Statistics and Its Interface Volume 1 (2008) 75–86

Semiparametric latent covariate mixed-effects models with application to a colon carcinogenesis study

Zonghui Hu and Naisyin Wang

We study a mixed-effects model in which the response and the main covariate are linked by position. While the covariate corresponding to the observed response is not directly observable, there exists a latent covariate process that represents the underlying positional features of the covariate. When the positional features and the underlying distributions are parametric, the expectation-maximization (EM) is the most commonly used procedure. Though without the parametric assumptions, the practical feasibility of a semiparametric EM algorithm and the corresponding inference procedures remain to be investigated. In this paper, we propose a semiparametric approach, and identify the conditions under which the semiparametric estimators share the same asymptotic properties as the unachievable estimators using the true values of the latent covariate; that is, the oracle property is achieved. We propose a Monte Carlo graphical evaluation tool to assess the adequacy of the sample size for achieving the oracle property. The semiparametric approach is later applied to data from a colon carcinogenesis study on the effects of cell DNA damage on the expression level of oncogene *bcl-2*. The graphical evaluation shows that, with moderate size of subunits, the numerical performance of the semiparametric estimator is very close to the asymptotic limit. It indicates that a complex EM-based implementation may at most achieve minimal improvement and is thus unnecessary.

Keywords and phrases: Carcinogenesis, Consistency, Generalized estimating equation, Local linear smoothing, Mixed-effects model.

1. INTRODUCTION

1.1 Colon carcinogenesis study

Recent researches on colon cancer have been focusing on linking colon tumor development to the inhibition of apoptosis (cell death; see Heemels et al. 2000). When the body is affected by carcinogen, apoptosis causes termination of the cells with irreparable genetic damages, and thus prevents them from proliferating to cancer cells. It consequently reduces the risk of cancer. Any inhibition of apoptosis, on the other hand, induces cancer development.

An oncogene closely linked to, but adversely affecting, apoptosis is *bcl-2*. Over-expression of *bcl-2* gene leads to suppression of apoptosis, thus allows tumor cells to survive and proliferate. During the initial stage of colon carcinogenesis, few apoptotic cells are formed and the main information about apoptosis is carried by apoptosis related gene, e.g., *bcl-2*. For the purpose of cancer prevention, it would be beneficial that the level of *bcl-2* gene expression decreases as cell DNA damage increases. Therefore, in this study, we focus on investigating the relationship between the cell DNA damage and *bcl-2* gene expression during the initial stage of colon cancer. Our primary interest is how the diet affects this relationship at different time post carcinogen exposure.

We now briefly describe the experiment. Thirty rats were divided evenly into two groups. Each group was fed with one of the two diets, fish oil supplemented or corn oil supplemented, for two weeks. After this, all 30 rats were injected with azoxymethane (AOM), a carcinogen to induce colon cancer. Three rats from each diet group were then euthanized at 0, 3, 6, 9, and 12 hours post injection to measure the cell DNA damage and *bcl-2* gene expression. In labs, the cell DNA damage is measured by the DNA adduct level. For each rat, 20 crypts were selected to measure *bcl-2*, and another group of 15 to 25 crypts were selected to measure the DNA adduct level. These two measurements were taken at each cell within the selected crypts. There are about 14 to 56 cells in each crypt.

Colon crypts are discrete units within the colon where colonic cells replicate. At the bottom of each crypt, there are the stem cells that generate all the cells within the crypt. Daughter cells are formed from stem cells, move up along the crypt and exfoliate into lumen as more cells are created. Thus, a cell's relative position within a crypt is an indicator of its age: cells at the bottom are younger, and cells near the top are older. In this study, the position of the cell was recorded by the relative cell position, as in Morris et al. (2001). The relative cell positions range from 0 at the bottom to 1 at the top.

Our goal is to understand the relationship between the response (*bcl-2* gene expression) and the covariate (cell DNA adduct level), as well as how this relationship changes with the diet. More precisely, we want to investigate, in comparison to the corn oil supplemented diet, whether the fish oil

Table 1. Colon Carcinogenesis Data within One Rat and Under One Treatment Condition: the Left Four Columns Are bcl-2 Observations from 20 Crypts, and the Right Four Columns Are DNA Adduct Observations from 23 Crypts; Crypt j Is the j-th Crypt for Observing bcl-2 and j' Is the j'-th Crypt for Observing DNA Adduct, These Are Two Different Sets of Crypts in the Same Rat.

bcl-2 observations				DNA adduct observations			
crypt (j)	cell	cell position	*bcl-2* reading	crypt (j')	cell	cell position	DNA adduct reading
1	1	0	$y_{1,1}$	1	1	0	$w_{1,1}$
1	2	1/31	$y_{1,2}$	1	2	1/39	$w_{1,2}$
1	3	2/31	$y_{1,3}$	1	3	2/39	$w_{1,3}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	32	1	$y_{1,32}$	1	32	31/39	$w_{1,32}$
				⋮	⋮	⋮	⋮
				1	40	1	$w_{1,40}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	1	0	$y_{20,1}$	20	1	0	$w_{20,1}$
20	2	1/28	$y_{20,2}$	20	2	1/35	$w_{20,2}$
20	3	2/28	$y_{20,3}$	20	3	2/35	$w_{20,3}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	29	1	$y_{20,29}$	20	29	29/35	$w_{20,29}$
				⋮	⋮	⋮	⋮
				20	36	1	$w_{20,36}$
				⋮	⋮	⋮	⋮
				23	1	1/30	$w_{23,1}$
				⋮	⋮	⋮	⋮
				23	31	1	$w_{23,31}$

supplemented diet helps to suppress the increasing trend of *bcl-2* gene expression when DNA damage increases. We need a mixed-effects model to accommodate the diet and time treatment effects, and also the random effects for rat and crypt. The special aspect about this study is: DNA adduct level and *bcl-2* gene expression were not measured in the same crypts, though from the same rats. This is because in this study, once a crypt was euthanized to take DNA adduct measurement, it could not be used again to measure *bcl-2*. Instead, a different crypt from the same rat was used. Since the number of cells varied from crypt to crypt, cells within different crypts had different relative positions. Consequently, the two measurements, *bcl-2* gene expression and DNA adduct level, were observed at different relative crypt depths (i.e., the relative cell positions) in addition to being from different crypts. It formed a problem of misaligned measurements. Conventional regression methods are not appropriate.

To illustrate the data structure, Table 1 lists the portion of data from one rat under one treatment condition (i.e., combination of diet and time post carcinogen exposure). It shows that, within this rat, *bcl-2* was observed from 20 crypts while DNA adduct observed from 23 crypts. These are two different groups of crypts; that is, crypt j for observing *bcl-2* is different from crypt j' for observing adduct, for all j and j'. Within the first crypt ($j = 1$) for *bcl-2* observation, *bcl-2* was measured over 32 cells, while within the first crypt ($j' = 1$) for DNA adduct observation, DNA adduct was measured over 40 cells.

1.2 Statistical background

When the covariate values are not directly available, imputation is the traditional practice, like the nearest neighbor method (NN; Pielou, 1961; Huang and Zhu, 2002) or the last observation carry-forward method(LOCF; Carroll, 2004). In the colon carcinogenesis study, these two methods are, within each rat, to use the DNA adduct values observed nearest to or immediately in front of the positions of the *bcl-2* measurement, respectively, as the matching DNA adduct measurement. The pitfall of the naive imputation approach is the attenuation effects as discussed in Carroll (2004).

Another possible approach is to assume a subject-level latent process $X_i(\cdot)$ for the positional feature behind the observed covariate from subject i and apply the maximum likelihood method through an expectation-maximization algorithm (EM; Dempster, Laird, and Rubin, 1977). The drawback of EM is two-fold. First, maximum likelihood is sensitive to model assumption; Second, due to the complexity

of the colon carcinogenesis data, the derivation of EM is not straightforward and the computational implementation is extremely intensive.

In this paper, we propose a semiparametric method. As in EM algorithm, we consider a latent covariate process $X_i(\cdot)$ for the DNA adduct versus the relative cell position within each rat. Unlike EM, we directly "estimate" the rat-level latent covariate by nonparametric regression and use the estimated DNA adduct for the estimation of the primary model. This semiparametric approach and the NN, LOCF methods are all "plug-in" based methods, which are often unfavorably considered. However, we prove that, under adequate within-subject sample sizes, the estimators from this semiparametric approach are asymptotically consistent and achieve the oracle property. An EM estimator, depending on its construction, may at most share the same asymptotic properties. We later propose a graphical tool to check the adequacy of the "effective" sample sizes.

