SEMIPARAMETRIC FUNCTIONAL DATA ANALYSIS FOR LONGITUDINAL/CLUSTERED DATA: THEORY AND APPLICATION

A Dissertation

by

ZONGHUI HU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2004

Major Subject: Statistics

SEMIPARAMETRIC FUNCTIONAL DATA ANALYSIS FOR LONGITUDINAL/CLUSTERED DATA: THEORY AND APPLICATION

A Dissertation

by

ZONGHUI HU

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Naisyin Wang (Chair of Committee)	
Raymond J. Carroll (Member)	Noah D. Cohen (Member)
Michael T. Longnecker (Member)	Michael T. Longnecker (Head of Department)

December 2004

Major Subject: Statistics

ABSTRACT

Semiparametric Functional Data Analysis for Longitudinal/Clustered Data: Theory and Application. (December 2004)

Zonghui Hu, B.S., Dalian University of Technology, Dalian, P.R.China;

M.S., Dalian University of Technology, Dalian, P.R. China;

M.S., Texas A& M University, College Station, TX

Chair of Advisory Committee: Dr. Naisyin Wang

Semiparametric models play important roles in the field of biological statistics. In this dissertation, two types of semiparametic models are to be studied. One is the partially linear model, where the parametric part is a linear function. We are to investigate the two common estimation methods for the partially linear models when the data is correlated — longitudinal or clustered. The other is a semiparametric model where a latent covariate is incorporated in a mixed effects model. We will propose a semiparametric approach for estimation of this model and apply it to the study on colon carcinogenesis.

First, we study the profile-kernel and backfitting methods in partially linear models for clustered/longitudinal data. For independent data, despite the potential root-*n* inconsistency of the backfitting estimator noted by Rice (1986), the two estimators have the same asymptotic variance matrix as shown by Opsomer and Ruppert (1999). In this work, theoretical comparisons of the two estimators for multivariate responses are investigated. We show that, for correlated data, backfitting often produces a larger asymptotic variance than the profile-kernel method; that is, in addition to its bias problem, the backfitting estimator does not have the same asymptotic efficiency as the profile-kernel estimator when data is correlated. Consequently, the common practice of using the backfitting method to com-

pute profile-kernel estimates is no longer advised. We illustrate this in detail by following Zeger and Diggle (1994), Lin and Carroll (2001) with a working independence covariance structure for nonparametric estimation and a correlated covariance structure for parametric estimation. Numerical performance of the two estimators is investigated through a simulation study. Their application to an ophthalmology dataset is also described.

Next, we study a mixed effects model where the main response and covariate variables are linked through the positions where they are measured. But for technical reasons, they are not measured at the same positions. We propose a semiparametric approach for this misaligned measurements problem and derive the asymptotic properties of the semiparametric estimators under reasonable conditions. An application of the semiparametric method to a colon carcinogenesis study is provided. We find that, as compared with the corn oil supplemented diet, fish oil supplemented diet tends to inhibit the increment of *bcl-2* (oncogene) gene expression in rats when the amount of DNA damage increases, and thus promotes apoptosis.

To my parents, my husband, and my daughter.

ACKNOWLEDGMENTS

During my study toward the Ph.D degree, people have given me a lot of help. In these acknowledgements, I hope to express my gratitude to some of them who have been most important to the fulfilment of this work.

First of all, I would like to thank my advisor Naisyin Wang for her guidance, support, and encouragement. She not only guided me through the research leading to this dissertation, she also provided me opportunities for further accomplishment in this field. She has been a patient mentor, letting me develop my ideas in addition to giving her advice. I am very grateful for what she has done for me.

I am equally grateful to committee members Raymond Carroll, Noah Cohen, and Michael Longnecker. They served on my committee even with their busy schedules. For Dr. Carroll, I sincerely thank him for his professional and career advice, which was extremely valuable to me. For Dr. Longnecker, he is always the resource I go to when I need advice on applied statistical methodologies. His patience and willingness to be accessible make his class an enjoyment for his students. As for Dr. Cohen, it has been a valuable experience to work with him. I can not thank him enough for providing me not only the financial support, but more importantly the chance to work with real clinical projects.

