y}_* = (I - \mathcal{S})\mathcal{Y}$ . Then for given  $\Sigma_\epsilon$ , the profile estimate of  $\beta$  is given by

$$\{\mathcal{W}_*^T \tilde{\Sigma}_\epsilon^{-1} \mathcal{W}_* - nG(\Sigma_\epsilon, \Sigma_{u0})\}^{-1} \mathcal{W}_*^T \tilde{\Sigma}_\epsilon^{-1} \mathcal{Y}_*.$$

A simple estimate of  $\Sigma_\epsilon$  is to form the working independence estimate of  $\beta$  and apply (23).

## 6 Discussion

This paper has described nonparametric and semiparametric methods in cases where the nonparametric function is evaluated repeatedly within a sampling unit. Examples discussed included old and new versions of marginal longitudinal and clustered data, matched case-control studies, generalized linear mixed models, common additive models linked by a parameter and multivariate measurement error models. The methodology is motivated by the use of a criterion function that would be used if the problem were a parametric one: if the criterion function is a likelihood, then our methods are semiparametric efficient. We showed that backfitting and profiling gave asymptotically the same results, although undersmoothing is needed for backfitting, and also showed how to use pseudolikelihood methods within our context when some of the parameters are more conveniently estimated by alternative algorithms.

Although we have motivated the methodology by basing it on criterion functions, the approach is considerably more general. Our approach really only requires the following. First, we need a set of unbiased estimating functions  $\mathcal{L}_{j\theta}\{\tilde{Y}, \tilde{X}, \theta_0(Z_1), \dots, \theta_0(Z_m), \mathcal{B}_0\}$  that satisfy (3). Second, we need an estimating function  $\Psi_{\mathcal{B}}\{\tilde{Y}, \tilde{X}, \theta_0(Z_1), \dots, \theta_0(Z_m), \mathcal{B}_0, \mathcal{B}_0\}$  taking the place of (12) and also satisfying (3): the double argument in  $\mathcal{B}_0$  is meant to allow for the possibility of using backfitting. It is useful to use the symbols  $\mathcal{L}$  and  $\Psi$  to emphasize that the derivative of the former with respect to  $\mathcal{B}$  need not be the same as the derivative of the latter with respect to the  $j^{\text{th}}$  component of  $\theta$ . It can be shown that Result #1 and (8) still hold with the same notation, as does the fundamental identity (15). The basic backfitting expansion (A.9) in the appendix, as well as the definition of  $\mathcal{F}$  in Result #3 also holds with  $\mathcal{L}$  replaced by  $\Psi$ . It then becomes straightforward to derive the asymptotic distribution of the estimate of  $\mathcal{B}_0$ : note here however that  $\mathcal{F}_1 + \mathcal{F}_2$  need no longer be symmetric. The asymptotic covariance matrix of the resulting estimator  $\hat{\mathcal{B}}$  is more complicated than that given in (16), because it involves the implicitly defined function  $\mathcal{G}$  in (6). However, the bootstrap method that bootstraps clusters can be used to estimate the covariance of  $\hat{\mathcal{B}}$  (Chen, et al., 2004).

## Acknowledgments

Lin's research was supported by a grant from the National Cancer Institute (CA-76404). Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We thank Naisyin Wang for many helpful suggestions.

## Appendix: Sketch of Technical Arguments

### A.1 A Key Technical Lemma

**Lemma A.1:** Let  $\hat{\theta}_{[\ell]}(\bullet)$  be the estimate at the  $\ell^{\text{th}}$  stage of the iteration. Then

$$\begin{aligned} \hat{\theta}_{[\ell]}(z) - \theta_0(z) &= (h^2/2)b_{[0]}(z) - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \epsilon_{ij} / \Omega(z) \\ &\quad - n^{-1} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j}^m \frac{K_h(Z_{ij} - z)}{\Omega(z)} \mathcal{L}_{jk\theta i}(\bullet) \left\{ \hat{\theta}_{[\ell-1]}(Z_{ik}) - \theta_0(Z_{ik}) \right\} + o_p(n^{-1/2}). \end{aligned} \tag{A.1}$$

Here is a brief sketch of (A.1). By Taylor expansion, we have

$$0 = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) \mathcal{L}_{ij\theta}(\bullet)$$

$$+n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) G_{ij}^T(z, h) \mathcal{L}_{ijj\theta}(\bullet) \begin{bmatrix} \hat{\alpha}_0 - \alpha_0 \\ \hat{\alpha}_1 - \alpha_1 \end{bmatrix} + o_p(n^{-1/2}),$$

where the argument is  $\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[\ell-1]}(Z_{i1}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \hat{\theta}_{[\ell-1]}(Z_{im})\}$ . It is easily seen that the sum in the second argument converges at the appropriate rate to  $\Omega(z)I_2$ , where  $I_2$  is the  $2 \times 2$  identity matrix (again, this is because  $K$  has variance 1.0). Hence,

$$\begin{aligned} -\Omega(z)(\hat{\alpha}_0 - \alpha_0) &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \mathcal{L}_{ijj\theta}(\bullet) + o_p(n^{-1/2}) = A_{1n} + A_{2n} + o_p(n^{-1/2}); \\ A_{1n} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \\ &\quad \times \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \theta_0(Z_{i1}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \theta_0(Z_{im}) \right\}; \\ A_{2n} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \\ &\quad \times \left[ \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[\ell-1]}(Z_{i1}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \hat{\theta}_{[\ell-1]}(Z_{im}) \right\} \right. \\ &\quad \left. - \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \theta_0(Z_{i1}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \theta_0(Z_{im}) \right\} \right]. \end{aligned}$$

By a direct calculation,  $A_{1n} = A_{1n1} - A_{1n2}$ , where  $A_{1n1} = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \epsilon_{ij}$  and

$$\begin{aligned} A_{1n2} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \left[ \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \theta_0(Z_{i1}), \dots, \theta_0(Z_{im}) \right\} \right. \\ &\quad \left. - \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \theta_0(Z_{i1}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \theta_0(Z_{im}) \right\} \right] \\ &= (h^2/2) \theta^{(2)}(z) n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \{(Z_{ij} - z)/h\}^2 \mathcal{L}_{ijj\theta}(\bullet) + o_p(n^{-1/2}) \\ &= (h^2/2) b_{[0]}(z) \Omega(z) + o_p(n^{-1/2}), \end{aligned}$$

the last since  $\int x^2 K(x) dx = 1$ . Equation (A.1) now follows since

$$A_{2n} = n^{-1} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j}^m K_h(Z_{ij} - z) \mathcal{L}_{jk\theta i}(\bullet) \left\{ \hat{\theta}_{[\ell-1]}(Z_{ik}) - \theta_0(Z_{ik}) \right\} + o_p(n^{-1/2}).$$

## A.2 Sketch of Result #1 and (8): Expansion for Nonparametric Estimator

The basic argument is the same as that in Wang (2003), namely repeated application of (A.1). At the technical level, in her argument she required that there be an initial estimator following the property that there are random variables  $\epsilon_{ij}^*$  and  $\epsilon_{ij}^{**}$  that have mean zero given  $\tilde{Z}_i$ , and functions  $(b^*, R_{j1}, R_{j2})$  for  $j = 1, \dots, m$  such that the initial estimator satisfies

$$\hat{\theta}_{[0]}(z) - \theta_0(z) = (h^2/2) b^*(z) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) R_{j1}(z, \tilde{Z}_i) \epsilon_{ij}^*$$

$$+n^{-1} \sum_{i=1}^n \sum_{j=1}^m R_{j2}(z, \tilde{Z}_i) \epsilon_{ij}^{**} + o_p(n^{-1/2}). \quad (\text{A.2})$$

As in her calculations, one starts with (A.2) and then applies (A.1) to get an expansion for the first step in the iteration. This new expansion is then substituted into (A.1) to get an expansion for the second step in the iteration, etc. Under the assumption that the algorithm converges, the effect of the initial estimator disappears. The calculations are merely extremely detailed rather than difficult, and in the interest of space we do not provide them.