Our semiparametric approach is basically to model the primary relationship between *bcl-2* and DNA adduct, and meanwhile, the positional feature of DNA adduct level. Joint modeling of parametric longitudinal features and a primary endpoint has been studied extensively, see Tsiatis and Davidian (2001), Li, Zhang, and Davidian (2004) and the works therein. However, identifying a parametric structure for longitudinal biomarkers is not always feasible. In this paper, the joint modeling is extended to accommodate nonparametric longitudinal covariate features and longitudinal response.

The semiparametric approach can also be considered as an extension of Carroll and Wand (1991) and Pepe and Fleming (1991) in that a nonparametric estimation is used to obtain the estimates of the unobserved covariate. However, there are three major differences. First, in these previous works, true covariate values were available for a portion of the subjects, but in our example, there were absolutely no matched measurements. Second, the DNA adduct measurement alone was from a nonparametric mixed-effects model with the marginal mean as a function of the relative cell position. Meanwhile, the response *bcl-2* was also measured repeatedly from the same crypt within the same subject. Therefore, the observations in our example were correlated while the previous works focused on independent cases. Third, the issue of "effective" sample size and the corresponding diagnostic tool had never been investigated in the previous papers.

The rest of this paper is organized as follows. § 2 formulates the mixed-effects models for the study and describes the proposed semiparametric method. We present the asymptotic properties of the semiparametric estimators and the conditions to reach the oracle property. § 3 gives the numerical results which include a simulation study, the application to the colon carcinogenesis, and a sensitivity analysis on the "effective" sample size. Finally, § 4 contains the concluding remarks.

2. THE MODEL, THE METHOD AND THE ASYMPTOTICS

2.1 Model specification

Let $X(\cdot)$ denote a latent covariate process. For the colon carcinogenesis study, $X_i(t)$ is the realization of the rat-level process for DNA adduct in rat i at cell position t. Hereafter, the cell position refers to the relative crypt depth of the cell within a colon crypt, thus $t \in [0, 1]$.

The following mixed-effects model describes a general relationship between the response Y and the latent covariate X,

$$Y^{\rm tr}_{ijk} = H(X^{\rm tr}_i(t_{ijk}), \beta^{\rm tr}) + Z^{\rm tr}_{ij} b^{\rm tr}_{ij} + \epsilon^{\rm tr}_{ijk}. \tag{1}$$

For better understanding of the notations, we associate them with the terminologies in the colon carcinogenesis example. That is, Y is the *bcl-2* gene expression, X is the DNA adduct latent covariate, i is the subject index of the rat, j is the subunit index of the crypt selected to measure *bcl-2*, k is the index of the cell in the selected crypt, and the sup-index "tr" is the treatment indicator for the diet and the time. The cell-level *bcl-2* gene expression $Y^{\rm tr}_{ijk}$ is linked to the rat-level covariate process $X^{\rm tr}_i(\cdot)$ through the cell position t_{ijk}. β is the unknown fixed effect parameter vector and H is the known link function. The zero mean and bounded variance random effect $b^{\rm tr}$, coupled with the rat- and the crypt-level observed covariate $Z^{\rm tr}$, lays out the hierarchical rat- and crypt-level dependency. Finally, $\gamma^{\rm tr}$ denotes the vector of variance parameters in the distribution of $b^{\rm tr}$ and the additive error $\epsilon^{\rm tr}$. For simplicity, we hereafter suppress the sup-index "tr" in the text.

The latent covariate $X_i(t)$ is completely unobservable but can be considered as the rat-level mean at cell position t. Let $W_{ij'k'}$ be the DNA adduct observed from cell k' of crypt j' in rat i. What follows is a natural model that links the observed $W_{ij'k'}$ to the rat-level latent covariate,

$$W_{ij'k'} = X_i(t_{ij'k'}) + d_{ij'}(t_{ij'k'}) + e_{ij'k'}. \tag{2}$$

In this nonparametric mixed-effects model, $d_{ij'}$ denotes the crypt-level variation and $e_{ij'k'}$ the additive error. Conditional on $X_i(t)$, we assume that measurements from different crypts are independent. We let γ^W_i denote the vector of variance parameters in model (2) for the covariate observation within rat i. Note that j' denotes the index of the crypt selected for measuring the DNA adduct, and k' is the index of the cell within that crypt. Due to the nature of this experiment, in no situation does $j' = j$ in (1) and (2). Throughout this paper, we use " $'$ " for the indices of covariate observations to distinguish them from the indices of response observations.

Since crypts are randomly selected from the same rat to measure the *bcl-2* and DNA adduct, the two groups of crypts should be biologically similar. As pointed out in Morris et al.

2001, the cells at the same positions of different crypts share the common characteristics. Therefore, it is reasonable to assume that the rat-level latent process for DNA adduct is the same for the two groups of crypts. This suggests that we can estimate the latent covariate $X_i(\cdot)$ in model (1) from the nonparametric model (2). In fact, the latent covariate process can only be assumed at rat level to link the observed surrogate covariate to the underlying covariate that corresponds to the response. This is because no covariate was observed from the same sub-unit (crypt) of response observation, consequently, no crypt-level process can be obtained.

2.2 Method description

Our semiparametric approach can be implemented in two steps. In step 1, we nonparametrically estimate the latent covariate X_i for each subject i based on model (2). In step 2, we use the nonparametrically estimated $X_i(t_{ijk})$ in the primary model (1) to estimate β and γ.

For the nonparametric estimation of $X_i(\cdot)$, we use local linear smoothing (Fan and Gijbels, 1996) with the working independence correlation (Lin and Carroll, 2000), and estimate within each rat separately. Other smoothing procedures can also be used. The choice of working independence approach is appropriate here because we intend to show that the simplest approach would still work here. For the parametric estimation of the primary model, we use the generalized estimating equation (GEE) with working covariance matrix. We introduce the semiparametric estimation in the case that H is quadratic. The approach which is designed for the generalized linear mixed-effects model and its associated properties have also been developed in the dissertation of the first author. We focus on presenting the more complex quadratic scenario here because it was the model used for analyzing the data.

Let n be the total number of rats. $\underline{Y_i} = (\underline{Y_{i,1}}^T, \ldots, \underline{Y_{i,J_i}}^T)^T$ is the vector of crypt by crypt *bcl-2* observations in rat i, with $\underline{Y_{i,j}}^T$ as the *bcl-2* observations from crypt j of rat i, for $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. $\underline{T_i} = (\underline{T_{i,1}}^T, \ldots, \underline{T_{i,J_i}}^T)^T$ denotes the vector of cell positions for observing *bcl-2*. $X_i(\underline{T_i})$ is the realization of $X_i(\cdot)$ at $\underline{T_i}$.

For the mixed-effects quadratic model, the semiparametric estimator for β is,

$$\hat{\beta} = \left\{ n^{-1} \sum_{i=1}^{n} [\underline{1}, \widetilde{X}_i(\underline{T_i}), \widetilde{X}_i^2(\underline{T_i})]^T \hat{\Sigma}_i^{-1} [\underline{1}, \widetilde{X}_i(\underline{T_i}), \widetilde{X}_i^2(\underline{T_i})] \right\}^{-1} \left\{ n^{-1} \sum_{i=1}^{n} [\underline{1}, \widetilde{X}_i(\underline{T_i}), \widetilde{X}_i^2(\underline{T_i})]^T \hat{\Sigma}_i^{-1} \underline{Y_i} \right\}, \tag{3}$$

where $\widetilde{X}_i(\underline{T_i})$ is the nonparametrically estimated $X_i(\underline{T_i})$ and $\hat{\Sigma}_i$ is the estimated covariance matrix in primary model (1).

To account for the nested experimental design in the colon carcinogenesis study for cells within a crypt and crypts within a rat, we consider a dependent covariance structure (4) as the working covariance for the primary model,

$$\Sigma_i = \sigma_r^2 \mathbf{J}_{N_i} + \sigma_c^2 \mathrm{diag}(\mathbf{J}_{K_{i,1}}, \ldots, \mathbf{J}_{K_{i,J_i}}) + \sigma_\epsilon^2 \mathbf{I}_{N_i}, \tag{4}$$

where σ_r^2 and σ_c^2 are the variance components for the rat- and crypt- level random effects, respectively, and σ_ϵ^2 is that for the random error. $\mathbf{J}$ is matrix of entry 1 and $\mathbf{I}$ is the identity matrix. All indices here refer to the *bcl-2* observation in rat i: N_i is the total number of *bcl-2* observations; J_i is the number of crypts selected for *bcl-2* observation and $K_{i,j}$ is the number of cells in crypt j.

