

Semiparametric estimation in general repeated measures problems

Xihong Lin

Harvard School of Public Health, Boston, USA

and Raymond J. Carroll

Texas A&M University, College Station, USA

[Received January 2005. Revised September 2005]

Summary. The 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 or 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. Although generally not true, it transpires that 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 methods proposed are evaluated by using simulation studies and applied to the Kenya haemoglobin data.

Keywords: Clustered and longitudinal data; Generalized estimating equations; Generalized linear mixed models; Kernel method; Marginal models; Measurement error; Nonparametric regression; Partially linear model; Profile method; Semiparametric efficient score; Semiparametric information bound; Time-dependent covariate

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 obtain some sense of the generality of our approach, consider the following examples, all of which can be solved by using our approach. 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 is known. In contrast, the fifth example has a large literature.

1.1. Example 1

One of the most common designs in epidemiology is the matched case–control study, which is 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

Address for correspondence: Xihong Lin, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA.
E-mail: xlin@hsph.harvard.edu

the 1–1 matched study, we consider 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 semi-parametric 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 $H(v) = 1/\{1 + \exp(-v)\}$ is the logistic distribution function, b_i is a nuisance parameter 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 by using a nonparametric smooth function $\theta_0(\cdot)$. 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 we do not want to model the stratum effects b_i , inference is based on the conditional likelihood function

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

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

1.2. Example 2

Hafner (1998) and Carroll *et al.* (2002) studied

$$Y_i = \sum_{j=1}^m \beta_0^{j-1} \theta_0(Z_{ij}) + \varepsilon_i,$$

a model that arises in finance. The algorithm that was 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.

1.3. 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 $\mu(\cdot)$ is the inverse of a link function and $b_i = \text{normal}(0, \sigma_0^2)$. Here the variance component σ_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.

1.4. Example 4

As discussed in a data example in Section 5.1.2, consider problems in which family i has m children, each of whom have a base-line 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 $\mu(\cdot)$, and a covariance matrix Σ reflecting the structure of the problem. In this case, note that the function $\theta_0(\cdot)$ is evaluated m times for different children per family.

1.5. Example 5

Consider a repeated measures Gaussian partially linear problem where for the i 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 $\mu(\cdot)$ and a true but unknown function $\theta_0(z)$,

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

where, given $(\tilde{X}_i, \tilde{Z}_i)$, $\tilde{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})^T$ has mean 0 and covariance matrix $\Sigma(\tau_0)$ for a parameter τ_0 . Note that the function $\theta_0(\cdot)$ 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 $\theta_0(\cdot)$ 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 $\theta_0(\cdot)$ 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 β_0 : these are semiparametric efficient when $\tilde{\varepsilon}_i$ is actually Gaussian.

These examples can be placed into a common framework. There is a *criterion function* $\mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}, \mathcal{B})$, 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

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

For example, consider the model that is given in equation (2). Here, $\mathcal{B}_0 = (\beta_0, \tau_0)$ and the criterion function is the Gaussian log-likelihood

$$-\frac{1}{2} \log[\det\{\Sigma(\tau_0)\}] - \frac{1}{2} (\tilde{Y} - \tilde{X}\beta_0 - \tilde{\eta}_0)^T \Sigma^{-1}(\tau_0) (\tilde{Y} - \tilde{X}\beta_0 - \tilde{\eta}_0).$$

The criterion function in example 1 is given in equation (1), and examples 2–4 also have explicit forms.

In this paper, we show how to compute efficient estimators of the nonparametric component $\theta_0(\cdot)$ for problems with and without the parametric component \mathcal{B}_0 . The method that is defined in Section 2 is based on a likelihood-type generalization of the basic kernel method of Wang (2003) to the general problem (3). The methods are applicable to likelihood and non-likelihood problems, the only constraint being that condition (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. Lin and Carroll (2001) and Wang *et al.* (2005) proposed estimating-equation-based profile kernel methods for a marginal generalized semiparametric model that was similar to the normal model (2) for clustered data. Hu *et al.* (2004) proposed a backfitting method under the normal model (2). Our profile likelihood method and the backfitting algorithm are likelihood-type extensions of the methods of Wang *et al.* (2005) and Hu *et al.* (2004) to the general setting in expression (3). We show that in our case, using the smoother of Section 2, profiling and backfitting have identical limit distributions. The 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 so (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 $\mathcal{L}(\cdot)$ is a likelihood function, and we 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 that was considered by Wang *et al.* (2005), the covariance matrix $\Sigma(\tau_0)$ depends on a parameter τ_0 . The parameter τ_0 is conveniently estimated by the simple device of ignoring the correlation of the data, forming residuals from the fit and then using the 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 by using the Kenya haemoglobin data and a simulation study. The second case that is considered in Section 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 the appendices and detailed proofs can be found at <http://www.bepress.com/harvardbiostat> and also at <http://www.stat.tamu.edu/~carroll/papers.php>.