### A.3 Proof of Result #2: Semiparametric Efficient Score

We use Begun, Hall, Huang and Wellner (1983). In their setup, their “ $f$ ” is our  $\exp(\mathcal{L})$ , their “ $\theta$ ” is our  $\mathcal{B}$ , their “ $g$ ” is our  $\theta$ . It is easily derived that their “ $2\rho_\theta/f^{1/2}$ ” is our  $\mathcal{L}_\mathcal{B}$ . Similarly, for an arbitrary function  $\gamma(\bullet)$ , their “ $2A\beta/f^{1/2}$ ” is  $\sum_{j=1}^m \mathcal{L}_{j\theta}(\bullet)\gamma(Z_j)$ . This means that their (3.1) is the following. The semiparametric optimal score is of the form  $\mathcal{L}_\mathcal{B}(\bullet) - \sum_{j=1}^m \mathcal{L}_{j\theta}(\bullet)\gamma_*(Z_j)$ , where  $\gamma_*(\bullet)$  is such that for all  $\gamma(\bullet)$ ,

$$0 = E\left[\left\{\mathcal{L}_\mathcal{B}(\bullet) - \sum_{j=1}^m \mathcal{L}_{j\theta}(\bullet)\gamma_*(\bullet)\right\} \sum_{k=1}^m \mathcal{L}_{k\theta}(\bullet)\gamma(Z_k)\right]. \quad (\text{A.3})$$

We now show that  $\gamma_*(\bullet) = -\theta_\mathcal{B}(\bullet)$  satisfies (A.3). To see this, interchange the indices  $j$  and  $k$  and note that (A.3) means that we must show that for arbitrary  $\gamma(\bullet)$

$$0 = E\left\{\sum_{j=1}^m \mathcal{L}_\mathcal{B}(\bullet)\mathcal{L}_{j\theta}(\bullet)\gamma(Z_j) + \sum_{j=1}^m \sum_{k=1}^m \mathcal{L}_{j\theta}(\bullet)\mathcal{L}_{k\theta}(\bullet)\theta_\mathcal{B}(Z_k)\gamma(Z_j)\right\}.$$

Condition on  $(\tilde{X}, \tilde{Z})$  and note that because  $\mathcal{L}(\bullet)$  is a likelihood function given  $\tilde{X}, \tilde{Z}$ ,

$$\begin{aligned} E\left\{\mathcal{L}_\mathcal{B}(\bullet)\mathcal{L}_{j\theta}(\bullet)|\tilde{X}, \tilde{Z}\right\} &= -E\left\{\mathcal{L}_{j\theta\mathcal{B}}(\bullet)|\tilde{X}, \tilde{Z}\right\}; \\ E\left\{\mathcal{L}_{j\theta}(\bullet)\mathcal{L}_{k\theta}(\bullet)|\tilde{X}, \tilde{Z}\right\} &= -E\left\{\mathcal{L}_{jk\theta}(\bullet)|\tilde{X}, \tilde{Z}\right\}. \end{aligned}$$

Thus we must show that for arbitrary  $\gamma(\bullet)$ ,

$$0 = \sum_{j=1}^m E[\gamma(Z_j)\{\mathcal{L}_{j\theta\mathcal{B}}(\bullet) + \sum_{k=1}^m \mathcal{L}_{jk\theta}(\bullet)\theta_\mathcal{B}(Z_k)\}] = \sum_{j=1}^m E\{\gamma(Z_j)\epsilon_{ij}^\#(\theta_0, \mathcal{B}_0)\}, \quad (\text{A.4})$$

where  $\epsilon_{ij}^\#(\theta_0, \mathcal{B}_0)$  is defined in Section 3.3. This last step follows by conditioning the expectation in (A.4) on  $Z_j$  and then applying (A.7).

#### A.4 Sketch Proof of (15): Fundamental Identity

Because  $n^{-1} \sum_{i=1}^n \{(Z_i - z)/h\} K_h(Z_i - z) = o_p(1)$ , it suffices to consider the slightly altered problem where for any  $\mathcal{B}$ , At convergence to  $\hat{\theta}(z, \mathcal{B})$ , this means that for any  $\mathcal{B}$ ,

$$0 = \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \mathcal{L}_{j\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B} \right\}. \quad (\text{A.5})$$

Differentiating (A.5) with respect to  $\mathcal{B}$ , we get

$$0 = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \left\{ \mathcal{L}_{j\theta \mathcal{B}}(\bullet) + \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\bullet) \hat{\theta}_{\mathcal{B}}(Z_{ik}, \mathcal{B}) \right\},$$

with argument  $\left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B} \right\}$ . Taking limits and evaluating at  $\mathcal{B}_0$  yields (15).

There is an important consequence of (15) that we will use repeatedly. Recall the definition of  $\epsilon_{ij}^{\#}(\theta, \mathcal{B})$  given in Section 3.3. Define

$$H_j(z) = E\{\epsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0) | Z_j = z\}. \quad (\text{A.6})$$

It follows from (15) that  $0 = \sum_{j=1}^m f_j(z) H_j(z)$ , and hence that for any function  $B(\bullet)$ ,

$$0 = E\left\{ \sum_{j=1}^m B(Z_j) H_j(Z_j) \right\}. \quad (\text{A.7})$$

#### A.5 Sketch Proof of Result #3: Asymptotic Distribution for Profiling

Recall that  $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$ , where  $\mathcal{F}_1 = E(\mathcal{L}_{\mathcal{B}\mathcal{B}})$  and  $\mathcal{F}_2 = E\{\sum_{j=1}^m \mathcal{L}_{j\theta \mathcal{B}}(\bullet) \theta_{\mathcal{B}}^{\top}(Z_j, \mathcal{B}_0)\}$ . Also, define

$$\mathcal{F}_3 = E\left\{ \sum_{j=1}^m \sum_{k=1}^m \mathcal{L}_{jk\theta}(\bullet) \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0) \theta_{\mathcal{B}}^{\top}(Z_k, \mathcal{B}_0) \right\}.$$

It is an easy consequence of (A.7) that  $\mathcal{F}_2 + \mathcal{F}_3 = 0$ , so that  $\mathcal{F} = \mathcal{F}_1 + 2\mathcal{F}_2 + \mathcal{F}_3$ .

Let  $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B}) = \partial \hat{\theta}(z, \mathcal{B}) / \partial \mathcal{B}$ , and let its limit as  $n \rightarrow \infty$  be  $\theta_{\mathcal{B}}(z, \mathcal{B})$ . Then the profile estimator solves the equation  $0 = A_1(\hat{\mathcal{B}}, \hat{\theta}) + A_2(\hat{\mathcal{B}}, \hat{\theta})$ , where

$$\begin{aligned} A_1(\hat{\mathcal{B}}, \hat{\theta}) &= n^{-1/2} \sum_{i=1}^n \mathcal{L}_{i\mathcal{B}} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \hat{\mathcal{B}}), \dots, \hat{\theta}(Z_{im}, \hat{\mathcal{B}}), \hat{\mathcal{B}} \right\}; \\ A_2(\hat{\mathcal{B}}, \hat{\theta}) &= n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta} \left\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \hat{\mathcal{B}}), \dots, \hat{\theta}(Z_{im}, \hat{\mathcal{B}}), \hat{\mathcal{B}} \right\} \hat{\theta}_{\mathcal{B}}(Z_{ij}, \hat{\mathcal{B}}_p). \end{aligned}$$

It is important to remember that by assumption,

$$E\{\mathcal{L}_{j\theta}(\bullet)\} = E[\mathcal{L}_{j\theta}\{\tilde{Y}, \tilde{X}, \theta(Z_1), \dots, \theta(Z_m), \mathcal{B}_0\} | \tilde{Z}] = 0. \quad (\text{A.8})$$