Besides, I am also very grateful to the professors who taught me the fundamental statistics. I would like to thank Tailen Hsing for his probability class, Daren Cline and Tom Wehrly for statistical theory, James Calvin for linear model, Clifford Spiegelman for multivariate analysis, Suojin Wang for large sample theory, Jeff Hart for nonparametric estimation, and Fred Dahm for biostatistics. The knowledge I obtained from them makes it possible for me to pursue my career as a statistician.

I thank Marilyn Randall for answering my questions on graduate study, Sandra Wood

and Pat Zeringue for help on financial issues. In fact, all the staff members work hard to give us a nice working environment. I am deeply grateful to them.

In closing, I want to thank my family — my parents, my husband and my daughter. Their love and support are an everlasting spiritual power for me. My parents brought me up to be righteous and hard working. They set a good example with their own life and successful careers. I want to thank my husband, who supports our family both financially and spiritually. With his understanding and patience, I could pursue my goal with concentration. I want to thank my daughter. She brings great joy to our life. She makes me motivated to grow stronger with her. I am lucky to have the best family in the world!

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

The first topic in this work is on the partially linear models. As a special case of semi-parametric models (Ruppert, Wand, and Carroll 2003), partially linear models have been studied intensively in the literature, see Härdle et al. (2000). Compared with parametric models, partially linear models provide great flexibility in modeling the data. This advantage of partially linear models is obvious when the main interest of the study is the linear effects, and the effects from other factors are unidentifiable or simply unimportant.

$$Y = X^{\mathrm{T}}\beta + \theta(T) + \varepsilon$$

Above is a general form of the partially linear model. It contains the linear term $X^{\mathrm{T}}\beta$, where β is unknown vector of parameters. It also contains a nonparametric term $\theta(T)$ where $\theta(\cdot)$ is unknown smooth function. In this model, Y is the response, X and T are the covariates, and ε is the random error.

There are two common methods for estimating the partially linear model, namely the profile-kernel method (Carroll et al. 1997) and the backfitting method (Buja et al. 1989). Both methods involve the nonparametric estimation on function $\theta(\cdot)$ and the parametric estimation on parameter β. For independent data, Rice (1986) pointed out that at the optimal bandwidth for nonparametric estimation, the backfitting estimator is not root-n consistent, while the profile-kernel method is consistent. However, Opsomer and Ruppert (1997) showed that these two estimators actually have the same asymptotic variances. There-

---

This dissertation follows the style and format of the *Journal of the American Statistical Association*.

fore, with under smoothing (a smaller bandwidth rather than the optimal one is adopted for nonparametric estimation), estimators from both methods can be consistent. More importantly, they are of the same asymptotic efficiency. Due to this equivalence between the two methods, people frequently use backfitting as a substitute for profile-kernel even for correlated data. They apply the backfitting method for estimation of the partially linear model, and consider the estimator has the same properties of the profile-kernel estimator. In fact, the properties of the backfitting estimator are not clear up to now when the data is correlated. Therefore, it is worth investigating whether the asymptotic equivalence between the backfitting and profile-kernel is still valid in case of correlated data. Also, we work on the asymptotics of the backfitting estimator for correlated data, this part of work is to be included in chapter II.

Another topic of the work is on the colon carcinogenesis. During development of colon cancer, the first event to happen is DNA damage in cells. It takes place within the first few hours after exposure to carcinogen. The DNA damage does not necessarily lead to formation of cancer cells, due to a surveillance system in the cell cycle, see Karp (2002). When the surveillance system detects the presence of DNA damage, it triggers a response that temporarily arrests further cell cycle progress. The cell then uses the delay to repair the damage or transmit a signal to kill the cell when the DNA damage is beyond repair. In this way, the body reduces the risk of damaged cells becoming cancerous. The function of " cell suicide" is called apoptosis. Apoptosis is one of the body's main weapons against cancer by getting rid of the defective cells. Thus, any alteration that diminishes a cell's ability of apoptosis would increase the risk of cancer. In the body, there are a groups of proteins called oncogenes. The oncogene most closely linked to apoptosis is *bcl-2* gene, which encodes a membrane - bound protein that inhibits apoptosis. Consequently, over-expression of *bcl-2* gene leads to suppression of apoptosis, allowing abnormal cells to proliferate and form cancer cells. Therefore, during initial stage of cancer development,