2. The nonparametric case

Before describing methods for the general semiparametric problem, we describe methods when there is no parametric component, which is a problem of interest in its own right. In the nonparametric case, the criterion function is $\mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}) = \mathcal{L}\{\tilde{Y}, \tilde{X}, \theta(Z_1), \dots, \theta(Z_m)\}$ where $\eta_j = \theta(Z_j)$ ($j = 1, \dots, m$). Define $\mathcal{L}_{j\theta}(\cdot) = \partial \mathcal{L}(\tilde{Y}, \tilde{X}, \eta_1, \dots, \eta_m) / \partial \eta_j$ and $\mathcal{L}_{jk\theta}(\cdot) = \partial^2 \mathcal{L}(\tilde{Y}, \tilde{X}, \eta_1, \dots, \eta_m) / \partial \eta_j \partial \eta_k$ ($j, k = 1, \dots, m$). 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 $K(\cdot)$ 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 $\theta(\cdot)$ by solving the kernel estimating equation

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

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 $\theta(\cdot)$ at the $(l-1)$ th step is $\hat{\theta}_{[l-1]}(\cdot)$. Then $\hat{\theta}_{[l]}(z) = \hat{\alpha}_0$, where $(\hat{\alpha}_0, \hat{\alpha}_1)$ solve

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

At convergence, $\hat{\theta}(z)$ solves the kernel estimating equation (4). In Gaussian cases such as in examples 1 and 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 $\mathcal{L}(\cdot) = \mathcal{L}\{\tilde{Y}, \tilde{X}, \theta(Z_1), \dots, \theta(Z_m)\}$, and similarly for its derivatives. Make the definitions

$$\Omega(z) = \sum_{j=1}^m f_j(z) E\{\mathcal{L}_{jj\theta}(\cdot) | Z_j = z\}$$

and

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

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

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

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

2.1. 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,

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

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

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

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

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

Remark 1. Equations (8)–(10) agree with the results of Wang (2003) in the special cases that were considered by her. In equation (8), since the first two terms are of order $O_p\{h^2 + (nh)^{-1/2}\}$ whereas the third is of order $O_p(n^{-1/2})$, the first two terms dominate. The proof of result (8) is similar to that of Wang (2003) and is given in the technical report that was mentioned at the end of Section 1.

Remark 2. Note that equation (9) has design-density-dependent bias. It is possible to remove this. Suppose that 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 that is 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 result (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 1 and equation (24) in Appendix A.1 starting from 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 $\mathcal{L}(\cdot)$ is a log-likelihood 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

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

To solve equation (11) we suggest the following iterative algorithm. Suppose that the current estimate in the iteration is $\hat{\theta}_{[l-1]}(z, \mathcal{B})$. Then we update to $\hat{\theta}_{[l]}(z, \mathcal{B})$ by solving (α_0, α_1) in the equation

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

Set $\hat{\theta}_{[l]}(z, \mathcal{B}) = \alpha_0$. 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

$$\sum_{i=1}^n \mathcal{L} \{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B} \}.$$

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 Appendix A.6, we show how to use an algorithm that is very similar to equation (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 owing 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 \mathcal{B} in the function

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

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

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

where

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

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 showed that in example 5 and equation (2) the use of the smoother that is defined in equation (5) leads to profiling and backfitting being asymptotically equivalent. Thus we would conjecture that the same equivalence holds in our general problem, a conjecture which is 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 $\mathcal{L}(\cdot)$ is a likelihood function.

3.2.1. Result 2: semiparametric efficiency bound

Assume that $(\tilde{Y}_i, \tilde{X}_i, \tilde{Z}_i)$ are independent and identically distributed, and that $\mathcal{L}(\cdot)$ is a likelihood function conditional on (\tilde{X}, \tilde{Z}) . Then the optimal semiparametric score function is

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

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

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

The proof of result (13) is given in Appendix A.2.

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 $\mathcal{L}(\cdot)$. 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 $\mathcal{L}_{j\theta\mathcal{B}}(\cdot) = \partial \mathcal{L}_{j\theta}(\tilde{Y}, \tilde{X}, \eta_1, \dots, \eta_m, \mathcal{B}) / \partial \mathcal{B}$, and

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

As we show in Appendix A.4, $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B}_0) = \theta_{\mathcal{B}}(z, \mathcal{B}_0) + o_p(1)$, where $\theta_{\mathcal{B}}(z, \mathcal{B}_0)$ satisfies

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

Define

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

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

3.3.1. Result 3: profile kernel method

Assume that $(\tilde{Y}_i, \tilde{X}_i, \tilde{Z}_i)$ are independent and identically distributed, and that $0 = E\{\mathcal{L}_B(\cdot)|\tilde{Z}\} = E\{\mathcal{L}_{j\theta}(\cdot)|\tilde{Z}\}$. Suppose further that the bandwidth $h \propto n^{-c}$ with $\frac{1}{5} \leq c \leq \frac{1}{3}$. Then

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

where $\varepsilon_{ij} = \mathcal{L}_{ij\theta}(\cdot)$ and \mathcal{V} is defined in equation (14). In the case that $\mathcal{L}(\cdot)$ is a log-likelihood 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 result (16) is given in Appendix A.4.

3.3.2. 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.5.

3.3.3. Result 5: covariance matrix estimation

Consistent estimates of \mathcal{F} and \mathcal{V} can be constructed as follows. Let $\hat{\mathcal{L}}_{iB}$, $\hat{\mathcal{L}}_{ij\theta}$, $\hat{\mathcal{L}}_{iBB}$ and $\hat{\mathcal{L}}_{ij\theta B}$ be the estimated versions of the quantities indicated. Let $\hat{\theta}_B(Z_{ij}, \hat{\mathcal{B}})$ be the solution of the kernel estimating equation (36) in Appendix A.6. Then a consistent estimator of \mathcal{V} is the sample covariance matrix of the terms

$$\hat{\mathcal{L}}_{iB} + \sum_{j=1}^m \hat{\mathcal{L}}_{ij\theta} \hat{\theta}_B(Z_{ij}, \hat{\mathcal{B}}).$$

Further, a consistent estimator of \mathcal{F} is

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

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 Σ_{ε_0} . In our notation, $\mathcal{B}_0 = \{\beta_0^T, \text{vec}^T(\Sigma_{\varepsilon_0})\}^T$. Wang *et al.* (2004) provided an initial estimate $\hat{\Sigma}_{\varepsilon p}$ of Σ_{ε_0} , and then applied our algorithm only to β while pretending that Σ_{ε_0} is known and equal to $\hat{\Sigma}_{\varepsilon p}$.