When this assumed structure is true, it is known that the covariance parameters can be consistently estimated under the assumption of normal random effects even if the distributions of the random effects are not normal (see Verbeke and Lesaffre, 1997, for a reference). Consequently, there is no need of distributional assumptions on these random effects. For estimation efficiency, we also assume the same variance parameters across all treatment groups as well as at different cell position. Thus $\gamma = (\sigma_r^2, \sigma_c^2, \sigma_\epsilon^2)$, together with (4), describes the covariance in the primary model. $\hat{\Sigma}_i$ is calculated by replacing γ with $\hat{\gamma}$.

Note that there are possible choices of Σ_i to allow for correlation within crypts. For example, instead of the term $\mathbf{I}_{N_i}$ in (4), we could have a Markov-structured matrix such that the correlation between two cells in the logarithm scale is inversely proportional to the distance between the two cells. We checked and found that the correlations between two adjacent cells within a crypt are low on average, it is thus reasonable to assume the covariance structure (4) for this study.

For the estimation of the variance components represented by γ, we focus on a simple regression-based method that uses the transformed response and covariates. The exact formulae are given in the Appendix. This is not a maximum likelihood method but it shares the same consistency property and performs numerically better for small sample sizes. An "equivalent" method was proposed in Henderson (1953) and was studied extensively by Fuller and Battese (1973) for nested designs. We choose this estimator for variance components not only because the role played by the latent covariate can be explicitly reflected, but also that the estimated parameters provide a summary of the variations at different levels even if the assumed variance structure is not exact.

2.3 Asymptotic properties of the semiparametric estimators

In this subsection, to simplify the notation, we assume that all rats have the same number of crypts J' and the same number of cells K' within each crypt for observing the covariate. Due to the structure of the colon, the number of

cells within each crypt is limited, while the number of crypts within a colon is nearly unbounded. This is because, compared with the dimension of a crypt, the colon is of nearly infinite length. Therefor, in this subsection, we focus on the asymptotic scenario that J' goes to ∞ and K' is bounded. We derive the asymptotic properties of the semiparametric estimator and identify the conditions under which the oracle property can be reached. Later in § 3.3, a bootstrap-based graphical tool will be used to evaluate the adequacy of crypt size J'.

For simplicity, we denote $X_i(\underline{T_i})$ as $\underline{X_i}$ in the following. Recall that $\gamma = (\sigma_r^2, \sigma_c^2, \sigma_\epsilon^2)$ is the vector of the variance components of the primary model (1), and let $\gamma^W = (\gamma_i^W, i = 1, \ldots, n)$ be the variance parameters in the nonparametric model (2) for the subject—level latent covariate processes.

For the mixed-effects quadratic model, we obtain the following properties:

Proposition 1. *As n and $J' \to \infty$, $h \to 0$ and $J'K'h \to \infty$, $\sqrt{n}(\hat{\beta} - \beta - B_\beta) \longrightarrow N(0, V_\beta)$, with*

$$B_\beta = V_0^{-1}\{h^2 \cdot C_h + (J'K'h)^{-1} \cdot C_{J'K'h} + (J')^{-1} \cdot C_{J'}\},$$
$$V_\beta = V_0^{-1} + V_0^{-1}\{h^2 \cdot D_h + (J'K'h)^{-1} \cdot D_{J'K'h} + (J')^{-1} \cdot D_{J'}\}V_0^{-1}.$$

where

$$V_0 \equiv V_0(X, \gamma) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} [\underline{1}, \underline{X_i}, \underline{X_i}^2]^T \Sigma_i^{-1} [\underline{1}, \underline{X_i}, \underline{X_i}^2],$$

C_h, $C_{J'K'h}$, $C_{J'}$, D_h, $D_{J'K'h}$, $D_{J'}$ are functions of $(X, \gamma, \gamma^W, \beta)$ and of order $O(1)$.

When $nh^4 \to 0$, $n(J'K'h)^{-2} \to 0$, and $n(J')^{-2} \to 0$, $\hat{\beta}$ is $\sqrt{n}$-consistent. The exact expressions of C_h, $C_{J'K'h}$, $C_{J'}$, D_h, $D_{J'K'h}$, and $D_{J'}$, as well as a sketch of proof, are in the Appendix.

Remarks.

1. The variance of $\hat{\beta}^*$ has two parts. $n^{-1}V_0^{-1}$ is the asymptotic variance from the regular GEE if X were observed. The second term of V_β represents the extra variation due to the nonparametric estimation of the latent covariate. As $h \to 0$, $J' \to \infty$, and $J'K'h \to \infty$, this term diminishes and the asymptotic variance of $\hat{\beta}^*$ is $n^{-1}V_0^{-1}$. That is, the estimator achieves the oracle property.
2. Both the bias B_β and the second term of the asymptotic variance V_β go to 0 as $(J'K'h)^{-1}$ and h go to 0, and J' goes to ∞. So, instead of requiring $K'h \to \infty$, the bandwidth selection for this semiparametric estimation is determined by $J'K'h \to \infty$. That is, we do not require that the observations (over cells) within each subunit (crypt) for observing the covariate to be dense enough, but only that the observations pooled over the subunits within each subject (rat) are dense. This is because the latent covariate process is assumed at subject instead of the sub-unit level. Therefore, in addition to the facts that the same positional feature is shared across the sub-units of a subject, latent covariate process assumed at subject level also allows feasible bandwidth selection for attaining the oracle property.
   In the colon carcinogenesis example, DNA adduct was measured within each selected crypt over all the cells, with the number of cells ranging from 14 to 56 per crypt. With about 20 crypts selected to observe the DNA adduct, the total number of cells within each rat ranged from several hundreds to one thousand. Thus, it is $J'K'h$ instead of $K'h$ that could be sufficiently large, which enables reasonable bandwidth selection.
   Due to the colon structure, the number of cells (K') within a crypt is limited while the number of crypts (J') within a colon is nearly infinite. Therefore, the only within-subject sample size that could be large enough is that of the crypt. It also indicates that, for the application of this semiparametric method, the major assumption to check is that the sub-unit number within each subject is sufficiently large. We describe a graphical diagnostic tool to check this assumption in § 3.3.
3. In Proposition 1, the bias and variance of the semiparametric estimators are determined by the variance parameters γ and γ^W from both models, the underlying true covariate, and the parameter β.
   We assumed that the variance components γ of the primary model are the same for all treatment groups and irrelevant to the cell positions. This assumption can be evaluated and further relaxed through a known parametric variance structure, as stated in § 2.2. In the nonparametric mixed-effects model for the latent covariate, we allow each rat to have its own variance parameters; in addition, the variance parameters can be functions of the cell position. Though in the proof, for presentation simplicity, we consider the case that $\sigma_{d,i}^2(\cdot) \equiv \sigma_{d,i}^2$, the consistency remains true for general scenarios, provided that either each $\sigma_{d,i}^2(\cdot)$ or their sum over the n rats divided by n are bounded.

For the estimation of the variance components in (4), we obtain the following consistency.

Proposition 2. *For the latent covariate X_i satisfying the conditions in Lemma 1, estimators of variance components σ_r^2, σ_c^2 and σ_ϵ^2 with the nonparametric estimate of X_i are consistent as $h \to 0$, $J' \to \infty$ and $J'K'h \to \infty$.*

3. NUMERICAL STUDY

We study the numerical performance of the semiparametric estimation by a simulation study in § 3.1, then present the application to the colon carcinogenesis example in § 3.2. Though the results developed in § 2.3 regard the asymptotic

performance in the case of $J' \to \infty$, we show graphically in § 3.3 that the moderate crypt numbers observed in the colon carcinogenesis example seem to be sufficiently large to achieve the oracle property; that is, the estimates and their variances remain roughly the same even when we artificially increase the crypt number.

3.1 Simulation results

Here, we investigate the finite sample performance of semiparametric estimators for a quadratic mixed-effects model with a latent covariate.

We simulate the data to mimic that of the colon carcinogenesis study. Thirty ($n = 30$) subjects are considered. For each subject, we generate $K = 15$ observations of response (Y) within each of the $J = 20$ crypts. Also, in the same subject, we generate the observed covariate (W) in $J' = 20$ crypts at $K' = 40$ positions. The positions for observing W are evenly spaced, and those for observing Y follow a uniform distribution, both in $[0, 1]$.

The underlying covariate process is $X_i(t) = 5 - 5\sin(3t \cdot r_{i1}) + r_{i2}$, with $r_{i1} \sim \text{unif}[0.9, 1.1]$ and $r_{i2} \sim N(0, 1)$. We generate the observed covariate W_i by model (2) with $d_{ij'}(\cdot) \equiv d_{ij'}$; $d_{ij'}$ and $e_{ij'k'}$ are independent with mean 0 and variance σ_d^2 and σ_e^2, respectively. In this simulation, we let $\sigma_d = 0.3$ and $\sigma_e = 0.7$. We generate the observed response Y by a quadratic mixed-effects model with $\beta_0 = 1$, $\beta_1 = -2$, $\beta_2 = 1$, and the covariance structure (4) with variance components $\sigma_r = 1$, $\sigma_c = 1$, and $\sigma_\epsilon = 3$. The data generation and the semiparametric estimation are repeated for 1000 times.