DNA damage may cause the formation of cancer cells depending on the functioning of apoptosis. Meanwhile, over-expression of *bcl-2* adversely affects apoptosis. Since there are very few cases of apoptosis at the initiation stage of colon cancer, in this work, we will investigate the relationship between DNA damage and *bcl-2* gene expression, instead of directly on apoptosis. A clear understanding of this relationship is important to that of colon cancer development.

The objective of the colon carcinogenesis study is to investigate the relationship between *bcl-2* gene expression and DNA damage, and also the effect of diet to this relationship. A difficulty in this study is that although the two measurements, *bcl-2* gene expression and DNA damage, are measured from the same experimental units — rats, they are measured at different subsampling units. They are measured over the cells from different crypts within the colon. As a result, the response and covariate are observed at varying locations depending on the cell number in each observed crypt. Furthermore, they are observed in different crypts. Due to these two reasons, conventional regression methods are not appropriate for this colon cancinogenesis study.

We propose a semiparametric approach for this study. We will apply a mixed effects model to study the relationship of *bcl-2* gene expression versus DNA damage and the diet effect. In this mixed effects model, a latent covariate is incorporated to stand for the unobservable DNA damage at the cell positions of *bcl-2* measurement. Consequently, a semiparametric estimation procedure is introduced for the relationship and the diet effect. For comparison, we will also apply the traditional methods (last observation carry-forward and nearest neighbor) to this misaligned measurements problem. Based on the regression outcomes, we are to find out how the diet affects the development of colon cancer during the initial stage. This second topic of the dissertation is to be presented in chapter III.

CHAPTER II

ESTIMATION OF PARTIALLY LINEAR MODELS FOR LONGITUDINAL/CLUSTERED DATA *

2.1 Introduction

The partially linear model has been investigated intensively in the literature and various extensions have been proposed; see for example Härdle et al.(2000). There have been two main classes of estimation methods for this model, namely the profile-kernel and backfitting methods. For independent data, Severini and Staniswalis (1994) and Carroll et al. (1997), among others, have studied the profile-kernel approach. Buja et al. (1989), Hastie and Tibshirani (1990) and Opsomer and Ruppert (1999) have investigated the backfitting approach. For clustered data, Severini and Staniswalis (1994) and Lin and Carroll (2001) extended the profile-kernel method to accommodate multivariate responses, as did Wang, Carroll and Lin (2004) in an unpublished report, while Zeger and Diggle (1994) studied the backfitting method.

On the theoretical front, the asymptotic properties of profile-kernel estimators were provided by Severini and Staniswalis (1994), Lin and Carroll (2001) and by Wang, Carroll and Lin in their report. Their results also cover the clustered data scenario. For independent data the bias problem of backfitting estimation was first noted by Rice (1986); see also Speckman (1988), Opsomer and Ruppert (1999). Their findings indicate that undersmoothing during nonparametric estimation is required for root-*n* consistent parametric estimation for the backfitting method. Meanwhile, Opsomer and Ruppert (1999) also showed that the

---

* Hu, Z., Wang, N., and Carroll, R.J., "Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data", *Biometrika*, 2004, Vol. 91 (2), 251-262, reproduced by permission of the *Biometrika* Trustees.

two estimators share the same asymptotic variance matrix.

In contrast to profile-kernel methods, properties of backfitting for clustered data are less well understood. In this chapter, we investigate the asymptotic properties of the backfitting method for clustered data. In practice, backfitting is often used as a substitute for profile-kernel estimation, perhaps because of their variance equivalence property in the independent case, as well as its simplicity. However, it is unclear whether or not this equivalence still holds for clustered data. The main purpose of this paper is to investigate this issue.