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 Appendix A.4 at equation (31), 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

In this section, we provide two examples to illustrate applications of our methods in the general likelihood-type framework that was described in Section 1. Our first example concerns multilevel hierarchical data where inference is based on a likelihood, whereas the second example is on longitudinal data with covariates that are measured with error where the likelihood inference is difficult and a non-likelihood criterion function is used.

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 base-line 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

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

where $i = 1, \dots, n$ (e.g. the i th family), $j = 1, \dots, L_i$ (e.g. the j th member in the i th family), $k = 1, \dots, m_{ij}$ (e.g. the k th time point). Equation (17) models the effect of the base-line subject level covariate Z_{ij}^* nonparametrically and other covariates X_{ijk} parametrically. Denote the covariance matrix of ε_i by Σ_i , which is a $\sum_{j=1}^{L_i} m_{ij} \times \sum_{j=1}^{L_i} m_{ij}$ matrix. Assuming that Σ_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} is an $m_{ij} \times 1$ vector of 1s. Let $\varepsilon_{ij} = (\varepsilon_{ij1}, \dots, \varepsilon_{ijm_{ij}})^T$, $\varepsilon_i = (\varepsilon_{i1}^T, \dots, \varepsilon_{iL_i}^T)^T$ and $\tilde{\varepsilon} = (\varepsilon_1^T, \dots, \varepsilon_n^T)^T$. Now partition Σ_i as follows: the (jk) th block $\Sigma_{i,jk} = \text{cov}(\varepsilon_{ij}, \varepsilon_{ik})$ and the dimension of $\Sigma_{i,jk}$ is $m_{ij} \times m_{ik}$. Denote $\Sigma_i^{-1} = \{\Sigma_i^{jk}\}$, where the partition of Σ_i^{-1} is the same as Σ_i . Chen and Jin (2005) considered a problem that was similar to our setting without the parametric component and proposed to apply Wang's (2003) smoothing algorithm, pretending that the repeated base-line values of Z_{ij}^* from the same subject were distinct over time. 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 β , 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 ε_{ij} , ε_i and $\tilde{\varepsilon}$. Define $Z_i^* = (Z_{i1}^*, \dots, Z_{iL_i}^*)^T$ and $\tilde{Z}^* = (Z_{11}^*, \dots, Z_{m, L_n}^*)^T$ and define $\tilde{X} = (X_{11}, \dots)^T$. Then, the linear kernel estimating equation at the l 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 $G_{ij}(z)$ is defined in Section 2 and

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

In Appendix A.7, we give an explicit closed form solution to equation (19): no iteration is necessary, and equation (19) is only a descriptive device. Indeed, we derive an explicit form of a

smoother matrix \mathcal{S} such that $\hat{\theta}(\tilde{Z}^*, \beta) = \mathcal{S} \tilde{Y}(\beta) = \mathcal{S} \tilde{Y} - \mathcal{S} \tilde{X} \beta$, where \mathcal{S} is given in equation (38). This means that the profile kernel estimator of β 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} that is 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 six observations per cluster, with $Z_{i1} = Z_{i2} = Z_{i3}$ and $Z_{i4} = Z_{i5} = Z_{i6}$, 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 that the correlation structure is autoregressive with correlation 0.60 between repeated measures over time and common between-subject (within-family) correlation 0.20: let Σ denote the resulting covariance matrix. The true function was $\theta_0(z) = \sin(8z - 2)$. The Z -values were generated as independent uniform distributions, whereas the X -values were bivariate independent uniform distributions 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 that were selected by using the method of Ruppert *et al.* (1995). The covariance matrix $\hat{\Sigma}$ of the ε_{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 β we formed $Y_{ij} - X_{ij}^T \beta$ and then calculated $\hat{\theta}(\cdot)$ by using the closed form expression (38). With \mathcal{S} as the smoother matrix, $\text{cov}\{\hat{\theta}(\cdot)\}$ 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 β was calculated by using the closed form formula (20).

In 1000 simulated data sets, both weighted methods achieved over 70% greater mean-squared error efficiency for estimating β_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 haemoglobin data

We applied our method to analyse a subset of the Kenya haemoglobin data to study the changes in haemoglobin level over time in the first year since birth and the risk factors of haemoglobin among Kenyan 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. Haemoglobin level was measured at each visit and visit times varied from child to child. The risk factors of interest include the mother's age at child birth, child sex and placental parasitemia density PDEN, a marker for malaria, which could affect haemoglobin. Log-transformation was applied to PDEN to make the normality assumption plausible. A preliminary analysis showed that the effect of

mother's age was non-linear. 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} to be the mother's age at birth, $X_{ijk} = \{\text{sex}, \text{logpden}, \text{month}, (\text{month}-4)_+\}$, where $\text{sex} = 1$ if female and $\text{sex} = 0$ if male, $\text{logpden} = \log(\text{PDEN}+1)$, the function $f_+ = f$ if $f > 0$ and $f_+ = 0$ if $f \leq 0$. Note that the terms $\{\text{months}, (\text{month}-4)_+\}$ model the time effect as a piecewise 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 Σ as obtained from a preliminary regression spline fit and then estimated the bandwidth by using leaving one mother out cross-validation, and thus obtained estimates of $\theta_0(\cdot)$ and β_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 the haemoglobin level. 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 cross-validation bandwidth was 0.23. The correlation was low or moderate in this example. In Fig. 1, we compared the estimated non-