We estimate the above mixed-effects model with latent covariate by three methods: (1) GEE with the true covariate values (True); (2) the semiparametric method (Semip): though the optimal bandwidth for nonparametric smoothing can be obtained through the "leave-one-subject-out" cross validation, we present the estimates over three different bandwidths around the optimal to show the influence of bandwidth; and (3) the last observation carry-forward method (LOCF), which sets the covariate value at the position of response, say t_{ijk}, to be the observed covariate at a position immediately in front of the target position; that is $W_{ij'_0k'_0}$ with $t_{ijk} - t_{ij'_0k'_0} = \min_{(j',k')}\{t_{ijk} - t_{ij'k'} : t_{ij'k'} < t_{ijk}\}$. Since NN estimates are similar to LOCF, only LOCF estimates are presented. Simulation results are in Table 2.

In Table 2, we observe that the LOCF estimates are biased toward zero, due to the attenuation effect (Carroll, Ruppert, and Stefanski, 1995). For the semiparametric estimators, the performance depends on the bandwidth h. At $h = 0.06$, the semiparametric estimates show negligible biases, and the coverage probabilities of the 95% Wald confidence intervals are very close to the nominal. In addition, the estimated standard errors based on the asymptotic normality in Proposition 1 are close to the Monte Carlo standard deviations. Though the semiparametric estimates have larger variation compared to the estimates using the true covariate values, which is unattainable in practice, this extra variation originates from the nonparametric estimation of the latent covariate and would diminish as the within-subject sample size J' increases. Overall, the simulation results indicate how close the proposed semiparametric estimators could approach the oracle property.

Table 2. Simulation Results Based on 1000 Repetitions: Primary Mixed-Effects Quadratic Model with $\beta_0 = 1$, $\beta_1 = -2$, $\beta_2 = 1$, and the Latent Covariate Process $X_i(t) = 5 - 5\sin(3t \cdot r_{i1}) + r_{i2}$ with $r_{i1} \sim \text{unif}[0.9, 1.1]$ and $r_{i2} \sim N(0, 1)$; RB (%): Relative Bias, SD: Monte Carlo Standard Deviation, SE: Average of the Estimated Standard Error from the Asymptotic Distribution, CP: Monte Carlo Coverage Probability of the 95% Wald Confidence Interval

Method		β_0	β_1	β_2
True	RB	1.024	0.056	0.020
	SD	0.192	0.047	0.009
	SE	0.194	0.046	0.009
	CP	94.8	94.7	94.3
Semip				
$h = 0.03$	RB	−0.521	−2.421	−1.406
	SD	0.199	0.073	0.015
	SE	0.193	0.065	0.013
	CP	95.7	91.1	92.8
$h = 0.06$	RB	2.071	−0.208	−0.238
	SD	0.199	0.073	0.015
	SE	0.195	0.065	0.013
	CP	95.5	94.4	95.1
$h = 0.09$	RB	5.217	1.920	0.483
	SD	0.200	0.074	0.015
	SE	0.196	0.066	0.013
	CP	95.0	93.6	93.7
LOCF	RB	−20.948	60.280	−37.085
	SD	0.241	0.121	0.028

For the estimation of the variance components in primary model (1), by semiparametric approach in (A.3) to (A.5), the estimates at $h = 0.06$ are $\tilde{\sigma}_r = 0.999$, $\tilde{\sigma}_c = 1.022$, and $\tilde{\sigma}_\epsilon = 3.003$, with the corresponding Monte Carlo standard deviation being 0.064, 0.03, and 0.019, respectively.

3.2 Analysis of colon carcinogenesis data

Here, we summarize the analysis of the colon carcinogenesis data. The goal of the study was to investigate whether *bcl-2* gene expression increases with the DNA adduct level, and whether the trend varies with diet. Recall that the response, *bcl-2*, and the covariate, the DNA adduct, were not observed from the same crypts within a rat. We assume the rat-level latent covariate process for the DNA adduct level as $X_i(\cdot)$.

Several features of this study should be noted. First, as discussed earlier, the relative crypt depth of a cell represents its physiologic function. If we divide the crypt into three sections—the bottom 1/3, the middle 1/3, and the top 1/3 section—these three sections roughly contain the

Table 3. Estimates of the Linear Mixed-Effects Model of bcl-2 Versus DNA Adduct: SE Is the Standard Error and the p-Value Is for the Comparison Between the Two Diets at Each Time Point

Time	Diet	Semip estimates		Comparison p-values		LOCF
		intercept (SE)	slope (SE)	intercept	slope	slope
0	fish	33.54 (2.79)	2.32 (0.53)	0.38	< 0.01	0.05
	corn	37.08 (2.94)	0.83 (0.43)			−0.08
3	fish	25.13 (2.79)	−0.79 (0.28)	0.04	< 0.01	−0.09
	corn	33.08 (2.80)	0.35 (0.28)			0.05
6	fish	25.57 (2.81)	0.18 (0.25)	0.43	< 0.01	0.08
	corn	28.51 (2.81)	2.25 (0.37)			0.12
9	fish	19.48 (2.88)	−1.28 (0.27)	0.54	< 0.01	−0.02
	corn	22.38 (2.89)	0.92 (0.36)			−0.02
12	fish	24.99 (2.72)	−0.52 (0.30)	0.72	< 0.01	0.07
	corn	26.42 (3.04)	0.42 (0.27)			0.09

stem cells, the proliferating cells and the differentiated cells, respectively. We accordingly carry out the analysis with respect to each section separately. Secondly, in the analysis of each section, we use the "centered" DNA adduct level around the section mean of each rat. The reason for centering is as follows: the rat to rat variation is fairly large in DNA adduct measurements, but the range of *bcl-2* is about the same for almost all rats. If we perform regression analysis rat by rat, we can see that the regression pattern is roughly shared by the rats within the same treatment group. Through "centering" in the DNA adduct, we can easily summarize the within-rat *bcl-2* versus DNA adduct relationship over all rats in the same treatment group. One the other hand, the regression that uses uncentered DNA adduct essentially models the trend between the rat-level averages of *bcl-2* and the rat-level averages of the DNA adduct, which is not the interest of this study.

Analysis is performed in all the three sections of the crypt. The results in Hong et al. (2000) indicated that the top section is where the proportions of apoptosis differ between fish oil enhanced and corn oil enhanced diets in the later stages of carcinogenesis. Therefore, we only report the results for the top section here. The results for the other two sections are either non-significant or similar to the findings for the top section.

For the semiparametric method, bandwidth selection is carried out by the "leave-one-subject-out" cross validation (Rice and Silverman, 1991). This method has been successfully applied by Hu, Wang, and Carroll (2004) and several other authors in semiparametric modeling of correlated data. For this study, the selected bandwidth is about $h = 0.05$. The values of the generalized cross validation function change little over bandwidths around 0.05, and the semiparametric estimates are very close with bandwidths in that neighborhood.

We report the results from linear mixed-effects model. Though a quadratic mixed-effects regression was conducted, we discovered that for all but one treatment groups, the quadratic trend was insignificant (with p values > 0.2). More importantly, a linear model allows easier interpretation of the diet effect on the *bcl-2* versus DNA adduct relationship. Therefore, we use the linear mixed-effects model as the primary model, and the properties in § 2.3 also apply here.

In Table 3, we report the estimated intercepts and slopes and their standard errors, as well as the p-values for the contrast between the two diets at each of the five time points. Both the standard errors and the p-values are obtained by the method of bootstrap, where the observed covariate is bootstrapped from model (2) with the underlying latent covariate as the rat-level smoothing estimate, and the response is bootstrapped from the primary model (1) with the semiparametrically estimated parameters. For comparison, we report the estimated slopes from the LOCF method. We can see that these estimated slopes shrink towards zero, as is the case in the simulation study.

From Table 3, we can see that, in the *bcl-2* versus DNA adduct relationship, the fish oil fed rats have significantly smaller slopes than the corn oil fed rats during the initial stage of colon carcinogenesis except at time 0. More specifically, as cell DNA damage increases, the *bcl-2* gene expression level always increases in the corn oil fed rats, while it either decreases as at time 3, 9, and 12 or remains relatively stable as at 6 in the fish oil fed rats.