We will make asymptotic comparisons between profile-kernel and backfitting estimation in two contexts, namely generally under a specific but widely applicable condition on the covariance matrix of the clustered data, and specifically under the scenario considered in Zeger and Diggle (1994), Lin and Carroll (2001). For the latter, we use a working independence correlation structure for the nonparametric estimation and a moment-based estimated covariance structure in parametric estimation: this estimation scheme is commonly used in practice. We will show that, besides the bias problem, for clustered data, the backfitting estimator tends to have larger variance than the profile-kernel estimator; that is, the asymptotic equivalence in variance no longer holds for the multivariate case.

The organization of this chapter is as following: we discuss the two estimation procedures, profile-kernel versus backfitting, in section 2.2 and summerize their asymptotic properties in section 2.3. We demonstrate our theoretical results with a simulation study in section 2.4, and an application in ophthalmology is given in section 2.5. Finally, concluding remarks are given in section 2.6, and proofs of the results in this chapter are provided in section 2.7.

2.2 Estimation Procedures

The partially linear model is

$$Y_{ij} = X_{ij}^{\mathrm{T}}\beta + \theta(T_{ij}) + \varepsilon_{ij}, \tag{2.1}$$

where the ith cluster, $i = 1, \cdots, n$, has m_i observations, β is a $p \times 1$ vector and $\theta(\cdot)$ is an unknown smooth function. Here, the ε_{ij} are random errors and we assume that the ε_{ij} from different clusters are independent. Without loss of generality, we let $m_i = m$ for all i. As in Lin and Carroll (2001), we assume that $E(Y_{ij}|X_i, T_i) = E(Y_{ij}|X_{ij}, T_{ij})$, where $X_i = (X_{i1}, \cdots, X_{im})^{\mathrm{T}}$, $T_i = (T_{i1}, \cdots, T_{im})^{\mathrm{T}}$ denote the covariates observed from the ith subject; see also Pepe and Couper (1997). Likewise, we assume that $E(Y_{ij}|T_i) = E(Y_{ij}|T_{ij})$ and denote it by $m_Y(T_{ij})$; $m_X(T_{ij})$ is defined equivalently.

For profile-kernel estimation, for a given β, the estimator of $\theta(\mathbf{T})$ is

$$\widehat{\theta}(\mathbf{T};\beta) = \widehat{m}_Y(\mathbf{T}) - \widehat{m}_X(\mathbf{T})\beta,$$

where $\mathbf{T} = (T_1^{\mathrm{T}}, \cdots, T_n^{\mathrm{T}})^{\mathrm{T}}$, and $\widehat{m}_Y(\mathbf{T})$, and $\widehat{m}_X(\mathbf{T})$ are nonparametric estimators of $m_Y(\mathbf{T})$ and $m_X(\mathbf{T})$, respectively. For a function with a scalar argument, for example, $\theta(\cdot)$, the notation $\theta(v)$ denotes a vector whose ith element is $\theta(v_i)$.

The parameter β is then estimated by a profile-kernel generalized estimating equation,

$$\sum_{i=1}^{n} \frac{\partial\{X_i\beta + \widehat{\theta}(T_i;\beta)\}^{\mathrm{T}}}{\partial\beta} V_i^{-1}(X_i, T_i)[Y_i - \{X_i\beta + \widehat{\theta}(T_i;\beta)\}] = 0,$$

where the V_i's are the working covariance matrices. The profile-kernel estimators of β and θ are, respectively,

$$\begin{aligned}
\widehat{\beta}_{\mathrm{P}} &= \left[\sum_{i=1}^{n}\{X_i - \widehat{m}_X(T_i)\}^{\mathrm{T}} V_i^{-1}\{X_i - \widehat{m}_X(T_i)\}\right]^{-1}\left[\sum_{i=1}^{n}\{X_i - \widehat{m}_X(T_i)\}^{\mathrm{T}} V_i^{-1}\{Y_i - \widehat{m}_Y(T_i)\}\right], \\
\widehat{\theta}(t) &= \widehat{m}_Y(t) - \widehat{m}_X(t)\widehat{\beta}_{\mathrm{P}}.
\end{aligned}$$