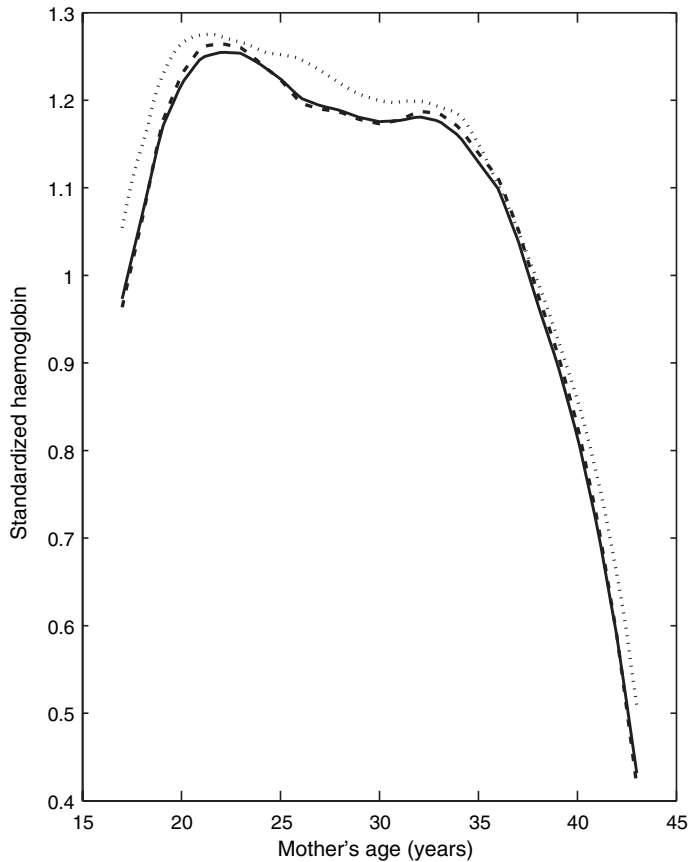


Fig. 1. Estimated nonparametric curve of the effect of mother's age at birth on child haemoglobin by fitting the semiparametric model (17) to the Kenya haemoglobin data: —, efficient estimate when common Z-values are ignored; - - - - -, method proposed; ·····, working independence fit

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

	Coefficients from the following models:		
	Working independence	Structured covariance (ignoring ties)	Structured covariance (accounting for ties)
Month	-0.418 (0.0378†)	-0.397 (0.039‡) (0.043§)	-0.397 (0.039‡) (0.043§)
(Month-4) ₊	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)

†Naïve standard error ignoring correlation.

‡Model-based standard error.

§Sandwich standard error.

parametric 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 or without accounting for ties in mother's age). The estimated curves were similar. Children's haemoglobin increased with mother's age at birth for mothers who were younger than 22 years old, then decreased slightly with mother's age until at age early 30 years and then started decreasing quickly with mother's age, indicating that children are likely to have much lower haemoglobin levels if mothers give birth after early 30 years of age, i.e. giving birth after early 30 years is likely to increase children's risk of anaemia (low haemoglobin) considerably.

As expected, since the correlation was not high, the estimates of the regression coefficients β were roughly the same for the working independence kernel fit with bandwidths selected by using the method of Ruppert *et al.* (1995), the method of Wang *et al.* (2004) ignoring the common Z -values and our method accounting for the common Z -values. 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. The haemoglobin level drops quickly after birth and decreases at a slower rate after month 4. Neither sex nor placental parasitemia density affects the haemoglobin level significantly.

5.2. Measurement error models

Here we consider the multivariate partially linear measurement error model

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

where $\tilde{\varepsilon}_i$ has covariance matrix $\Sigma_{\varepsilon 0}$. 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 0 and the property that $\text{cov}\{\text{vec}(\tilde{U}_i)\} = \Sigma_{u0}$, which is 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 $G(\Sigma_{\varepsilon}, \Sigma_{u0}) = E(\tilde{U}^T \Sigma_{\varepsilon}^{-1} \tilde{U})$ and define $\mathcal{K}(\Sigma_{u0}, \beta) = E(\tilde{U} \beta \beta^T \tilde{U}^T)$. Note that

$$\beta^T G(\Sigma_{\varepsilon}, \Sigma_{u0}) \beta = \text{tr}\{\Sigma_{\varepsilon}^{-1} E(\tilde{U} \beta \beta^T \tilde{U}^T)\} = \text{tr}\{\Sigma_{\varepsilon}^{-1} \mathcal{K}(\Sigma_{u0}, \beta)\}.$$

In equation (21), $\mathcal{B} = (\beta, \tau, \Sigma_{\varepsilon})$ and the criterion function is

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

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

For symmetric matrices Σ , $\partial\{\log(|\Sigma|)\}/\partial\Sigma = 2\Sigma^{-1} - \text{diag}(\Sigma^{-1})$ and $\partial\{\text{tr}(\Sigma A)\}/\partial\Sigma = 2A - \text{diag}(A)$. It is readily seen that the derivative of expression (22) with respect to β , Σ_ϵ and θ evaluated at the true parameters has expectation 0, and thus expression (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 (2003) 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) showed, there is a smoother matrix $S = S(\Sigma_\epsilon)$ such that $\hat{\theta}(\mathcal{Z}, \mathcal{B}) = S(\mathcal{Y} - \mathcal{W}\beta)$. 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 $\mathcal{W}_* = (I - S)\mathcal{W}$ and $\mathcal{Y}_* = (I - S)\mathcal{Y}$. Then, for given Σ_ϵ , the profile estimate of β is given by

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

A simple estimate of Σ_ϵ is to form the working independence estimate of β and to apply equation (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 that are 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 problem: 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. We also showed how to use pseudolikelihood methods within our context when some of the parameters are more conveniently estimated by alternative algorithms. In a very different problem, namely nonparametric regression of additive models, Mammen *et al.* (1999) proposed a ‘smooth backfitting’ algorithm that does not require undersmoothing. It is of future research interest to extend this method to our setting.