As we know, the main function of *bcl-2* is to suppress apoptosis activity. It can consequently prevent premature cell death when the DNA damage is within a normal range, but leads to less active self-termination of the cancer-prone cells where the DNA damage is genetically irreparable. Our findings in Table 3 suggest that during the first 12 hours post carcinogen exposure, the fish oil diet, compared with the corn oil diet, suppresses the increment in the gene expression level of *bcl-2* when the DNA damage increases. Therefore, the fish oil supplemented diet is more advantageous in promoting apoptosis and potentially reducing the risk of colon cancer. On the other hand, since there is no cell DNA damage caused by the carcinogen at time 0, the positive slope of the fish oil diet at this time suggests that the fish oil diet is also good at preventing premature cell death in case of no abnormal cell damage.

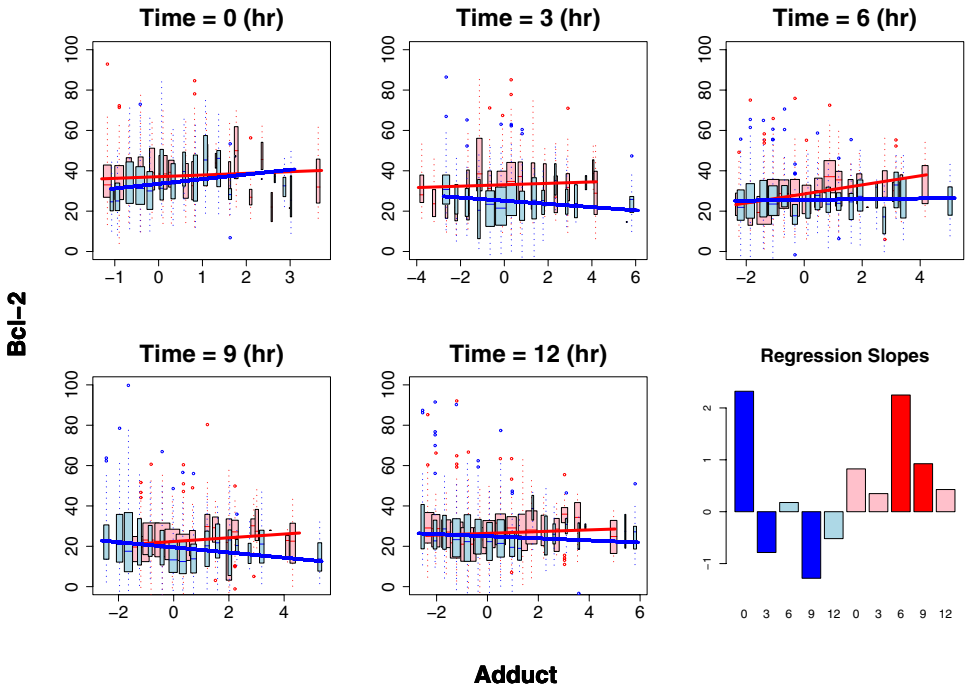

Figure 1. Fitted Regression Lines for bcl-2 vs. the Centered Smooth Estimates of DNA Adduct. Observations Are from the Top $1/3$ *Section at 0, 3, 6, 9, 12 Hours Post Carcinogen Exposure; Box Plots Are Produced for 25 Equal-Distance Regions for Corn Oil (Red) and Fish Oil (Blue) Observations, Respectively. The Plot in the Right-Lower Corner Provides Regression Slopes for All Groups; a Color of Dark Red (Blue) Indicates that the Slope Significantly Differs from 0 (at Significance Level* $\alpha = 0.05$*).*

The variance component estimates of the primary linear mixed-effects model are the rat-level variation $\hat{\sigma}_r = 4.89$, the crypt-level variation $\hat{\sigma}_c = 1.80$, and the cell-level error $\hat{\sigma}_\epsilon = 12.48$. It shows that the crypt-level variation is relatively small compared with the other two sources of variations.

Figure 1 illustrates the fitted lines from semiparametric regression at bandwidth $h = 0.05$ in the top section; the Y axis indicates the observed *bcl-2* gene expression and the X axis indicates the centered nonparametrically estimated cell DNA adduct level. Because the data points within each time group are extremely dense, we present the data in the following way. That is, for each time group, we divide the range of the centered DNA adduct values into 25 segments of equal length. Then for the centered DNA adduct values within each segment, we produce the box plot of the corresponding *bcl-2* observations.

3.3 A graphical assumption-checking tool

The key point of this work is that for a hierarchical mixed-effects model with subject-level process for latent covariate, a simple semiparametric estimation can replace the complicated EM computation. This is often the case when the number of total observations per subject is sufficiently large. Coupling the proposed semiparametric estimation with a graphical assumption-checking tool can be very useful in practice.

In Proposition 1, the asymptotics for the semiparametric estimators are built upon the condition that the crypt number for covariate observation within each rat is big, i.e., $J' \to \infty$. That is, we expect both the estimates and the corresponding variation to stabilize when J' is sufficiently large. Next, we investigate how the crypt number J' affects the semiparametric estimates, and roughly check whether the crypt number is large enough in the colon carcinogenesis example. We do this by the following bootstrap procedure, where the first two steps are to estimate the crypt effects and the random errors in model (2).

1. Within each subject, subtract the covariate values observed at t by its nonparametric estimate $\widetilde{X}_i(t)$ and obtain the best linear unbiased predictor (BLUP) of the crypt-level random effects $d_{ij'}(\cdot)$. Here, we focus on the scenario that the crypt effect does not change with the cell position; that is, $d_{ij'}(\cdot) \equiv d_{ij'}$. We then construct the kernel estimate of the crypt effect density f_d.
2. Denote the corresponding residual process at crypt j' by $r_{ij'}(\cdot)$.
3. Let J^* denote the crypt size of consideration, which is the number of crypts within each rat for covariate observation. Sample independent crypt-level random effects $d_{i\ell}^{<b>}$, $\ell = 1, \ldots, J^*$, from the estimated crypt effect density f_d. This can be achieved by letting $d_{i\ell}^{<b>} = d_{i\ell}^{<b>*} + h_d U_K^{<b>}$, where $d_{i\ell}^{<b>*}$ is sampled with replacement from the original set of BLUP's of the

crypt-level random effects; h_d is the bandwidth; $K(\cdot)$ is the kernel density function in the estimation of f_d, and $U_K^{<b>}$ is a randomly generated number from $K(\cdot)$.

4. Create bootstrapped surrogate covariates by letting $W_{i\ell}^{<b>}(\cdot) = \widetilde{X}_i(\cdot) + d_{i\ell}^{<b>} + r_{i\ell}^{<b>*}(\cdot)$, where $r_{i\ell}^{<b>*}(\cdot)$ is sampled with replacement from the original residual processes of $\{r_{ij'}(\cdot)\}$.
   Though, in step 1, we did not take the cell position into account in the estimation of the crypt effect, the effect of cell position is absorbed into the residuals and contributes to the generation of the bootstrapped $W^{<b>}$ at this step.
5. Create parametrically the corresponding bootstrapped responses $Y_i^{<b>}$ from the primary model with the estimated parameter $\hat{\beta}$ and $\widetilde{X}_i(\cdot)$.
6. Obtain semiparametric estimate $\hat{\beta}^{<b>}$ from $Y_i^{<b>}$ and $\{W_{i\ell}^{<b>}, \ell = 1, \ldots, J^*\}$, $i = 1, \ldots, n$.

The above procedure (step 3 to step 6) is repeated for $b = 1, \ldots, B$.

Based on the semiparametric estimates in Table 3, we obtain a sequence of the bootstrap estimates $\{\hat{\beta}^{<b>}, b = 1, \ldots, B\}$ for each desired crypt size J^*, with J^* ranging from 3 to 50. We keep all other setups the same as the original data and let $B = 1000$. Figure 2 presents the bootstrap estimates for the intercepts and the slopes under both diets at 9 hours post carcinogen exposure. For all plots, the X-axis indicates the crypt sizes J^*. In the top two panels, the Y-axis corresponds to the estimated intercepts, and in the bottom two are the estimated slopes. The fish oil diet plots are on the left, and the corn oil diet plots are on the right.

In Figure 2, it appears that the estimated fish oil diet slope is the most sensitive to the crypt size among the four estimates, whose bootstrap estimated standard errors decrease about 8.2% from $J^* = 5$ to $J^* = 50$. At crypt size $J^* = 24$, which is the "typical" crypt number per rat for DNA adduct observation in this example, the corresponding estimates and standard errors are within 1.01% of those obtained at $J^* = 50$, respectively, and are apparently close to the asymptotic limits. This implies that, in this example, the semiparametric estimates are close to the asymptotic results. While the bias and variation of the semiparametric estimates could be further reduced by having a larger number of crypts, the improvement would be very limited.