A Taylor series expansion shows that

$$\begin{aligned}
A_1(\widehat{\mathcal{B}}, \widehat{\theta}) &= n^{-1/2} \sum_{i=1}^n \mathcal{L}_{i\mathcal{B}}(\bullet) + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta\mathcal{B}}(\bullet) \left\{ \widehat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta(Z_{ij}) \right\} \\
&\quad + (\mathcal{F}_1 + \mathcal{F}_2) n^{1/2} (\widehat{\mathcal{B}} - \mathcal{B}_0) + o_p(1).
\end{aligned} \tag{A.9}$$

where the symbol “ $\bullet$ ” here means evaluated at  $\theta$  and  $\mathcal{B}_0$ . Similarly, we have that

$$\begin{aligned}
A_2(\widehat{\mathcal{B}}, \widehat{\theta}) &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta\mathcal{B}}(\bullet) \theta_{\mathcal{B}}^{\mathbb{T}}(Z_{ij}) n^{1/2} (\widehat{\mathcal{B}} - \mathcal{B}_0) \\
&\quad + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) \theta_{\mathcal{B}\mathcal{B}}(Z_{ij}) n^{1/2} (\widehat{\mathcal{B}} - \mathcal{B}_0) \\
&\quad + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\bullet) \theta_{\mathcal{B}}(Z_{ij}) \theta_{\mathcal{B}}^{\mathbb{T}}(Z_{ik}) n^{1/2} (\widehat{\mathcal{B}} - \mathcal{B}_0) \\
&\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta} \left\{ \widetilde{Y}_1, \widetilde{X}_i, \widehat{\theta}(Z_{i1}, \mathcal{B}_0), \dots, \widehat{\theta}(Z_{im}, \mathcal{B}_0), \mathcal{B}_0 \right\} \widehat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) + o_p(1).
\end{aligned}$$

The first and third terms sum to  $(\mathcal{F}_2 + \mathcal{F}_3) n^{1/2} (\widehat{\mathcal{B}} - \mathcal{B}_0) + o_p(1)$ . Because  $E\{\mathcal{L}_{ij\theta}(\bullet) | \widetilde{Z}_i\} = 0$ , the second term is  $o_p(1)$ . The last term can be decomposed, so that

$$\begin{aligned}
A_2(\widehat{\mathcal{B}}, \widehat{\theta}) &= (\mathcal{F}_2 + \mathcal{F}_3) n^{1/2} (\widehat{\mathcal{B}} - \mathcal{B}_0) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) \theta_{\mathcal{B}}(Z_{ij}) \\
&\quad + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\bullet) \theta_{\mathcal{B}}(Z_{ij}) \left\{ \widehat{\theta}(Z_{ik}, \mathcal{B}_0) - \theta(Z_{ik}) \right\} \\
&\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) \left\{ \widehat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) - \theta_{\mathcal{B}}(Z_{ij}) \right\} + o_p(1).
\end{aligned}$$

Recall that  $\epsilon_{ij} = \mathcal{L}_{ij\theta}(\bullet)$  and that  $H_j(z)$  is defined in (A.6). Hence, if  $P_{ij} = \mathcal{L}_{ij\theta\mathcal{B}}(\bullet) + \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\bullet) \theta_{\mathcal{B}}(Z_{ik})$ , we have shown that

$$\begin{aligned}
-\mathcal{F} n^{1/2} (\widehat{\mathcal{B}}_p - \mathcal{B}_0) &= n^{-1/2} \sum_{i=1}^n \left\{ \mathcal{L}_{i\mathcal{B}} + \sum_{j=1}^m \epsilon_{ij} \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0) \right\} \\
&\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m H_j(Z_{ij}) \left\{ \widehat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta(Z_{ij}) \right\} \\
&\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \left\{ P_{ij} - H_j(Z_{ij}) \right\} \left\{ \widehat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta(Z_{ij}) \right\} \\
&\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) \left\{ \widehat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) - \theta_{\mathcal{B}}(Z_{ij}) \right\} + o_p(1).
\end{aligned} \tag{A.10}$$

We have to show that the second, third and fourth terms of (A.10) are all  $o_p(1)$ .

First consider the second term of (A.10). Substituting in (8), this second term is

$$\begin{aligned} & (h^2/2)n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m b(Z_{ij})H_j(Z_{ij}) - n^{-1/2} \sum_{r=1}^n \sum_{s=1}^m \epsilon_{rs}n^{-1} \sum_{i=1}^n \sum_{j=1}^m H_j(Z_{ij})K_h(Z_{ij} - Z_{rs})/\Omega(Z_{ij}) \\ & + n^{-1/2} \sum_{r=1}^n \sum_{s=1}^m \epsilon_{rs}n^{-1} \sum_{i=1}^n \sum_{j=1}^m H_j(Z_{ij})\mathcal{G}(Z_{ij}, Z_{rs})/\Omega(Z_{ij}). \end{aligned}$$

Each of these terms is easily handled. From (A.7),  $E\{\sum_{j=1}^m b(Z_j)H_j(Z_j)\} = 0$ , so that the first term is  $O_p(h^2) = o_p(1)$ . The inner sums in each of the last two terms are  $o_p(1)$  by (A.7), so these two terms are also  $o_p(1)$ . This completes the argument for the second term of (A.10).

Now turn to the third term of (A.10). Write  $T_{ij} = G_{ij} - H_j(Z_{ij})$ . Substitute in (8) to get that this third term is

$$\begin{aligned} & (h^2/2)n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m b(Z_{ij})T_{ij} - n^{-1/2} \sum_{r=1}^n \sum_{s=1}^m \epsilon_{rs}n^{-1} \sum_{i=1}^n \sum_{j=1}^m T_{ij}K_h(Z_{ij} - Z_{rs})/\Omega(Z_{ij}) \\ & + n^{-1/2} \sum_{r=1}^n \sum_{s=1}^m \epsilon_{rs}n^{-1} \sum_{i=1}^n \sum_{j=1}^m T_{ij}\mathcal{G}(Z_{ij}, Z_{rs})/\Omega(Z_{ij}). \end{aligned}$$

Because  $E(T_{ij}|Z_{ij}) = 0$ , the first term is obviously  $o_p(1)$ , and the inner sums of the last two terms are also  $o_p(1)$ . This completes the argument for the third term of (A.10).

Now turn to the fourth term of (A.10). In Section A.7, we will show a result for  $\hat{\theta}_{\mathcal{B}}(\bullet, \mathcal{B}_0)$  similar to (8), namely that for  $j = 1, 2$ , there are functions  $b_j(\bullet)$ ,  $\Omega_j(\bullet)$  and  $\mathcal{G}_j(\bullet)$  such that

$$\begin{aligned} \hat{\theta}_{\mathcal{B}}(z, \mathcal{B}_0) - \theta_{\mathcal{B}}(z, \mathcal{B}_0) &= (h^2/2)\{b_1(z) + b_2(z)\} - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z)\epsilon_{ij}\Omega_1(z) \\ & - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z)\epsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0)\Omega_2(z) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij}\mathcal{G}_1(z, Z_{ij}) \\ & + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0)\mathcal{G}_2(z, Z_{ij}) + o_p(n^{-1/2}). \end{aligned} \tag{A.11}$$

That the fourth term of (A.10) is  $o_p(1)$  is an easy consequence of (15), (A.7) and (A.11).

## A.6 Sketch Proof of Result #4: Asymptotic Distribution for Backfitting

Using the notation of Section A.5, for backfitting we are solving the equation  $0 = A_1(\hat{\mathcal{B}}_b, \hat{\theta})$ . We have already shown in that section that this means that

$$-\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_b - \mathcal{B}_0) = n^{-1/2} \sum_{i=1}^n \mathcal{L}_{i\mathcal{B}}(\bullet) + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta_{\mathcal{B}}}(\bullet)\{\hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta_0(Z_{ij})\} + o_p(1).$$

However, we also showed in that section that

$$n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \{\mathcal{L}_{ij\theta_{\mathcal{B}}}(\bullet) + \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\bullet)\theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}_0)\}\{\hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta_0(Z_{ij})\} = o_p(1),$$

so that to terms of  $o_p(1)$ ,

$$\begin{aligned} -\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_b - \mathcal{B}_0) &= n^{-1/2} \sum_{i=1}^n \mathcal{L}_{i\mathcal{B}}(\bullet) \\ &\quad - n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\bullet) \theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}_0) \{\hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta_0(Z_{ij})\}. \end{aligned} \quad (\text{A.12})$$

Since the profile estimator satisfies

$$-\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) = n^{-1/2} \sum_{i=1}^n \{\mathcal{L}_{i\mathcal{B}}(\bullet) + \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) \theta_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0)\} + o_p(1), \quad (\text{A.13})$$

we see that we must show that the second terms in (A.12) and (A.13) are asymptotically equivalent.