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 condition (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 equation (12) and also satisfying condition (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 Ψ 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 th component of θ . It can be shown that result 1 and equation (8) still hold with the same notation, as does the fundamental identity (15). The basic backfitting expansion (30) in Appendix A.3, as well as the definition of \mathcal{F} in result 3, also holds with \mathcal{L} replaced by Ψ . 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 equation (16), because it involves the implicitly defined function \mathcal{G} in equation (6). However, the bootstrap method that bootstraps clusters can be used to estimate the covariance of $\hat{\mathcal{B}}$ (Chen *et al.*, 2003).

Acknowledgements

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 A: Sketch of technical arguments

Detailed proofs are given in the technical report at <http://www.bepress.com/harvardbiostat/> and also at <http://www.stat.tamu.edu/~carroll/papers.php>.

A.1. A key technical lemma

Lemma 1. Let $\hat{\theta}_{[l]}(\cdot)$ be the estimate at the l th stage of the iteration. Then

$$\begin{aligned} \hat{\theta}_{[l]}(z) - \theta_0(z) &= \frac{h^2}{2} b_{[0]}(z) - n^{-1} \sum_{i=1}^n \sum_{j=1}^m \frac{K_h(Z_{ij} - z) \varepsilon_{ij}}{\Omega(z)} - n^{-1} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j}^m \frac{K_h(Z_{ij} - z)}{\Omega(z)} \\ &\quad \times \mathcal{L}_{ijk\theta}(\cdot) \{ \hat{\theta}_{[l-1]}(Z_{ik}) - \theta_0(Z_{ik}) \} + o_p(n^{-1/2}), \end{aligned} \tag{24}$$

where $b_{[0]}(z) = \theta^{(2)}(z)$, and the argument is $\{ \tilde{Y}_i, \tilde{X}_i, \theta(z_{i1}), \dots, \alpha_0 + \alpha_1(z_{ij} - z)/R, \dots, \theta(z_{im}) \}$. Here is a brief sketch of equation (24). By Taylor series expansion, we have

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) \mathcal{L}_{ij\theta}(\cdot) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) G_{ij}^T(z, h) \\ &\quad \times \mathcal{L}_{ijj\theta}(\cdot) \begin{pmatrix} \hat{\alpha}_0 - \alpha_0 \\ \hat{\alpha}_1 - \alpha_1 \end{pmatrix} + o_p(n^{-1/2}), \end{aligned}$$

where the argument is $\{ \tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[l-1]}(Z_{i1}), \dots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \dots, \hat{\theta}_{[l-1]}(Z_{im}) \}$. 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 2×2 identity matrix (again, this is because K has variance 1.0). Hence,

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

Some calculation shows that

$$A_{1n} = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \varepsilon_{ij} + (h^2/2) b_{[0]}(z) \Omega(z) + o_p(n^{-1/2}),$$

and A_{2n} is equal to the third term in equation (24).

A.2. Proof of result 2: semiparametric efficient score

We use Begun *et al.* (1983). In their set-up, their ‘ f ’ is our $\exp(\mathcal{L})$, their ‘ θ ’ is our \mathcal{B} and their ‘ g ’ is our θ . It is easily derived that their ‘ $2\rho_\theta/f^{1/2}$ ’ is our $\mathcal{L}_\mathcal{B}$. Similarly, for an arbitrary function $\gamma(\cdot)$, their ‘ $2A\beta/f^{1/2}$ ’ is $\sum_{j=1}^m \mathcal{L}_{j\theta}(\cdot) \gamma(Z_j)$. This means that their equation (3.1) is the following. The semiparametric optimal score is of the form

$$\mathcal{L}_\mathcal{B}(\cdot) - \sum_{j=1}^m \mathcal{L}_{j\theta}(\cdot) \gamma_*(Z_j),$$

where $\gamma_*(\cdot)$ is such that, for all $\gamma(\cdot)$,

$$0 = E\left[\left\{\mathcal{L}_\mathcal{B}(\cdot) - \sum_{j=1}^m \mathcal{L}_{j\theta}(\cdot) \gamma_*(\cdot)\right\} \sum_{k=1}^m \mathcal{L}_{k\theta}(\cdot) \gamma(Z_k)\right]. \quad (25)$$

We now show that $\gamma_*(\cdot) = -\theta_\mathcal{B}(\cdot)$ satisfies condition (25). To see this, interchange the indices j and k and note that condition (25) means that we must show that for arbitrary $\gamma(\cdot)$

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

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

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

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

$$\begin{aligned} 0 &= \sum_{j=1}^m E[\gamma(Z_j) \{\mathcal{L}_{j\theta\mathcal{B}}(\cdot) + \sum_{k=1}^m \mathcal{L}_{jk\theta}(\cdot) \theta_\mathcal{B}(Z_k)\}] \\ &= \sum_{j=1}^m E\{\gamma(Z_j) \varepsilon_{ij}^\#(\theta_0, \mathcal{B}_0)\}, \end{aligned} \quad (26)$$

where $\varepsilon_{ij}^\#(\theta_0, \mathcal{B}_0)$ is defined in Section 3.3. This last step follows by conditioning the expectation in equation (26) on Z_j and then applying equation (29) below.