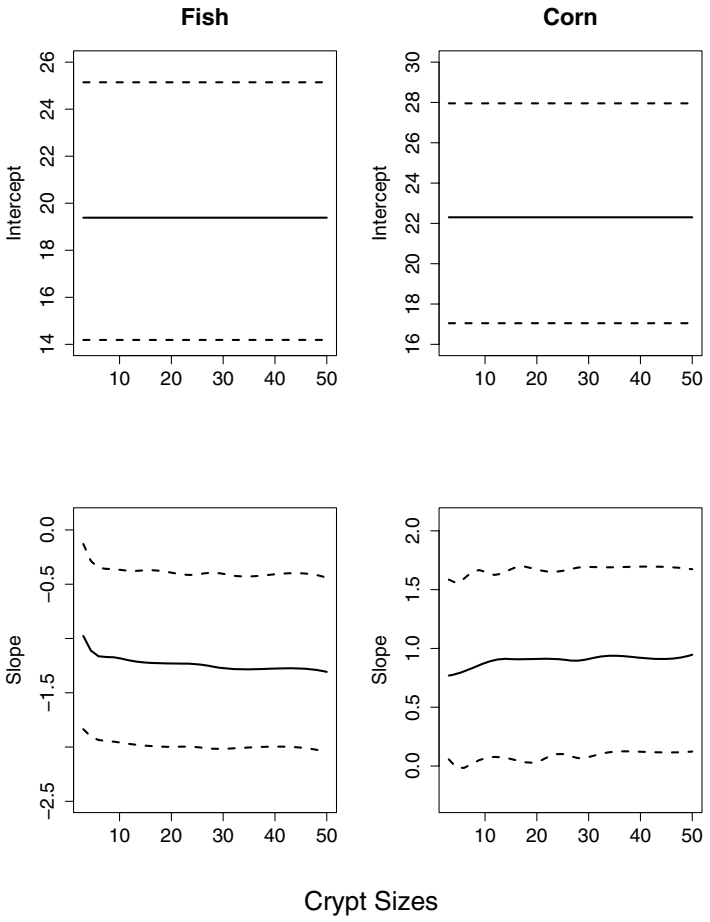

Figure 2. Diagnostic Plots of the Semiparametrically Estimated Parameters Versus the Crypt Sizes at 9 Hours Post Carcinogen Exposure. Left Panel Plots: Fish Oil; Right Panel Plots: Corn Oil; Top Panel Plots: Intercepts; Bottom Panel Plots: Slopes. The Center, Top and Bottom Curves in Each Plot Correspond to the Bootstrap Averages and the 95% Confidence Limits, Respectively.

4. CONCLUDING REMARKS

We propose a semiparametric approach for estimating a hierarchical mixed-effects model with an unobservable latent functional covariate. Compare to a parametric approach, like the maximum likelihood, semiparametric approach is computationally easier to implement for applications with hierarchical structure; meanwhile, it avoids the problem of model misspecification for the latent covariate. When the "effective" sample size is large enough, the semiparametric estimator attains the oracle property, which is the best a maximum likelihood estimator can do.

In the colon carcinogenesis study, the interest is on the relationship between *bcl-2* gene expression level and DNA adduct, while the latent covariate functional—the positional feature of DNA adduct—is not the interest. This research objective makes the semiparametric approach, parametric modeling for the primary relationship and nonparametric modeling for the positional feature, appropriate. It obviates model specification for the positional feature of DNA adduct, yet leads to consistent estimation of the primary model.

Due to the long-standing belief that within a rat, cells at the same position of different crypts share the same biological characteristics, we propose the estimation of the positional feature of DNA adduct at rat-level. This is done by

nonparametric estimation of the positional feature over all the adduct observation crypts within each rat. The advantage of this nonparametric estimation scheme is two-fold. First, it makes possible to relate the response *bcl-2* to the covariate DNA adduct, which were observed over two different groups of colon crypts in each rat. Second, the basic observation units—the cells were of fixed number within each crypt, while the subunits—the crypts were randomly selected from the rat and technically of infinite number. If we lump the crypts within each rat to estimate the rat-level DNA adduct positional feature, the total number of observation cells can be sufficiently large and the relative cell positions are densely distributed in $[0,1]$, which allows for a reasonable bandwidth selection in the nonparametric estimation. Based on this pooling-over-subunit nonparametric estimation scheme, the consistency of the estimators for both the primary parameters and the latent covariate functional requires only the number of the crypts to be sufficiently large. We also propose a bootstrap based diagnostic tool to check on the sufficiency of this "effective" sample size—the number of crypts within each rat.

Though the method developed here is motivated by the colon carcinogenesis study, it has potentially wider applications. In biological studies, it is common that two measurements of interest can not be taken from the same subunit of a subject due to the constructive nature of the experiment. It is even more common that true covariates are not directly observable but can only be postulated as coefficients or functions of another regression model (see Li, Zhang, and Davidian, 2004, for a parametric example). However, it is not always feasible to identify a parametric model for the covariate, therefore, the semiparametric approach is of more flexibility.

ACKNOWLEDGMENTS

This research was supported by a grant, CA74552, from the US National Cancer Institute. It was also partially supported by the Texas A&M Foundation LINK program.

The authors thank Drs. Hong and Lupton for kindly providing the dataset.

APPENDIX

1 Lemma and definitions for Proposition 1

We assume a common marginal distribution of the cell positions for the covariate observation cells, i.e., density $f_T(t)$ for cell position t. Similar to $\underline{T}_i$, $\underline{T}_i'$ denotes the vector of cell positions for covariate observation. Based on the mixed-effects nonparametric model (2), the local linear smoothing estimator $X_i(\cdot)$ at t has the following expression, with the subscript i suppressed.

Lemma 1. *For $X(\cdot) \in C^2(0,1)$ and a kernel density function $\mathbf{K}$ satisfying $\int s\mathbf{K}(s)ds = 0$, $\int s^2\mathbf{K}(s)ds = 1$,*

$$\widetilde{X}(t) = X(t) + W_2^{-1}(t)\frac{1}{J'}\sum_{j'=1}^{J'}\sum_{k'=1}^{K'}\mathbf{K}_h(T_{j'k'} - t)\eta_{j'k'} + \zeta(t)h^2/2 + o_p\{h^2 + (J'K'h + J')^{-1/2}\} \tag{A.1}$$

where $\mathbf{K}_h(v) = h^{-1}\mathbf{K}(v/h)$ and h is the smoothing bandwidth. $W_2(t) = K'f_T(t)$ and $\zeta(t) = X^{(2)}(t) + f_T(t)^{-1}X(t) \times f_T^{(2)}(t)$, where $X^{(2)}(\cdot)$ and $f_T^{(2)}(\cdot)$ denote the second derivatives of $X(\cdot)$ and $f_T(\cdot)$, respectively, and $\eta_{j'k'} = d_{j'} + e_{j'k'}$.

The above expression (A.1) can be derived similarly as done in Lin and Carroll (2000).

For the nonparametric estimate $\widetilde{X}_i(\underline{T}_i)$, it is the realization of (A.1) at the response observation positions $\underline{T}_i$. Thus the second term of (A.1) at $t = \underline{T}_i$ is the random error $\mathcal{W}_i(\underline{T}_i)$, and the third term is the bias $h^2/2\zeta(\underline{T}_i)$.

The covariance of $\widetilde{X}_i(\underline{T}_i)$ is

$$\mathrm{cov}\{\widetilde{X}_i(\underline{T}_i)\} = \mathrm{cov}\{\widetilde{\mathcal{W}}_i(\underline{T}_i)\} = (J'K'h)^{-1}\Sigma_i^{\mathcal{W}_1} + (J')^{-1}\Sigma_i^{\mathcal{W}_2}$$

In $\Sigma_i^{\mathcal{W}_1}$, the l-th diagonal entry corresponding to position t is $\frac{\gamma_K(0)(\sigma_{d,i}^2+\sigma_{e,i}^2)}{f_T(t)}$, and the (l,m)-th off-diagonal entry corresponding to paired position (t_l, t_m) is $\frac{(\sigma_{d,i}^2+\sigma_{e,i}^2)K(t_l-t_m)}{f_T(t_l)f_T(t_m)}$; In $\Sigma_i^{\mathcal{W}_2}$, the diagonal entry is $\frac{\sigma_{d,i}^2 f_{TT}(t,t)}{f_T^2(t)}$ and the off-diagonal entry is $\frac{\sigma_{d,i}^2 f_{TT}(t_l,t_m)}{f_T(t_l)f_T(t_m)}$. In the following, denote $\Upsilon_i^{\mathcal{W}_1}$ as the vector of diagonal entries of $\Sigma_i^{\mathcal{W}_1}$ and $\Upsilon_i^{\mathcal{W}_2}$ as the vector of diagonal entries of $\Sigma_i^{\mathcal{W}_2}$.