Make the definitions  $\Omega(z_0) = \sum_{j=1}^m f_j(z_0) E\{\mathcal{L}_{j\theta}(\bullet) | Z_j = z_0\}$ ;

$$\begin{aligned} P_1(z_0) &= \sum_{j=1}^m f_j(z_0) E\{\mathcal{L}_{j\theta\mathcal{B}}(\bullet) | Z_j = z_0\} / \Omega(z_0); \\ P_2(z_0) &= E\left[\sum_{j=1}^m E\{\mathcal{L}_{j\theta\mathcal{B}}(\bullet) | Z_j\} \mathcal{G}(Z_j, z_0) / \Omega(Z_j)\right]; \\ P_3(z_0) &= \sum_{j=1}^m \sum_{k \neq j}^m \{f_j(z_0) / \Omega(z_0)\} E\{\mathcal{L}_{jk\theta}(\bullet) \theta_{\mathcal{B}}(Z_k, \mathcal{B}_0) | Z_j = z_0\}. \end{aligned}$$

Recalling (15), we see that

$$P_1(z_0) = - \sum_{j=1}^m \sum_{k=1}^m \{f_j(z_0) / \Omega(z_0)\} E\{\mathcal{L}_{jk\theta}(\bullet) \theta_{\mathcal{B}}(Z_k, \mathcal{B}_0) | Z_j = z_0\} = -\theta_{\mathcal{B}}(z_0, \mathcal{B}_0) - P_3(z_0).$$

In addition, it is easy to see that

$$P_2(z_0) = \int P_1(z) \mathcal{G}(z, z_0) dz = - \int \theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{G}(z, z_0) dz - \int P_3(z) \mathcal{G}(z, z_0) dz.$$

We now plug in (8) into the second term of (A.12). Noting the assumption that  $nh^4 \rightarrow 0$ , this second term is asymptotically equivalent to  $C_{n1} + C_{n2}$ , where

$$\begin{aligned} C_{n1} &= -n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta\mathcal{B}}(\bullet) n^{-1} \sum_{r=1}^n \sum_{s=1}^m K_h(Z_{rs} - Z_{ij}) \mathcal{L}_{rs\theta}(\bullet) / \Omega(Z_{ij}); \\ C_{n2} &= n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta\mathcal{B}}(\bullet) n^{-1} \sum_{r=1}^n \sum_{s=1}^m \mathcal{L}_{rs\theta}(\bullet) \mathcal{G}(Z_{ij}, Z_{rs}) / \Omega(Z_{ij}). \end{aligned}$$

Interchanging the summations over  $(i, j)$  and  $(r, s)$ , it is easily shown that

$$\begin{aligned} C_{n1} &= -n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) P_1(Z_{ij}) + o_p(1); \\ C_{n2} &= n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) P_2(Z_{ij}) + o_p(1). \end{aligned}$$

Collecting the expressions for  $P_1(z)$  and  $P_2(z)$ , it thus follows that

$$\begin{aligned} R_n &= -\mathcal{F}\{n^{1/2}(\widehat{\mathcal{B}}_b - \mathcal{B}_0) - n^{1/2}(\widehat{\mathcal{B}}_p - \mathcal{B}_0)\} + o_p(1); \\ R_n &= n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) \{P_3(Z_{ij}) - \int P_3(z) \mathcal{G}(z, Z_{ij}) dz + \int \theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{G}(z, Z_{ij}) dz\}. \end{aligned}$$

Now refer to  $Q(z_1, z_2)$  defined above (6). Let  $f_{j|k}(\bullet)$  be the conditional density of  $Z_j$  given  $Z_k$ . Then

$$\begin{aligned} & \int \theta_{\mathcal{B}}(z, \mathcal{B}_0) Q(z, z_0) dz \\ &= \int \sum_{j=1}^m \sum_{k \neq j}^m \{f_k(z_0)/\Omega(z_0)\} f_{j|k}(z|z_0) E\{\mathcal{L}_{jk\theta}(\bullet) \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0) | Z_j = z, Z_k = z_0\} dz \\ &= \sum_{j=1}^m \sum_{k \neq j}^m \{f_k(z_0)/\Omega(z_0)\} E\{\mathcal{L}_{jk\theta}(\bullet) \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0) | Z_k = z_0\} \\ &= \sum_{j=1}^m \sum_{k \neq j}^m \{f_j(z_0)/\Omega(z_0)\} E\{\mathcal{L}_{jk\theta}(\bullet) \theta_{\mathcal{B}}(Z_k, \mathcal{B}_0) | Z_j = z_0\} = P_3(z_0), \end{aligned}$$

the next-to-last equality following from the fact that  $\mathcal{L}_{jk\theta}(\bullet) = \mathcal{L}_{kj\theta}(\bullet)$ . This means that by the definition of  $\mathcal{A}(\bullet)$  defined just above (6),

$$\int \theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{G}(z, z_0) dz = P_3(z_0) - \int \theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{A}(\mathcal{G}, z, z_0) dz,$$

and hence that

$$R_n = n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\bullet) \int \{\theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{A}(\mathcal{G}, z, Z_{ij}) - P_3(z) \mathcal{G}(z, Z_{ij})\} dz.$$

We thus show that  $R_n = 0$  if we can show that for all  $z_0$ ,

$$0 = \int \{\theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{A}(\mathcal{G}, z, z_0) - P_3(z) \mathcal{G}(z, z_0)\} dz.$$

Now,

$$\begin{aligned} & \int \theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{A}(\mathcal{G}, z, z_0) dz \\ &= \int \sum_{j=1}^m \sum_{k \neq j}^m \theta_{\mathcal{B}}(z, \mathcal{B}_0) f_j(z) E\{\mathcal{L}_{jk\theta}(\bullet) \mathcal{G}(Z_k, z_0)/\Omega(Z_k) | Z_j = z\} dz \\ &= \int \sum_{j=1}^m \sum_{k \neq j}^m \theta_{\mathcal{B}}(z, \mathcal{B}_0) f_{jk}(z, z_*) \{\mathcal{G}(z_*, z_0)/\Omega(z_*)\} E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z, Z_k = z_*\} dz dz_*. \end{aligned}$$

Now remember though that  $f_{jk}(z, z_*) = f_{kj}(z_*, z)$  and  $\mathcal{L}_{jk\theta}(\bullet) = \mathcal{L}_{kj\theta}(\bullet)$ . This means that

$$\int \theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{A}(\mathcal{G}, z, z_0) dz$$

$$\begin{aligned}
&= \int \sum_{j=1}^m \sum_{k>j}^m \frac{\theta_{\mathcal{B}}(z) \mathcal{G}(z_*, z_0)}{\Omega(z_*)} \left[ f_{jk}(z, z_*) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z, Z_k = z_*\} \right. \\
&\quad \left. + f_{kj}(z, z_*) E\{\mathcal{L}_{kj\theta}(\bullet) | Z_k = z, Z_j = z_*\} \right] dz dz_* \\
&= \int \sum_{j=1}^m \sum_{k>j}^m \frac{\theta_{\mathcal{B}}(z) \mathcal{G}(z_*, z_0)}{\Omega(z_*)} \left[ f_{jk}(z, z_*) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z, Z_k = z_*\} \right. \\
&\quad \left. + f_{jk}(z_*, z) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z_*, Z_k = z\} \right] dz dz_*.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\int P_3(z_*) \mathcal{G}(z_*, z_0) dz_* \\
&= \int \sum_{j=1}^m \sum_{k \neq j}^m \frac{\mathcal{G}(z_*, z_0) f_j(z_*)}{\Omega(z_*)} E\{\mathcal{L}_{jk\theta}(\bullet) \theta_{\mathcal{B}}(Z_k, \mathcal{B}_0) | Z_j = z_*\} dz_* \\
&= \int \sum_{j=1}^m \sum_{k \neq j}^m \frac{\mathcal{G}(z_*, z_0) \theta_{\mathcal{B}}(z, \mathcal{B}_0)}{\Omega(z_*)} f_{jk}(z_*, z) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z_*, Z_k = z\} dz dz_* \\
&= \int \sum_{j=1}^m \sum_{k>j}^m \frac{\mathcal{G}(z_*, z_0) \theta_{\mathcal{B}}(z, \mathcal{B}_0)}{\Omega(z_*)} [f_{jk}(z_*, z) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z_*, Z_k = z\} \\
&\quad + f_{kj}(z_*, z) E\{\mathcal{L}_{kj\theta}(\bullet) | Z_k = z_*, Z_j = z\}] dz_* dz \\
&= \int \sum_{j=1}^m \sum_{k>j}^m \frac{\mathcal{G}(z_*, z_0) \theta_{\mathcal{B}}(z, \mathcal{B}_0)}{\Omega(z_*)} [f_{jk}(z_*, z) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z_*, Z_k = z\} \\
&\quad + f_{kj}(z_*, z) E\{\mathcal{L}_{kj\theta}(\bullet) | Z_k = z_*, Z_j = z\}] dz_* dz \\
&= \int \sum_{j=1}^m \sum_{k>j}^m \frac{\mathcal{G}(z_*, z_0) \theta_{\mathcal{B}}(z, \mathcal{B}_0)}{\Omega(z_*)} [f_{jk}(z_*, z) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z_*, Z_k = z\} \\
&\quad + f_{jk}(z, z_*) E\{\mathcal{L}_{jk\theta}(\bullet) | Z_j = z, Z_k = z_*\}] dz_* dz \\
&= \int \theta_{\mathcal{B}}(z, \mathcal{B}_0) \mathcal{A}(\mathcal{G}, z, z_0) dz,
\end{aligned}$$

thus completing the argument.