A.3. Sketch proof of equation (15): fundamental identity

Since

$$n^{-1} \sum_{i=1}^n \{(Z_i - z)/h\} K_h(Z_i - z) = o_p(1)$$

one can show that, for any \mathcal{B} ,

$$0 = \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \mathcal{L}_{j\theta} \{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \dots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B}\}. \quad (27)$$

Differentiating equation (27) with respect to \mathcal{B} , we obtain

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

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

Recall the definition of $\varepsilon_{ij}^\#(\theta, \mathcal{B})$ that is given in Section 3.3. Define

$$H_j(z) = E\{\varepsilon_{ij}^\#(\theta_0, \mathcal{B}_0) | Z_j = z\}. \quad (28)$$

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

$$0 = E\left\{\sum_{j=1}^m B(Z_j) H_j(Z_j)\right\}. \quad (29)$$

We shall use this equality repeatedly.

A.4. 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 $\mathcal{F}_2 = E\{\sum_{j=1}^m \mathcal{L}_{j\theta\mathcal{B}}(\cdot) \theta_{\mathcal{B}}^\top(Z_j, \mathcal{B}_0)\}$. Also, define

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

It is an easy consequence of equation (29) 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 $0 = A_1(\hat{\mathcal{B}}_p, \hat{\theta}) + A_2(\hat{\mathcal{B}}_p, \hat{\theta})$, where

$$A_1(\hat{\mathcal{B}}_p, \hat{\theta}) = n^{-1/2} \sum_{i=1}^n \mathcal{L}_{i\mathcal{B}_p} \{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \hat{\mathcal{B}}_p), \dots, \hat{\theta}(Z_{im}, \hat{\mathcal{B}}_p), \hat{\mathcal{B}}_p\},$$

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

A Taylor series expansion shows that

$$A_1(\hat{\mathcal{B}}_p, \hat{\theta}) = n^{-1/2} \sum_{i=1}^n \mathcal{L}_{i\mathcal{B}}(\cdot) + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta\mathcal{B}}(\cdot) \{\hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta(Z_{ij})\} + (\mathcal{F}_1 + \mathcal{F}_2)n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) + o_p(1), \quad (30)$$

where the symbol ‘ \cdot ’ here means evaluated at θ and \mathcal{B}_0 . Similarly, we have that

$$\begin{aligned} A_2(\hat{\mathcal{B}}_p, \hat{\theta}) &= n^{-1} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta\mathcal{B}}(\cdot) \theta_{\mathcal{B}}^\top(Z_{ij}) n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\cdot) \theta_{\mathcal{B}\mathcal{B}}(Z_{ij}) n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) \\ &\quad + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\cdot) \theta_{\mathcal{B}}(Z_{ij}) \theta_{\mathcal{B}}^\top(Z_{ik}) n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) \\ &\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta} \{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}_0), \dots, \hat{\theta}(Z_{im}, \mathcal{B}_0), \mathcal{B}_0\} \hat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) + o_p(1). \end{aligned}$$

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

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

Recall that $\varepsilon_{ij} = \mathcal{L}_{ij\theta}(\cdot)$ and $H_j(z)$ as defined in equation (28). If

$$P_{ij} = \mathcal{L}_{ij\theta\mathcal{B}}(\cdot) + \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\cdot) \theta_{\mathcal{B}}(Z_{ik}),$$

we have

$$\begin{aligned}
 -\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) &= n^{-1/2} \sum_{i=1}^n \{ \mathcal{L}_{i\mathcal{B}} + \sum_{j=1}^m \varepsilon_{ij} \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0) \} + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m H_j(Z_{ij}) \{ \hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta(Z_{ij}) \} \\
 &\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \{ P_{ij} - H_j(Z_{ij}) \} \{ \hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta(Z_{ij}) \} \\
 &\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \mathcal{L}_{ij\theta}(\cdot) \{ \hat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) - \theta_{\mathcal{B}}(Z_{ij}) \} + o_p(1).
 \end{aligned} \tag{31}$$

We can show that the last three terms of equation (31) are all $o_p(1)$. The proof of the last term uses the asymptotic expansion

$$\begin{aligned}
 \hat{\theta}_{\mathcal{B}}(z, \mathcal{B}_0) - \theta_{\mathcal{B}}(z, \mathcal{B}_0) &= (h^2/2) \{ b_1(z) + b_2(z) \} - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \varepsilon_{ij} \Omega_1(z) \\
 &\quad - n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) \varepsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0) \Omega_2(z) + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \varepsilon_{ij} \mathcal{G}_1(z, Z_{ij}) \\
 &\quad + n^{-1} \sum_{i=1}^n \sum_{j=1}^m \varepsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0) \mathcal{G}_2(z, Z_{ij}) + o_p(n^{-1/2}),
 \end{aligned} \tag{32}$$

for some functions $b_j(\cdot)$, $\Omega_j(\cdot)$ and $\mathcal{G}_j(\cdot)$ ($j = 1, 2$). The detailed proofs are given in the technical report that is mentioned at the end of Section 1.