In matrix form, the profile-kernel estimator of β can be written as

$$\widehat{\beta}_{\mathrm{P}} = \{\mathbf{X}^{\mathrm{T}}(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^{\mathrm{T}}(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\mathbf{Y}, \tag{2.2}$$

where $\mathbf{S}$ is a smoother matrix with respect to $\mathbf{T}$ (c.f. Opsomer and Ruppert 1997), and $\mathbf{V} = \mathrm{diag}(V_1, \cdots, V_n)$ is the block diagonal matrix containing the n working covariance matrices.

For backfitting, at the current value of $\beta = \widehat{\beta}_c$, the updated estimator of θ is

$$\widehat{\theta}(\mathbf{T}; \widehat{\beta}_c) = \widehat{m}_Y(\mathbf{T}) - \widehat{m}_X(\mathbf{T})\widehat{\beta}_c,$$

and the updated value of β is obtained by a generalized least squares regression of $Y_i - \widehat{\theta}(T_i; \widehat{\beta}_c)$ on X_i with the argument β minimizing

$$\sum_{i=1}^{n} \{Y_i - \widehat{\theta}(T_i; \widehat{\beta}_c) - X_i\beta\}^{\mathrm{T}} V_i^{-1} \{Y_i - \widehat{\theta}(T_i; \widehat{\beta}_c) - X_i\beta\}.$$

At convergence, the backfitting estimators of β and θ are, respectively,

$$\begin{aligned} \widehat{\beta}_{\mathrm{BF}} &= \left[\sum_{i=1}^{n} X_i^{\mathrm{T}} V_i^{-1} \{X_i - \widehat{m}_X(T_i)\}\right]^{-1} \left[\sum_{i=1}^{n} X_i^{\mathrm{T}} V_i^{-1} \{Y_i - \widehat{m}_Y(T_i)\}\right], \\ \widehat{\theta}(t) &= \widehat{m}_Y(t) - \widehat{m}_X(t)\widehat{\beta}_{\mathrm{BF}}. \end{aligned}$$

In matrix form, the backfitting estimator of β is

$$\widehat{\beta}_{\mathrm{BF}} = \{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\mathbf{Y}. \tag{2.3}$$

For independent data where $\mathbf{V} = I$ is used, the two estimators for β are

$$\begin{aligned} \widehat{\beta}_{\mathrm{P}} &= \{\mathbf{X}^{\mathrm{T}}(I-\mathbf{S})^{\mathrm{T}}(I-\mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^{\mathrm{T}}(I-\mathbf{S})^{\mathrm{T}}(I-\mathbf{S})\mathbf{Y}, \\ \widehat{\beta}_{\mathrm{BF}} &= \{\mathbf{X}^{\mathrm{T}}(I-\mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^{\mathrm{T}}(I-\mathbf{S})\mathbf{Y}. \end{aligned} \tag{2.4}$$

2.3 Asymptotic Properties

Throughout, the number of observations for each subject, m, is regarded as fixed. The usual regularity assumptions on the kernel function are assumed, including that the second moment is assumed to equal 1. We also assume that $(Y_i, X_i, T_i), i = 1, \cdots, n$, are independent and identically distributed with $f_j(t)$ denoting the marginal density of T_{ij}. Throughout this section, we assume the regularity conditions as in Lin and Carroll (2001) and suppress the index i in the presentation.

The results concerning the comparison of the asymptotic variances of the two estimators can be constructed based on (2.2) and (2.3); that is, the results are not restricted to the case of the local linear smoother.