## A.7 Sketch Proof of (A.11)

Here we derive an asymptotic expansion of  $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B}) - \theta_{\mathcal{B}}(z; \mathcal{B})$ . The idea is to work with the estimating equation of  $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B})$  directly. We first derive the 1st degree polynomial kernel estimating equation for  $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B})$ . Differentiating (11) with respect  $\mathcal{B}$  gives the linear kernel estimating equation for  $\theta_{\mathcal{B}}(z; \mathcal{B})$ . Let  $\Theta_i(\tilde{Z}_i, \mathcal{B}) = \{\theta(Z_{i1}, \mathcal{B}), \dots, \theta(Z_{im}, \mathcal{B})\}^T$  and  $\Theta_{i\mathcal{B}}(\tilde{Z}_i, \mathcal{B}) = \{\theta_{\mathcal{B}}(Z_{i1}, \mathcal{B}), \dots, \theta_{\mathcal{B}}(Z_{im}, \mathcal{B})\}^T$ . Denote

the estimating function

$$e_{ij}(\tilde{Y}_i, \tilde{X}_i, \Theta_i, \Theta_{i\mathcal{B}}) = \mathcal{L}_{ij\theta\mathcal{B}}(\bullet) + \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\bullet)\theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}), \quad (\text{A.14})$$

where  $\bullet = \{\tilde{Y}_i, \tilde{X}_i, \theta(Z_{i1}, \mathcal{B}), \dots, \theta(Z_{im}, \mathcal{B})\}$ . Note that (A.14) is the same as  $\epsilon_{ij}^{\#}(\theta, \mathcal{B})$  defined in Section 3.3, but as shown below a slightly different notation is needed in our arguments. Then  $\sum_{j=1}^m E\{e_j(\bullet)|Z_j = z\}f_j(z) = 0$ , see (A.7). The kernel estimating equation for  $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B})$  can be written as

$$R_n = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z)G_{ij}(z, h)e_{ij}\{\tilde{Y}_i, \tilde{X}_i, \hat{\Theta}_{ij}(z, \tilde{Z}_i, \mathcal{B}), \hat{\Theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B})\} = 0, \quad (\text{A.15})$$

where

$$\begin{aligned} \hat{\Theta}_{ij}(z, \tilde{Z}_i, \mathcal{B}) &= \{\hat{\theta}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}(z, \mathcal{B}) + h\hat{\theta}^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h, \dots, \hat{\theta}(Z_{im}, \mathcal{B})\}^{\text{T}}; \\ \hat{\Theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B}) &= \{\hat{\theta}_{\mathcal{B}}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}_{\mathcal{B}}(z, \mathcal{B}) + h\hat{\theta}_{\mathcal{B}}^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h, \dots, \hat{\theta}_{\mathcal{B}}(Z_{im}, \mathcal{B})\}^{\text{T}}. \end{aligned}$$

Denote  $e_{ijk\theta}(\bullet) = \partial e_{ij}(\bullet)/\partial\theta(Z_{ik})$ . An expansion of  $R_n$  about  $\Theta_i(\tilde{Z}_i, \mathcal{B})$  gives that  $\hat{\theta}_{\mathcal{B}}(t, \mathcal{B})$  satisfies

$$R_n = R_{1n} + R_{2n} + o_p(n^{-1/2}) + O_p(h^3) = 0,$$

where

$$\begin{aligned} R_{1n} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z)G_{ij}(z, h)e_{ij}\{\tilde{Y}_i, \tilde{X}_i, \Theta_i(\tilde{Z}_i, \mathcal{B}), \hat{\Theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B})\}; \\ R_{2n} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z)G_{ij}(z, h) \\ &\quad \left[ e_{ijj\theta}\{\tilde{Y}_i, \tilde{X}_i, \Theta_i(\tilde{Z}_i, \mathcal{B}), \hat{\Theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B})\} \left\{ \hat{\theta}(z, \mathcal{B}) + h\hat{\theta}^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h - \theta(Z_{ij}, \mathcal{B}) \right\} \right. \\ &\quad \left. + \sum_{k \neq j} e_{ijk\theta}\{\tilde{Y}_i, \tilde{X}_i, \Theta_i(\tilde{Z}_i, \mathcal{B}), \hat{\Theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B})\} \left\{ \hat{\theta}(Z_{ik}, \mathcal{B}) - \theta(Z_{ik}, \mathcal{B}) \right\} \right]. \end{aligned}$$

A further expansion of  $R_{2n}$  about  $\Theta_{i\mathcal{B}}(\tilde{Z}_i, \mathcal{B})$  gives

$$\begin{aligned} R_{2n} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z)G_{ij}(z, h) \\ &\quad \left[ e_{ijj\theta}(\bullet) \left\{ \hat{\theta}(z, \mathcal{B}) + h\hat{\theta}^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h - \theta(Z_{ij}, \mathcal{B}) \right\} \right. \\ &\quad \left. + \sum_{k \neq j} e_{ijk\theta}(\bullet) \left\{ \hat{\theta}(Z_{ik}, \mathcal{B}) - \theta(Z_{ik}, \mathcal{B}) \right\} \right] + o_p(n^{-1/2}) + O_p(h^3), \end{aligned}$$

where  $\bullet = \{\tilde{Y}_i, \tilde{X}_i, \Theta_i(\tilde{Z}_i, \mathcal{B}), \Theta_{i\mathcal{B}}(\tilde{Z}_i, \mathcal{B})\}$ . Note that the residual  $o_p(n^{-1/2}) + O_p(h^3)$  is due to the fact that the leading residual term involves higher order products  $\{\hat{\theta}(Z_{ik}, \mathcal{B}) - \theta(Z_{ik}, \mathcal{B})\}\{\hat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}) - \theta_{\mathcal{B}}(Z_{ij}, \mathcal{B})\}$ , which are of order  $o_p(n^{-1/2}) + O_p(h^3)$ .

We first focus on  $R_{2n}$  and find its Taylor expansion. One can rewrite  $R_{2n}$  as

$$R_{2n} = R_{21n} + R_{22n} + R_{23n} + o_p(n^{-1/2}) + O_p(h^3), \quad (\text{A.16})$$

where

$$\begin{aligned} R_{21n} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) e_{ijj\theta}(\bullet) \times \\ &\quad \left\{ \hat{\theta}(z, \mathcal{B}) + h\hat{\theta}^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h - \theta(z, \mathcal{B}) - h\theta^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h \right\}; \\ R_{22n} &= -n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) e_{ijj\theta}(\bullet) \\ &\quad \times \left\{ \theta(Z_{ij}, \mathcal{B}) - \theta(z, \mathcal{B}) - h\theta^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h \right\}; \\ R_{23n} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) \sum_{k \neq j} e_{ijk\theta}(\bullet) \left\{ \hat{\theta}(Z_{ik}, \mathcal{B}) - \theta(Z_{ik}, \mathcal{B}) \right\}. \end{aligned}$$