With the notations above, in Proposition 1, $C_h = C_1\beta$, $C_{J'K'h} = C_{2,1}\beta$, $C_{J'} = C_{2,2}\beta$, $D_h = C_1 + C_1^T$, $D_{J'K'h} = C_{2,1} + C_{2,1}^T + C_{3,1}$, $D_{J'} = C_{2,2} + C_{2,2}^T + C_{3,2}$, where

$$C_1(X,\beta,\gamma) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}[\underline{0}, \zeta(\underline{T}_i)/2, \underline{X}_i * \zeta(\underline{T}_i)]^T \times \Sigma_i^{-1}[\underline{1}, \underline{X}_i, \underline{X}_i{}^2],$$

$$C_{2,s}(X,\beta,\gamma,\gamma^W) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}[\underline{0}, \underline{0}, \underline{\Upsilon}_i^{\mathcal{W}_s}]^T\Sigma_i^{-1}[\underline{1}, \underline{X}_i, \underline{X}_i{}^2],$$

$$C_{3,s}(X,\beta,\gamma,\gamma^W) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\begin{bmatrix} 0 & 0 & 0 \\ 0 & \underline{p}_i{}^T\Sigma_i^{\mathcal{W}_s}\underline{p}_i & 2\underline{p}_i{}^T\Sigma_i^{\mathcal{W}_s}\mathrm{diag}(\underline{X}_i)\underline{p}_i \\ 0 & 2\underline{p}_i{}^T\Sigma_i^{\mathcal{W}_s}\mathrm{diag}(\underline{X}_i)\underline{p}_i & 4\underline{p}_i{}^T\mathrm{diag}(\underline{X}_i)\Sigma_i^{\mathcal{W}_s}\mathrm{diag}(\underline{X}_i)\underline{p}_i \end{bmatrix} + \lim_{n\to\infty} n^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathrm{tr}(\Sigma_i^{-1}\Sigma_i^{\mathcal{W}_s}) & 2\mathrm{tr}\{\Sigma_i^{-1}\mathrm{diag}(\underline{X}_i)\Sigma_i^{\mathcal{W}_s}\} \\ 0 & 2\mathrm{tr}\{\Sigma_i^{-1}\mathrm{diag}(\underline{X}_i)\Sigma_i^{\mathcal{W}_s}\} & 4\mathrm{tr}\{\mathrm{diag}(\underline{X}_i)^T\Sigma_i^{-1}\mathrm{diag}(\underline{X}_i)\Sigma_i^{\mathcal{W}_s}\} \end{bmatrix},$$

for $s = 1, 2$, $\underline{p}_i = \Sigma_i^{-1}[\underline{1}, \underline{X}_i, \underline{X}_i{}^2]\beta$, and $\mathrm{diag}(\underline{X}_i)$ as the $N_i \times N_i$ diagonal matrix of $\underline{X}_i$.

2 Proof of Proposition 1

Following (A.1) and suppressing subject index i,

$$\widetilde{X}^2(t) = \big[X^2(t) + 2X(t)\mathcal{W}(t) + X(t)\zeta(t)h^2 + \mathcal{W}^2(t) + o_p\{(J')^{-1/2}\}\big]\{1 + o_p(1)\}$$

Denote $\widetilde{\underline{X_i}} \equiv \widetilde{X_i}(\underline{T_i})$ as the nonparametric estimate of $\underline{X_i}$ and $\underline{\mathcal{W}_i}$ as $\mathcal{W}_i(\underline{T_i})$. Since $\mathrm{E}(\underline{\mathcal{W}_i}) = 0$ and $\mathrm{cov}(\underline{\mathcal{W}_i}) = (J'K'h)^{-1}\Sigma_i^{\mathcal{W}_1} + (J')^{-1}\Sigma_i^{\mathcal{W}_2}$, each entry in $\underline{\mathcal{W}_i}$ is $O_p\{(J'K'h)^{-1/2}\} + O_p\{(J')^{-1/2}\}$, and each entry in $\underline{\mathcal{W}_i}^2$ is $O\{(J'K'h)^{-1}\} + O\{(J')^{-1}\}$.

For the semiparametric estimator $\hat{\beta}^*$ in (3), $\hat{\beta}^* = A_1^{-1}A_2$, where

$$A_1 = n^{-1}\sum_{i=1}^{n}[\underline{1}, \widetilde{\underline{X_i}}, \widetilde{\underline{X_i}}^2]^T\Sigma_i^{-1}[\underline{1}, \widetilde{\underline{X_i}}, \widetilde{\underline{X_i}}^2],$$

$$A_2 = n^{-1}\sum_{i=1}^{n}[\underline{1}, \widetilde{\underline{X_i}}, \widetilde{\underline{X_i}}^2]^T\Sigma_i^{-1}Y_i.$$

Denote

$$A_1^i = [\underline{1}, \widetilde{\underline{X_i}}, \widetilde{\underline{X_i}}^2]^T\Sigma_i^{-1}[\underline{1}, \widetilde{\underline{X_i}}, \widetilde{\underline{X_i}}^2],$$

then, A_1^i is the matrix with entry $(A_1^i)_{r,s} = (\widetilde{\underline{X_i}}^{r-1})^T \times \Sigma_i^{-1}\widetilde{\underline{X_i}}^{s-1}$, for $r, s = 1, 2, 3$.

$$(\mathrm{A.2}) \quad (A_1^i)_{r,s} \to (\underline{X_i}^{r-1})^T\Sigma_i^{-1}\underline{X_i}^{s-1}, \quad \text{for } i = 1, \ldots, n.$$

in probability, as $J' \to \infty$, $h \to 0$ and $J'K'h \to \infty$. Consequently, $A_1 \to V_0$ in probability.

Write

$$A_2 = A_{21} + A_{22} + A_{23} + A_{24} + A_{25}, \quad \text{where}$$

$$A_{21} = n^{-1}\sum_{i=1}^{n}[\underline{1}, \underline{X_i}, \underline{X_i^2}]^T\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i^2}]\beta,$$

$$A_{22} = n^{-1}\sum_{i=1}^{n}[\underline{1}, \underline{X_i}, \underline{X_i^2}]^T\Sigma_i^{-1}\underline{\epsilon_i},$$

$$A_{23} = n^{-1}\sum_{i=1}^{n}[\underline{0}, \zeta(\underline{T_i})h^2/2, \underline{X_i} * \zeta(\underline{T_i})h^2]^T\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i^2}]\beta,$$

$$A_{24} = n^{-1}\sum_{i=1}^{n}[\underline{0}, \zeta(\underline{T_i})h^2/2, \underline{X_i} * \zeta(\underline{T_i})h^2]^T\Sigma_i^{-1}\underline{\epsilon_i},$$

$$A_{25} = n^{-1}\sum_{i=1}^{n}[\underline{0}, \underline{\mathcal{W}_i}, 2\underline{X_i} * \underline{\mathcal{W}_i} + \underline{\mathcal{W}_i}^2]^T\Sigma_i^{-1}\underline{Y_i}.$$

Note that A_{21} and A_{22} correspond to the mean and variance terms for the quadratic regression if X were observed. The bias of $\hat{\beta}^*$ originates from A_{23} and A_{25}, the extra variance is from A_{24} and A_{25} and their covariance with A_{22}. Therefore, the bias terms are as following,

$$E(A_{23}) = C_1\beta, \quad \text{and}$$

$$E(A_{25}) = n^{-1}\sum_{i=1}^{n}[\underline{0}, \underline{0}, E(\underline{\mathcal{W}_i}^2)]^T\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i}^2]\beta$$
$$= (J'K'h)^{-1} \cdot C_{2,1}\beta + (J')^{-1} \cdot C_{2,2}\beta.$$