For independent data, as observed in expression (2.4), the profile-kernel estimator and the backfitting estimator are identical if the smoother matrix $\mathbf{S}$ is symmetric and idempotent. They are generally different otherwise. However, the two estimators have the same asymptotic variance matrix; see Opsomer and Ruppert (1999). For clustered data, the comparison of the variances of the two estimators can be simplified when V and Σ are functions only of T.

Proposition II.1. Under the assumption that both the working covariance matrix V and the true covariance matrix Σ depend only on T, the asymptotic variance of the backfitting estimator is at least as large as that of profile-kernel estimator. That is, $V_{\mathrm{BF}} - V_{\mathrm{P}}$ is positive semidefinite.

A sketch proof is given in the Appendix. Proposition 1 shows that, for clustered data, the two estimators may not share the same asymptotic variance matrix, in contrast to the independent case. This result is completely general and does not require a specific choice of working covariance matrix beyond that it does not depend on X. The result also applies to general nonparametric smoothers.

To appreciate better the differences between the two estimators, we now concentrate on the following commonly-used estimation scheme. For nonparametric estimation, we assume a working independence correlation matrix, and, for parametric estimation, we use a working covariance matrix V_i estimated by data. Wang and Wang (2001), Lin and Carroll (2001) discuss the advantage of variance reduction in using the correlation for parametric estimation versus ignoring the correlation.

The following proposition concerning the profile-kernel method is given in Lin and Carroll (2001). We quote it here to ease comparison with properties of the backfitting method given in Proposition II.3. In the next two propositions, the results are based on using a local linear smoother with working independence in nonparametric estimation. This estimation scheme is also taken for the numerical studies in the following sections.

Proposition II.2. (Lin and Carroll, 2001) Suppose that $h \propto n^{-\alpha}$, $1/5 \leq \alpha \leq 1/3$ and $n \to \infty$ and define

$$\widetilde{X} \;=\; X + \lim_{n\to\infty} \partial\widehat{\theta}(T;\beta)/\partial\beta.$$

Then $\widehat{\beta}_{\rm P}$ converges in distribution: $\sqrt{n}\{\widehat{\beta}_{\rm P} - \beta + h^2 b_{\rm P}(\beta,\theta)/2\} \longrightarrow {\rm N}(0, V_{\rm P})$, where

$$\begin{aligned} b_{\rm P}(\beta,\theta) &= E(\widetilde{X}^{\rm T}V^{-1}\widetilde{X})^{-1}E\{\widetilde{X}^{\rm T}V^{-1}\theta^{(2)}(T)\}, \\ V_{\rm P} &= E(\widetilde{X}^{\rm T}V^{-1}\widetilde{X})^{-1}E\{(Z_1 - Z_2)^{\rm T}\Sigma(Z_1 - Z_2)\}E(\widetilde{X}^{\rm T}V^{-1}\widetilde{X})^{-1}. \end{aligned}$$

Here $\widetilde{X} = \{X - m_X(T)\}$, $\Sigma = {\rm var}(Y|X,T), Z_1 = V^{-1}\widetilde{X}$, $Z_2 = (Z_2^1, \cdots, Z_2^m)^{\rm T}$, with $Z_2^j = \{\sum_{k=1}^m \sum_{l=1}^m E(\widetilde{X}^k V^{kl}|T^l = T^j)\}\, f_j(T^j)/\sum_{l=1}^m f_l(T^j)$, and V^{kl} denotes the (k,l) entry of V^{-1}.

Proposition II.3. Under the same conditions as those of Proposition II.2, the backfitting estimator $\widehat{\beta}_{\rm BF}$ converges in distribution: $\sqrt{n}\{\widehat{\beta}_{\rm BF} - \beta + h^2 b_{\rm BF}(\beta,\theta)/2\} \longrightarrow {\rm N}(0, V_{\rm BF})$, where

$$b_{\rm BF}(\beta,\theta) \;=\; E(\widetilde{X}^{\rm T}V^{-1}\widetilde{X})^{-1}E\{X^{\rm T}V^{-1}\theta^{(2)}(T)\},$$

$$V_{\mathrm{BF}} \;=\; E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1}E\{(Z_1^*-Z_2^*)^T\Sigma(Z_1^*-Z_2^*)\}E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1},$$

and $Z_1^* = V^{-1}X$, $Z_2^* = (Z_2^{*1},\cdots,Z_2^{*m})^{\mathrm{T}}$, with $Z_2^{*j} = \{\sum_{k=1}^m \sum_{l=1}^m E(X^kV^{kl}|T^l = T^j)\}\, f_j(T^j)$ $/\sum_{l=1}^m f_l(T^j)$.