Using the expansion of  $\hat{\theta}(t, \mathcal{B}) - \theta(t, \mathcal{B})$  given in equation (8), some detailed calculations give

$$\begin{aligned} R_{21n} &= -n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \epsilon_{ij} \tilde{\Omega}(z, \mathcal{B}) / \Omega(z, \mathcal{B}) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} \mathcal{G}(z, Z_{ij}, \mathcal{B}) \tilde{\Omega}(z, \mathcal{B}) / \Omega(z) \\ &\quad + (h^2/2) \tilde{\Omega}(z, \mathcal{B}) b(z, \mathcal{B}) + o_p(n^{-1/2}) + O_p(h^3) \\ R_{22n} &= -(h^2/2) \tilde{\Omega}(z, \mathcal{B}) \theta^{(2)}(z, \mathcal{B}) + O_p(h^3) \\ R_{23n} &= -n^{-1} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} D(z, Z_{ij}) \\ &\quad + (h^2/2) \sum_{j=1}^m \sum_{k \neq j}^m E \{ e_{jk\theta}(\bullet) b(Z_k, \mathcal{B}) | Z_j = z \} f_j(z) + o_p(n^{-1/2}) + O_p(h^3), \end{aligned}$$

where

$$\begin{aligned} \tilde{\Omega}(z, \mathcal{B}) &= \sum_{j=1}^m E \{ e_{jj\theta}(\bullet) | Z_j = z \} f_j(z); \\ D(z_1, z_2) &= \sum_{j=1}^m \sum_{k \neq j}^m E \left\{ \frac{e_{jk\theta}(\bullet)}{\Omega(Z_k, \mathcal{B})} \middle| Z_j = z_1, Z_k = z_2 \right\} f_{jk}(z_1, z_2) \\ &\quad - \sum_{j=1}^m \sum_{k \neq j}^m E \left\{ \frac{e_{jk\theta}(\bullet) \mathcal{G}(Z_k, z_2, \mathcal{B})}{\Omega(Z_k, \mathcal{B})} \middle| Z_j = z_1 \right\} f_j(z). \end{aligned}$$

One can easily show that  $\tilde{\Omega}(z, \mathcal{B}) = \partial \Omega(z, \mathcal{B}) / \partial \mathcal{B}$ . Combining these three expansions, we have

$$R_{2n} = -\frac{h^2}{2} \tilde{b}(z, \mathcal{B}) + n^{-1} \sum_{i=1}^n \epsilon_{ij} \tilde{\mathcal{C}}(z, \mathcal{B}) - n^{-1} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} \tilde{\mathcal{G}}(z, Z_{ij}, \mathcal{B}) + o_p(n^{-1/2}) + O_p(h^3), \quad (\text{A.17})$$

where

$$\tilde{b}(z, \mathcal{B}) = \tilde{\Omega}(z, \mathcal{B}) \left\{ \theta^{(2)}(z, \mathcal{B}) - b(z, \mathcal{B}) \right\} - \sum_{j=1}^m \sum_{k=1}^m E \{ e_{jk\theta}(\bullet) b(Z_k, \mathcal{B}) | Z_j = z \} f_j(z);$$

$$\begin{aligned}
\tilde{\mathcal{C}}(z, \mathcal{B}) &= -\tilde{\Omega}(z, \mathcal{B})/\Omega(z, \mathcal{B}); \\
\tilde{\mathcal{G}}(z_1, z_2, \mathcal{B}) &= \sum_{j=1}^m \sum_{k \neq j}^m E \left\{ \frac{e_{jk\theta}(\bullet)}{\Omega(Z_k, \mathcal{B})} \middle| Z_j = z_1, Z_k = z_2 \right\} f_{jk}(z_1, z_2) \\
&\quad - \sum_{j=1}^m \sum_{k=1}^m E \left\{ \frac{e_{jk\theta}(\bullet) \mathcal{G}(Z_k, z_2, \mathcal{B})}{\Omega(Z_k, \mathcal{B})} \middle| Z_j = z \right\} f_j(z).
\end{aligned}$$

Now revisit the kernel estimating equation of  $\hat{\theta}_{\mathcal{B}}(t, \mathcal{B})$  given in (A.15), which can be rewritten as  $R_{1n}\{\hat{\theta}_{\mathcal{B}}(\cdot), \hat{\theta}_{\mathcal{B}}^{(1)}(\cdot)\} = -R_{2n} + o_p(n^{-1/2}) + O_p(h^3)$ . Note that the estimators  $\{\hat{\theta}_{\mathcal{B}}(\cdot), \hat{\theta}_{\mathcal{B}}^{(1)}(\cdot)\}$  only enter into  $R_{1n}$ , and that the right hand side of  $R_{2n}$  as given in (A.17) does not involve these unknown estimators.

First consider the solution  $\hat{\theta}_{\mathcal{B}}^*(t, \mathcal{B})$  of the kernel estimating equation  $R_{1n}\{\hat{\theta}_{\mathcal{B}}(\cdot), \hat{\theta}_{\mathcal{B}}^{(1)}(\cdot)\} = 0$ , i.e.,

$$\begin{aligned}
R_{1n}\{\hat{\theta}_{\mathcal{B}}(\cdot), \hat{\theta}_{\mathcal{B}}^{(1)}(\cdot)\} &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) e_{ij} \times \\
&\quad \{e_{ij}(\tilde{Y}_i, \tilde{X}_i, \Theta_i(\tilde{Z}_i, \mathcal{B}), \hat{\theta}_{\mathcal{B}}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}_{\mathcal{B}}(z, \mathcal{B}) + h\hat{\theta}_{\mathcal{B}}^{(1)}(z, \mathcal{B})(Z_{ij} - z)/h, \dots, \hat{\theta}_{\mathcal{B}}(Z_{im}, \mathcal{B}))\}.
\end{aligned} \tag{A.18}$$

One can easily see that equation (A.18) takes the same form as the kernel estimating equation for  $\hat{\theta}(t, \mathcal{B})$  in (11) except that  $\mathcal{L}_{ij}(\bullet)$  in (11) is replaced by  $e_{ij}(\bullet)$ , and  $\{\hat{\theta}_{\mathcal{B}}(\cdot), \hat{\theta}_{\mathcal{B}}^{(1)}(\cdot)\}$  are now unknown parameters. Its solution  $\hat{\theta}_{\mathcal{B}}^*(t, \mathcal{B})$  hence takes the same Taylor expansion as (8) in the paper except that  $\epsilon_{ij}$  in (8) is replaced by  $e_{ij}$  and  $\{b(z, \mathcal{B}), \Omega(z, \mathcal{B}), \mathcal{G}(z_1, z_2, \mathcal{B})\}$  in (8) are modified by replacing  $\mathcal{S}_{jk}$  by  $\partial e_j(\tilde{\eta}, \tilde{\delta})/\partial \delta_k$  in their definitions. Suppose we call these modified terms  $\{b_{\mathcal{B}}(z, \mathcal{B}), \Omega_{\mathcal{B}}(z, \mathcal{B}), \mathcal{G}_{\mathcal{B}}(z_1, z_2, \mathcal{B})\}$ .

Using (A.14), one can easily see that  $\partial e_j(\tilde{\eta}, \tilde{\delta})/\partial \theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}) = \mathcal{L}_{jk\theta}(\bullet)$ . It follows that  $\Omega_{\mathcal{B}}(z, \mathcal{B}) = \Omega(z, \mathcal{B})$ ,  $\mathcal{G}_{\mathcal{B}}(z_1, z_2, \mathcal{B}) = \mathcal{G}(z_1, z_2, \mathcal{B})$  and  $b_{\mathcal{B}}(z, \mathcal{B})$  differs from  $b(z, \mathcal{B})$  by replacing  $\theta^{(2)}(t, \mathcal{B})$  by  $\theta_{\mathcal{B}}^{(2)}(t, \mathcal{B})$ , i.e.,  $b_{\mathcal{B}}(z, \mathcal{B})$  solves

$$b_{\mathcal{B}}(z, \mathcal{B}) = \theta_{\mathcal{B}}^{(2)}(z, \mathcal{B}) - \Lambda(b_{\mathcal{B}}, z, \mathcal{B}).$$