The additional variance is as following,

$$\mathrm{cov}(A_{24}) + \mathrm{cov}(A_{25}) + \mathrm{cov}(A_{22}, A_{24}) + \mathrm{cov}^T(A_{22}, A_{24}) + \mathrm{cov}(A_{22} + A_{24}, A_{25}) + \mathrm{cov}^T(A_{22} + A_{24}, A_{25})$$

where,

$$\mathrm{cov}(A_{25}) = n^{-2}\sum_{i=1}^{n}\mathrm{cov}([\underline{0}, \underline{\mathcal{W}_i}, 2\underline{X_i} * \underline{\mathcal{W}_i} + \underline{\mathcal{W}_i}^2]^T \times \Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i}^2]\beta)$$
$$+ n^{-2}\sum_{i=1}^{n}E([\underline{0}, \underline{\mathcal{W}_i}, 2\underline{X_i} * \underline{\mathcal{W}_i} + \underline{\mathcal{W}_i}^2]^T \times \Sigma_i^{-1}[\underline{0}, \underline{\mathcal{W}_i}, 2\underline{X_i} * \underline{\mathcal{W}_i} + \underline{\mathcal{W}_i}^2]),$$

with

$$\text{1st term} = n^{-2}\sum_{i=1}^{n}\Bigg\{(J'K'h)^{-1} \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & \underline{p_i}^T\Sigma_i^{\mathcal{W}_1}\underline{p_i} & 2\underline{p_i}^T\Sigma_i^{\mathcal{W}_1}\mathrm{diag}(X_i)\underline{p_i} \\ 0 & 2\underline{p_i}^T\Sigma_i^{\mathcal{W}_1}\mathrm{diag}(X_i)\underline{p_i} & 4\underline{p_i}^T\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_1}\mathrm{diag}(X_i)\underline{p_i} \end{bmatrix} + (J')^{-1} \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & \underline{p_i}^T\Sigma_i^{\mathcal{W}_2}\underline{p_i} & 2\underline{p_i}^T\Sigma_i^{\mathcal{W}_2}\mathrm{diag}(X_i)\underline{p_i} \\ 0 & 2\underline{p_i}^T\Sigma_i^{\mathcal{W}_2}\mathrm{diag}(X_i)\underline{p_i} & 4\underline{p_i}^T\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_2}\mathrm{diag}(X_i)\underline{p_i} \end{bmatrix}\Bigg\}$$
$+ \text{higher order term}$

$$\text{2nd term} = n^{-2}\sum_{i=1}^{n}\Bigg\{(J'K'h)^{-1} \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathrm{tr}(\Sigma_i^{-1}\Sigma_i^{\mathcal{W}_1}) & 2\mathrm{tr}(\Sigma_i^{-1}\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_1}) \\ 0 & 2\mathrm{tr}(\Sigma_i^{-1}\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_1}) & 4\mathrm{tr}(\mathrm{diag}(X_i)^T\Sigma_i^{-1}\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_1}) \end{bmatrix} + (J')^{-1} \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathrm{tr}(\Sigma_i^{-1}\Sigma_i^{\mathcal{W}_2}) & 2\mathrm{tr}(\Sigma_i^{-1}\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_2}) \\ 0 & 2\mathrm{tr}(\Sigma_i^{-1}\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_2}) & 4\mathrm{tr}(\mathrm{diag}(X_i)^T\Sigma_i^{-1}\mathrm{diag}(X_i)\Sigma_i^{\mathcal{W}_2}) \end{bmatrix}\Bigg\}$$
$+ \text{higher order term}$

Thus, $\mathrm{cov}(A_{25}) = n^{-1}\{(J'K'h)^{-1} \cdot C_{3,1} + (J')^{-1} \cdot C_{3,2}\}$.

$$
\begin{aligned}
&\mathrm{cov}\{(A_{22} + A_{24}), A_{25}\} \\
&\quad = n^{-2} \sum_{i=1}^{n} E\{[\underline{1}, \underline{X_i} + \zeta(\underline{T_i})h^2/2, \underline{X_i}^2 + \underline{X_i} * \zeta(\underline{T_i})h^2]^T \\
&\qquad \times \Sigma_i^{-1}[\underline{0}, \underline{\mathcal{W}_i}, 2\underline{X_i} * \underline{\mathcal{W}_i} + \underline{\mathcal{W}_i}^2]\} \\
&\quad = n^{-1}\{(J'K'h)^{-1} \cdot C_{2,1} + (J')^{-1} \cdot C_{2,2}\} \\
&\qquad + n^{-1}[O\{h/(J'K')\} + O\{h^2/J'\}].
\end{aligned}
$$

Since $\mathrm{cov}(A_{22}, A_{24}) = n^{-1}h^2 \cdot C_1$ and $\mathrm{cov}(A_{24}) = O(h^4)$, Proposition 1 follows from the above derivations.

3 Proof of Proposition 2

Fuller and Battese (1973) gave the variance components estimators for nested design. The estimator of $\gamma = (\sigma_r^2, \sigma_c^2, \sigma_\epsilon^2)$ has the following expression, should the true covariate is known:

$$\hat{\sigma_\epsilon}^2 = \hat{\tau}^T\hat{\tau}/(N_2 - N_1 - p + \lambda_{12}) \tag{A.3}$$

$$\hat{\sigma_c}^2 = \frac{\hat{u}^T\hat{u} - (N_2 - n - p + \lambda_1)\hat{\sigma_\epsilon}^2}{N_2 - \mathrm{tr}(H_b)} \tag{A.4}$$

$$\hat{\sigma_r}^2 = \frac{\hat{v}^T\hat{v} - (N_2 - p)\hat{\sigma_\epsilon}^2 - \{N_2 - \mathrm{tr}(H_{a_1})\}\hat{\sigma_c}^2}{N_2 - \mathrm{tr}(H_{a_2})} \tag{A.5}$$

Denote $\mathbf{X}$ as the design matrix with the true covariates. H_b, H_{a_1}, and H_{a_2} are the hat matrices in the estimation of variance components. Here the notations are the same as in Fuller and Battese (1973).

The semiparametric variance component estimators $\tilde{\sigma_r}^2$, $\tilde{\sigma_c}^2$, and $\tilde{\sigma_\epsilon}^2$ are of the same expression as in (A.5), (A.4), and (A.3), except that $\mathbf{X}$ is replaced by $\widetilde{\mathbf{X}}$, which is the design matrix of the nonparametrically estimated covariates. To study these semiparametric variance components estimators, we need only to focus on the effects from the nonparametric estimation of the covariates, which are contained in the following terms Q_x=$\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ and $H_x = \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T$, where Q_x appears in hat matrices H_b, H_{a1}, H_{a2}; H_x appears in the estimated sum of squared errors $\hat{\tau}^T\hat{\tau}$, $\hat{u}^T\hat{u}$, and $\hat{v}^T\hat{v}$.

It is easy to see that $n^{-1}Q_x \to n^{-1}\mathbf{X}^T\mathbf{X}$ in probability as $J' \to \infty$, $h \to 0$ and $J'K'h \to \infty$. Similarly, $n^{-1}H_x \to n^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ in probability. Thus, the semiparametric variance components estimators $(\tilde{\sigma_r}^2, \tilde{\sigma_c}^2, \tilde{\sigma_\epsilon}^2)$ converge to $(\hat{\sigma_r}^2, \hat{\sigma_c}^2, \hat{\sigma_\epsilon}^2)$ in probability. Since $(\hat{\sigma_r}^2, \hat{\sigma_c}^2, \hat{\sigma_\epsilon}^2)$ is unbiased estimator for $(\sigma_r^2, \sigma_c^2, \sigma_\epsilon^2)$, the semiparametric variance components estimators are thus consistent.

Received 23 October 2007

REFERENCES

Carroll, R. J. (2004). Discussion of two important missing data issues. *Statistical Sinica* **14** 627–629.

Carroll, R. J. and Lin, X. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association* **95** 520–535.

Carroll, R. J., Ruppert D. and Stefanski L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society B* **53** 573–585.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39** 1–38.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.

Fuller, W. A. and Battese, G. E. (1973). Transformations of estimation of linear models with nested-error structure. *Journal of the American Statistical Association* **68** 626–632.

Heemels M. T., Dhand R. and Allen L. (2000). Apoptosis. *Nature* **407** 769–769.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9** 226–252.

Hong, M. Y., Lupton, J. R., Morris, J. S., Wang, N., Carroll, R. J., Davidson, L. A., Elder R. H. and Chapkin R. S. (2000). Dietary fish oil reduces 06-methylguanine DNA adduct levels in rat colon in part by increasing apoptosis during tumor initiation. *Cancer Epidemiology Biomarkers & Prevention* **9** 819–826.

Hu, Z., Wang, N. and Carroll, R. J. (2004). Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika* **91** 251–262.

Huang, X. and Zhu, Q. (2002). A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets. *Pattern Recognition Letters* **23** 1613–1622.

Li, E., Zhang, D. and Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics* **60** 1–7.

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.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D., Hong, M. Y. and Carroll, R. J. (2001). Parametric and nonparametric methods for understanding the relationship between carcinogen-induced DNA adduct levels in distal and proximal regions of the colon. *Journal of the American Statistical Association* **96** 816–827.

Pepe, M. S. and Fleming, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate data. *Journal of the American Statistical Association* **86** 108–113.

Pielou, E. C. (1961). Segregation and symmetry in two species populations as studied by nearest neighbor methods. *Journal of Ecology* **49** 255–269.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society B* **53** 233–243.

Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88** 447–458.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed effects models for longitudinal data. *Computational Statistics and Data Analysis* **23** 541–556.

Zonghui Hu
NIAID, National Institutes of Health
Bethesda, MD 20817, USA
E-mail address: huzo@niaid.nih.gov

Naisyin Wang
Department of Statistics
Texas A&M University
College Station, TX 77843, USA
E-mail address: nwang@stat.tamu.edu