A.5. Sketch proof of result 4: asymptotic distribution for backfitting

Using the notation of Appendix A.4, for backfitting we are solving the equation $0 = A_1(\hat{\mathcal{B}}_b, \hat{\theta})$. Using the results in Appendix A.4, we have

$$-\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_b - \mathcal{B}_0) = n^{-1/2} \sum_{i=1}^n \mathcal{L}_{i\mathcal{B}}(\cdot) - n^{-1/2} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\cdot) \theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}_0) \{ \hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta_0(Z_{ij}) \}. \tag{33}$$

Since the profile estimator satisfies

$$-\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) = n^{-1/2} \sum_{i=1}^n \{ \mathcal{L}_{i\mathcal{B}}(\cdot) + \sum_{j=1}^m \mathcal{L}_{ij\theta}(\cdot) \theta_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) \} + o_p(1), \tag{34}$$

we see that we must show that the second terms in equations (33) and (34) are asymptotically equivalent. Make the definitions

$$\Omega(z_0) = \sum_{j=1}^m f_j(z_0) E \{ \mathcal{L}_{jj\theta}(\cdot) | Z_j = z_0 \},$$

$$P_1(z_0) = \sum_{j=1}^m f_j(z_0) E \{ \mathcal{L}_{j\theta\mathcal{B}}(\cdot) | Z_j = z_0 \} / \Omega(z_0),$$

$$P_2(z_0) = E \left[\sum_{j=1}^m E \{ \mathcal{L}_{j\theta\mathcal{B}}(\cdot) | Z_j \} \mathcal{G}(Z_j, z_0) / \Omega(Z_j) \right],$$

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

Recalling equation (15), we see that

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

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

We now plug in result (8) into the second term of equation (33). Noting the assumption that $nh^4 \rightarrow 0$, some calculation shows that this second term is asymptotically equivalent to $C_{n1} + C_{n2}$, where

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

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

$$-\mathcal{F}\{n^{1/2}(\hat{\mathcal{B}}_b - \mathcal{B}_0) - n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0)\} = S_n + o_p(1)$$

where

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

Now, using the definition of $\mathcal{Q}(z_1, z_2)$ above equation (6) and the definition of $\mathcal{A}(\cdot)$ just above equation (6), one can show that

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

Hence

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

We thus need to show that, for all z_0 ,

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

Its proof is given in the technical report that was mentioned at the end of Section 1.

A.6. Computation of $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$

We first derive the first-degree polynomial kernel estimating equation for $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B})$. Differentiating equation (11) with respect to \mathcal{B} gives the linear kernel estimating equation for $\theta_{\mathcal{B}}(z; \mathcal{B})$. Let $\theta(z, \mathcal{B})$ be the asymptotic limit of $\hat{\theta}(z, \mathcal{B})$. 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

$$e_{ij}(\tilde{Y}_i, \tilde{X}_i, \Theta_i, \Theta_{i\mathcal{B}}) = \mathcal{L}_{ij\theta\mathcal{B}}(\cdot) + \sum_{k=1}^m \mathcal{L}_{ijk\theta}(\cdot) \theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}), \quad (35)$$

where $\cdot = \{\tilde{Y}_i, \tilde{X}_i, \theta(Z_{i1}, \mathcal{B}), \dots, \theta(Z_{im}, \mathcal{B})\}$. Equation (35) is the same as $\varepsilon_{ij}^{\#}(\theta, \mathcal{B})$ that was defined in Section 3.3, but as shown below a slightly different notation is needed in our arguments. Then

$$\sum_{j=1}^m E\{e_{ij}(\cdot) | Z_{ij} = z\} f_j(z) = 0;$$

see equation (29). The kernel estimating equation for $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B})$ can be written as

$$R_n = n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(Z_{ij} - z) G_{ij}(z, h) e_{ij}\{\tilde{Y}_i, \tilde{X}_i, \hat{\Theta}_{ij}(z, \tilde{Z}_i, \mathcal{B}), \hat{\Theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B})\} = 0, \quad (36)$$

where

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

Equation (36) can be used to show that $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$ has the asymptotic expansion (32), and it can be computed by a similar algorithm to that which was 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 $G_{ij}(z, h) \mathcal{L}_{ij\theta\mathcal{B}}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{ij}(z, \tilde{Z}_i, \mathcal{B}), \mathcal{B}\}$. 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.* (2004), their same method can be used to find an explicit, closed form solution for $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$.

A.7. Explicit algorithm for method in Section 5.1

Equation (19) can be rewritten as

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

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

$$\begin{aligned} & \left\{ \sum_{i=1}^n \sum_{j=1}^{L_i} K_h(Z_{ij}^* - z_0) G_{ij}(z_0) e_{ij}^T \sum_i^{jj} e_{ij} G_{ij}^T(z_0) \right\} \hat{\alpha} \\ &= \sum_{i=1}^n \sum_{j=1}^{L_i} K_h(Z_{ij}^* - z_0) G_{ij}(z_0) \left[e_{ij}^T \sum_i^{jj} e_{ij} \hat{\theta}_{[l-1]}(Z_{ij}^*) + e_{ij}^T \sum_{k=1}^{L_i} \sum_i^{jk} \{\mathcal{Y}_{ik} - \hat{\theta}_{[l-1]}(Z_{ik}^*) e_{ik}\} \right]. \end{aligned} \quad (37)$$