It follows that the expansion of  $\hat{\theta}_{\mathcal{B}}^*(t, \mathcal{B}) - \theta_{\mathcal{B}}(t, \mathcal{B})$  is

$$\begin{aligned}
\hat{\theta}_{\mathcal{B}}^*(t, \mathcal{B}) - \theta_{\mathcal{B}}(t, \mathcal{B}) &= (h^2/2)b_{\mathcal{B}}(z, \mathcal{B}) - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) e_{ij}/\Omega(z, \mathcal{B}) \\
&\quad + n^{-1} \sum_{i=1}^m \sum_{j=1}^m e_{ij} \mathcal{G}(z, Z_{ij}, \mathcal{B})/\Omega(z, \mathcal{B}) + o_p(n^{-1/2}) + O_p(h^3).
\end{aligned} \tag{A.19}$$

Now study how  $\hat{\theta}_{\mathcal{B}}^*(t, \mathcal{B})$  and  $\hat{\theta}_{\mathcal{B}}(t, \mathcal{B})$  are related. The estimator  $\hat{\theta}_{\mathcal{B}}(t, \mathcal{B})$  solves

$$R_{1n}\{\hat{\theta}_{\mathcal{B}}(t, \mathcal{B}), \hat{\theta}_{\mathcal{B}}^{(1)}(t, \mathcal{B})\} = -R_{2n} + o_p(n^{-1/2}) + O_p(h^3)$$

instead of  $R_{1n}\{\hat{\theta}_B(t, \mathcal{B}), \hat{\theta}_B^{(1)}(t, \mathcal{B})\} = 0$ , where  $R_{2n}$  is given in (A.17). Using the same iterative proof as that used for  $\hat{\theta}(t, \mathcal{B})$ , one can easily see that the Taylor expansion of  $\hat{\theta}_B(t, \mathcal{B}) - \theta_B(t, \mathcal{B})$  only differs from  $\hat{\theta}_B^*(t, \mathcal{B}) - \theta_B(t, \mathcal{B})$  by adding an extra expansion  $-R_{2n}/\Omega(z)$  to the expansion of  $\hat{\theta}_B^*(t, \mathcal{B})$ . Combining the expansions (A.17) and (A.19), the Taylor expansion of  $\hat{\theta}_B(t, \mathcal{B}) - \theta_B(t, \mathcal{B})$  is, to terms of order  $o_p(n^{-1/2}) + O_p(h^3)$ ,

$$\begin{aligned} \hat{\theta}_B(t, \mathcal{B}) - \theta_B(t, \mathcal{B}) &= (h^2/2) \left\{ b_B(z, \mathcal{B}) + \tilde{b}(z, \mathcal{B})/\Omega(z, \mathcal{B}) \right\} \\ &\quad - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) e_{ij} / \Omega(z, \mathcal{B}) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m e_{ij} \mathcal{G}(z, Z_{ij}, \mathcal{B}) / \Omega(z, \mathcal{B}) \\ &\quad - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \epsilon_{ij} \tilde{C}(z, \mathcal{B}) / \Omega(z, \mathcal{B}) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \epsilon_{ij} \tilde{\mathcal{G}}(z, Z_{ij}, \mathcal{B}) / \Omega(z, \mathcal{B}), \end{aligned}$$

where  $b_B(\cdot)$ ,  $\tilde{C}(\cdot)$ ,  $\tilde{\mathcal{G}}(\cdot)$  are defined above, and  $\Omega(\cdot)$ ,  $\mathcal{G}(\cdot)$  are defined in the paper.

Note that interestingly,  $\tilde{C}(z, \mathcal{B})/\Omega(z, \mathcal{B}) = \partial\Omega^{-1}(z, \mathcal{B})/\partial\mathcal{B}$ . This suggests that the above Taylor expansion takes the same form as that obtained by differentiating the Taylor expansion of  $\theta(z, \mathcal{B})$  in equation (8) of the paper with respect to  $\mathcal{B}$  except that  $\partial b(z, \mathcal{B})/\partial\mathcal{B}$  is not  $b_B(z, \mathcal{B}) + \tilde{C}(z, \mathcal{B})\tilde{b}(z, \mathcal{B})$ , and  $\partial\{\mathcal{G}(z, Z_{ij}, \mathcal{B})\Omega^{-1}(z, \mathcal{B})\}/\partial\mathcal{B}$  is not  $\tilde{\mathcal{G}}(z, Z_{ij}, \mathcal{B})/\Omega(z, \mathcal{B})$ . The other terms are the same.

## A.8 Computation of $\hat{\theta}_B(z, \mathcal{B})$

Equation (A.15) can be used to show that  $\hat{\theta}_B(z, \mathcal{B})$  can be computed by a similar algorithm as that used to compute  $\hat{\theta}(z, \mathcal{B})$ . If we refer to equation (2) of Lin, et al. (2004), we can make the following substitutions. First replace their  $B_{ij}^T(t)V^{-1}Y_i$  by the  $G_{ij}(z, h)\mathcal{L}_{ij\theta\mathcal{B}}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{ij}(z, \tilde{Z}_i, \mathcal{B}), \mathcal{B}\}$ . Then replace  $B_{ij}^T(t)V^{-1}\mu_{i(j)}(t)$  by  $G_{ij}(z, h)\sum_{k=1}^m \mathcal{L}_{ijk\theta\mathcal{B}}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{ij}(z, \tilde{Z}_i, \mathcal{B}), \mathcal{B}\}\hat{\theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B})$ . Although this is a vector form rather than the scalar form in Lin, et al., their same method can be used to find an explicit, closed form solution for  $\hat{\theta}_B(z, \mathcal{B})$ .

## A.9 Explicit Algorithm for Method in Section 5.1

Equation (19) can be rewritten as

$$\sum_{i=1}^n \sum_{j=1}^{L_i} K_h(Z_{ij}^* - z_0) G_{ij}(z_0) [e_{ij}^T \Sigma_i^{jj} \{\mathcal{Y}_{ij} - G_{ij}(z_0)^T \alpha e_{ij}\} + e_{ij}^T \sum_{k \neq j}^{L_i} \Sigma_i^{jk} \{\mathcal{Y}_{ik} - \hat{\theta}_{[\ell-1]}(Z_{ik}^*) e_{ik}\}],$$

where  $\mathcal{Y}_{ij} = (\mathcal{Y}_{ij1}, \dots, \mathcal{Y}_{ijm_{ij}})^T$  is a  $m_{ij} \times 1$  vector and  $\mathcal{Y}_i = (\mathcal{Y}_{i1}^T, \dots, \mathcal{Y}_{iL_i}^T)^T$ . It follows that

$$\left\{ \sum_{i=1}^n \sum_{j=1}^{L_i} K_h(Z_{ij}^* - z_0) G_{ij}(z_0) e_{ij}^T \Sigma_i^{jj} e_{ij} G_{ij}^T(z_0) \right\} \hat{\alpha} \quad (\text{A.20})$$

$$= \sum_{i=1}^n \sum_{j=1}^{L_i} K_h(Z_{ij}^* - z_0) G_{ij}(z_0) \{e_{ij}^T \Sigma_i^{jj} e_{ij} \hat{\theta}_{[\ell-1]}(Z_{ij}^*) + e_{ij}^T \sum_{k=1}^{L_i} \Sigma_i^{jk} (\mathcal{Y}_{ik} - \hat{\theta}_{[\ell-1]}(Z_{ik}^*) e_{ik})\}.$$

Denote by  $M = \sum_{i=1}^n \sum_{j=1}^{L_i} m_{ij}$  the total sample size and  $L = \sum_{i=1}^n L_i$  the total number of family members, i.e., the number of levels of the second hierarchical level). Let  $\tilde{G}(z_0) = \{G_{11}(z_0), \dots, G_{nL_n}(z_0)\}^T$ , which is a  $L \times p$  design matrix,  $\tilde{Z} = (Z_{11}^*, \dots, Z_{nL_n}^*)^T$  be a  $L \times 1$  vector containing distinct observed values of  $Z$ 's,  $K_{dh}(z_0) = \text{diag}\{K_h(Z_{11}^* - z_0), \dots, K_h(Z_{nL_n}^* - z_0)\}$ , which is a  $L \times L$  matrix,  $E = \text{diag}(e_{11}, \dots, e_{nL_n})$ , which is an  $M \times L$  matrix,  $\tilde{\Sigma}^d = \text{diag}(\Sigma_1^d, \dots, \Sigma_n^d)$  and  $\Sigma_i^d = \text{diag}(\Sigma_i^{11}, \dots, \Sigma_i^{L_i L_i})$ , and  $\tilde{\Sigma} = \text{diag}(\Sigma_1, \dots, \Sigma_n)$ ,  $\mathcal{Y} = (\mathcal{Y}_1^T, \dots, \mathcal{Y}_n^T)^T$ . Notice that  $\hat{\theta}^{(l+1)}(z_0) = \hat{\alpha}_0$ . Writing equation (A.20) in a matrix form, simple calculations show that

$$\begin{aligned} \hat{\theta}^{(l+1)}(z_0) &= \delta^T \left\{ \tilde{G}(z_0)^T K_{dh}(z_0) E^T \tilde{\Sigma}^d E \tilde{G}(z_0) \right\}^{-1} \tilde{G}(z_0)^T K_{dh}(z_0) \\ &\quad \times \left\{ E^T \tilde{\Sigma}^{-1} \mathcal{Y} + E^T (\tilde{\Sigma}^d - \tilde{\Sigma}^{-1}) E \hat{\theta}_{[\ell-1]}(\tilde{Z}^*) \right\}, \end{aligned}$$

where  $\delta = (1, 0, \dots, 0)^T$ . Let  $K_{wh}^T(z_0) = \delta^T \left\{ \tilde{G}(z_0)^T K_{dh}(z_0) E^T \tilde{\Sigma}^d E \tilde{G}(z_0) \right\}^{-1} \tilde{G}(z_0)^T K_{dh}(z_0)$ , and  $K_w = \{K_{wh}(Z_{11}^*), \dots, K_{wh}(Z_{nL_n}^*)\}^T$ , which is a  $L \times L$  matrix. Then we have

$$\hat{\theta}^{(l+1)}(\tilde{Z}^*) = K_w \left\{ E^T \tilde{\Sigma}^{-1} \mathcal{Y} + E^T (\tilde{\Sigma}^d - \tilde{\Sigma}^{-1}) E \hat{\theta}_{[\ell-1]}(\tilde{Z}^*) \right\}.$$

Write  $\hat{\theta}_{[\ell]}(\tilde{Z}^*) = \mathcal{S}_{[\ell]} E^T \tilde{\Sigma}^{-1} \mathcal{Y}$ . Note that  $\mathcal{S}_{[\ell]}$  is a  $L \times L$  square matrix. At convergence  $\mathcal{S}_{[\ell]} \rightarrow \mathcal{S}$ , where  $\mathcal{S}$  satisfies  $\mathcal{S} = K_w \{I + E^T (\tilde{\Sigma}^d - \tilde{\Sigma}^{-1}) E \mathcal{S}\}$ . It follows that  $\mathcal{S} = \{I + K_w E^T (\tilde{\Sigma}^{-1} - \tilde{\Sigma}^d) E\}^{-1} K_w$ . Hence at convergence

$$\hat{\theta}(\tilde{Z}^*) = \{I + K_w E^T (\tilde{\Sigma}^{-1} - \tilde{\Sigma}^d) E\}^{-1} K_w E^T \tilde{\Sigma}^{-1} \mathcal{Y}. \quad (\text{A.21})$$

If  $m_{ij} \equiv 1$  then  $E = I$ . The results then reduce to those in Lin et al (2004).

Notice that  $E$ ,  $\tilde{\Sigma}^{-1}$  and  $\tilde{\Sigma}^d$  are all block diagonal matrices. The above matrix calculations can then be greatly simplified. Specifically, partition  $K_w$  as an  $n \times n$  block matrix with the  $(i, i')$ <sup>th</sup> block denoted by  $K_{w,ii'}$  which is a  $L_i \times L_i$  matrix. Write  $E = \text{diag}(E_1, \dots, E_n)$  and  $K_{dh} = \text{diag}\{K_{dh,1}, \dots, K_{dh,n}\}$ , where  $E_i = \text{diag}(e_{i1}, \dots, e_{iL_i})$  and  $K_{dh,i}(z_0) = \text{diag}\{K_h(Z_{i1}^* - z_0), \dots, K_h(Z_{iL_i}^* - z_0)\}$ . Write  $\tilde{G}(z_0) = \{\tilde{G}_i(z_0)^T, \dots, \tilde{G}_n(z_0)^T\}^T$ . Then

$$K_{wh}^T(z_0) = \delta^T \left\{ \sum_{i=1}^n \tilde{G}_i(z_0)^T K_{dh,i}(z_0) E_i^T \tilde{\Sigma}_i^d E_i \tilde{G}_i(z_0) \right\}^{-1} \{\tilde{G}_1(z_0)^T K_{dh,1}(z_0), \dots, \tilde{G}_n(z_0)^T K_{dh,n}(z_0)\}.$$

For equation (A.21), partition the matrix  $K_w E^T (\tilde{\Sigma}^{-1} - \tilde{\Sigma}^d) E$  in the same fashion as  $K_w$  into an  $n \times n$  block matrix.

## References

- Begun, J. H., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.*, 11, 432-452.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models, *Journal of the American Statistical Association*, 88, 9-25.
- Carroll, R. J., Härdle, W. and Mammen, E. (2002). Estimation in an additive model when components are linked parametrically. *Econometric Theory*, 18, 886-912.
- Chen, K. and Jin, Z. (2001). Local polynomial regression analysis of clustered data. Preprint.
- Chen, X., Linton, O., Keilegom, I. V. (2003) Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71, 1591-1608.
- Fan, J. and Li, R. (2004) New estimation and model selection procedures for semiparametric modeling in longitudinal data. *Journal of the American Statistical Association*, 99, 710-723.
- Hafner, C. M. (1998). *Nonlinear Time Series Analysis With Applications To Foreign Exchange Rate Volatility*. Heidelberg: Physica.
- Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika*, 88, 973-985.
- 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.
- Hu, Z., Wang, N. & Carroll, R. J. (2004). Profile-kernel versus backfitting in the partially linear model for longitudinal/clustered data. *Biometrika*, 91, 251-262.
- 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.
- Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96, 1045-1056.
- 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.
- Lin, X., Wang, N., Welsh, A. H. and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for longitudinal/clustered data. *Biometrika*, 91, 177-194.
- Linton, O. B. and Nielson, J. P. (1995). A kernel method for estimating structured nonparametric

- regression based on marginal integration. *Biometrika*, 82, 93-101.
- Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57, 253-259.
- Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression (Corr: 96V91 p1380). *Journal of the American Statistical Association*, 90, 1257-1270.
- Schaid, D. J. (1999). Case-parents design for gene-environment interaction. *Genetic Epidemiology*, 16, 261-273.
- Severini, T. A. and Staniswalis, J. G. (1994). Quasilikelihood 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.
- Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, in press.
- Wang, Y. (1998). Mixed-effects smoothing spline ANOVA. *Journal of the Royal Statistical Society, Series B*, 60, 159-174.
- Wild, C. J. and Yee, T. W. (1996). Additive extensions to generalized estimating equation methods. *Journal of the Royal Statistical Society, Series B*, 58, 711-725.
- Wu, H. and Zhang, J. Y. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97, 883-897.
- 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.
- Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93, 710-719.

Table 1 Profile-kernel estimates regression coefficients of the semiparametric model (17) applied to the Kenya hemoglobin data

	Working Independence	Structured Covariance (Ignoring ties)	Structured Covariance (Accounting for ties)
Month	-0.418(0.0378 <sup>a</sup> )	-0.397(0.039 <sup>b</sup> )(0.043 <sup>c</sup> )	-0.397(0.039 <sup>b</sup> )(0.043 <sup>c</sup> )
(Month-4) <sub>+</sub>	0.147(0.028)	0.129(0.028)(0.028)	0.129(0.028)(0.028)
Sex	-0.122(0.072)	-0.122(0.080)(0.087)	-0.122(0.080)(0.087)
LNPDEN	-0.010(0.013)	-0.009(0.015)(0.017)	-0.009(0.014)(0.016)

<sup>a</sup>: Naive SE ignoring correlation

<sup>b</sup>: Model-based SE

<sup>c</sup>: Sandwich SE

### List of Figure

Figure 1. Estimated nonparametric curve of the effect of mother age at birth on child hemoglobin by fitting the semiparametric model (17) to the Kenya hemoglobin data. The solid line is the efficient estimate when common  $Z$ -values are ignored, the dashed line is the proposed method, and the dotted line is the working independence fit.