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 $G(z_0) = (G_{11}(z_0), \dots, G_{nL_n}(z_0))^T$, which is an $L \times p$ design matrix, $\tilde{Z} = (Z_{11}^*, \dots, Z_{nL_n}^*)^T$ be an $L \times 1$ 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 an $L \times L$ matrix, $E = \text{diag}(e_{11}, \dots, e_{nL_n})$, which is an $M \times L$ matrix, $\tilde{\Sigma}^d = \text{diag}(\Sigma_1^d, \dots, \Sigma_n^d)$, $\Sigma_i^d = \text{diag}(\Sigma_i^{11}, \dots, \Sigma_i^{L_i L_i})$ and $\tilde{\Sigma} = \text{diag}(\Sigma_1, \dots, \Sigma_n)$, $\mathcal{Y} = (\mathcal{Y}_1^T, \dots, \mathcal{Y}_n^T)^T$. Note that $\hat{\theta}^{(l+1)}(z_0) = \hat{\alpha}_0$. Writing equation (37) in a matrix form, simple calculation shows that

$$\hat{\theta}^{(l+1)}(z_0) = \delta^T \{ \tilde{G}(z_0)^T K_{dh}(z_0) E^T \tilde{\Sigma}^d E \tilde{G}(z_0) \}^{-1} \tilde{G}(z_0)^T K_{dh}(z_0) \{ E^T \tilde{\Sigma}^{-1} \mathcal{Y} + E^T (\tilde{\Sigma}^d - \tilde{\Sigma}^{-1}) E \hat{\theta}_{[l-1]}(\tilde{Z}^*) \},$$

where $\delta = (1, 0, \dots, 0)^T$. Let

$$K_{wh}^T(z_0) = \delta^T \{ \tilde{G}(z_0)^T K_{dh}(z_0) E^T \tilde{\Sigma}^d E \tilde{G}(z_0) \}^{-1} \tilde{G}(z_0)^T K_{dh}(z_0),$$

and $K_w = (K_{wh}(Z_{11}^*), \dots, K_{wh}(Z_{nL_n}^*))^T$, which is an $L \times L$ matrix. Then we have

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

Write $\hat{\theta}_{[l]}(\tilde{Z}^*) = S_{[l]} E^T \tilde{\Sigma}^{-1} \mathcal{Y}$. Note that $S_{[l]}$ is an $L \times L$ square matrix. At convergence $S_{[l]} \rightarrow S$, where S satisfies

$$S = K_w \{ I + E^T (\tilde{\Sigma}^d - \tilde{\Sigma}^{-1}) \} E S.$$

It follows that

$$S = \{ I + K_w E^T (\tilde{\Sigma}^{-1} - \tilde{\Sigma}^d) E \}^{-1} K_w.$$

Hence at convergence

$$\hat{\theta}(\tilde{Z}^*) = \{ I + K_w E^T (\tilde{\Sigma}^{-1} - \tilde{\Sigma}^d) E \}^{-1} K_w E^T \tilde{\Sigma}^{-1} \mathcal{Y}. \quad (38)$$

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

Note 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') th block denoted

by $K_{w,ij'}$ which is an $L_i \times L_{i'}$ matrix. Write $E = \text{diag}(E_1, \dots, E_n)$ and $K_{dh} = \text{diag}(K_{dh,1}, \dots, K_{dh,n})$, where $E_i = \text{diag}(e_{i1}, \dots, e_{im_i})$ and $K_{dh,i}(z_0) = \text{diag}\{K_h(Z_{i1}^* - z_0), \dots, K_h(Z_{iL_i}^* - z_0)\}$. Write $\tilde{G}(z_0) = (\tilde{G}_1(z_0)^T, \dots, \tilde{G}_n(z_0)^T)^T$. Then

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

For equation (38), 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 and the computation can be simplified in a similar way.

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. *J. Am. Statist. Ass.*, **88**, 9–25.
- Carroll, R. J., Härdle, W. and Mammen, E. (2002) Estimation in an additive model when components are linked parametrically. *Econometr. Theory*, **18**, 886–912.
- Chen, K. and Jin, Z. (2005) Local polynomial regression analysis of clustered data. *Biometrika*, **92**, 59–74.
- Chen, X., Linton, O. and Van Keilegom, I. (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. *J. Am. Statist. Ass.*, **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. and Carroll, R. J. (2004) Profile-kernel versus backfitting in the partially linear model for longitudinal/clustered data. *Biometrika*, **91**, 251–262.
- Lin, D. Y. and Ying, Z. (2001) Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *J. Am. Statist. Ass.*, **96**, 103–126.
- Lin, X. and Carroll, R. J. (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Ass.*, **95**, 520–534.
- Lin, X. and Carroll, R. J. (2001) Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Ass.*, **96**, 1045–1056.
- 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.
- Mammen, E., Linton, O. and Nielsen, J. (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, **27**, 1443–1490.
- 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. *J. Am. Statist. Ass.*, **90**, 1257–1270; correction, **91** (1996), 1380.
- Schaid, D. J. (1999) Case-parents design for gene-environment interaction. *Genet. Epidemiol.*, **16**, 261–273.
- Severini, T. A. and Staniswalis, J. G. (1994) Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Ass.*, **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. *J. Am. Statist. Ass.*, **100**, 147–157.
- Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *J. R. Statist. Soc. B*, **60**, 159–174.
- Wild, C. J. and Yee, T. W. (1996) Additive extensions to generalized estimating equation methods. *J. R. Statist. Soc. B*, **58**, 711–725.
- Wu, H. and Zhang, J. Y. (2002) Local polynomial mixed-effects models for longitudinal data. *J. Am. Statist. Ass.*, **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. *J. Am. Statist. Ass.*, **93**, 710–719.