A sketch proof of Proposition II.3 is provided in the Appendix.

For clustered data under the estimation scheme considered, the profile-kernel estimator is in general root-*n* inconsistent. An exception occurs when working independence is assumed throughout (Lin and Carroll 2001).

Corollary II.1. Under the assumption that the working covariance matrix V depends only on T, when h is of regular order $n^{-1/5}$, the profile-kernel estimator is root-*n* consistent, while the backfitting estimator is root-*n* inconsistent; under the assumed conditions, $E\{X^{\mathrm{T}}V^{-1}\theta^{(2)}(T)\}$ in b_{BF} remains non-zero.

Corollary II.1 is a direct consequence of (A.3) with straightforward conditional expectation calculations.

As shown in Proposition II.1, the results concerning asymptotic variance matrices of the two estimators apply to general nonparametric smoothers. For independent data, Opsomer and Ruppert (1999) point out that the two estimators have the same asymptotic variance matrix. This is also an easy consequence of Propositions II.2 and II.3. To see this, note that, for independent data, both Σ and V equal $\sigma^2 I$. In this case, $Z_1 = \sigma^{-2}\widetilde{X}$, $Z_2 = \sigma^{-2}E(\widetilde{X}|T) = 0$, $Z_1^* = \sigma^{-2}X$ and $Z_2^* = \sigma^{-2}E(X|T)$. Consequently, $Z_1 - Z_2 = Z_1^* - Z_2^* = \sigma^{-2}\widetilde{X}$ and the asymptotic variance matrices of the two estimators are $V_{\mathrm{P}} = V_{\mathrm{BF}} = \sigma^2[E\{\mathrm{cov}(X|T)\}]^{-1}$.

For clustered data, the results in the Appendix indicate that the two asymptotic variance matrices will be the same if and only if $E\{m_X(\mathbf{T})^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\,\Sigma\,(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}m_X(\mathbf{T})\}$ is zero; that is, a specific structure is required of the smoother matrix. In Lemma 1 of Wang, Carroll and Lin's report, it is shown that the nonparametric smoother of Wang (2003) pos-

sesses such a property. The above propositions and Corollary II.1 clearly indicate that, under the currently most commonly used estimation scheme, backfitting in general has a larger asymptotic variance than the profile-kernel estimator and is often more biased.

2.4 Simulation Study

We conducted a simulation study to evaluate the finite sample performance of the profile-kernel method versus the backfitting method, again in the specific context that the nonparametric estimation uses working independence. Of course, from our results, we expect the profile-kernel method to have smaller variance in general, not just for this particular choice of smoother.

For the case of clustered data, we generated 500 datasets, each comprising $n = 100$ subjects with $m = 5$ observations per subject. The covariate vectors $(T_{ij}, X_{ij}), j = 1, \cdots, m$, were independently generated from the bivariate normal distribution with mean 0, variance 1 and correlation coefficient $0.75^{1/2}$. The Y_{ij} were generated from the partially linear model (2.1), where $\theta(t) = \sin(2t)$ and $\beta = 1$, with normally distributed error with variance 1 and exchangeable correlation 0.4. For nonparametric estimation, we used local linear kernel estimation with the bandwidth choices 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6, and we assumed working independence. For parametric estimation, the working covariance V_i was set to be the true within-subject covariance of Y_i.

Table 1 reports the empirical biases and standard deviations, SD, of the estimated β from the profile-kernel and backfitting methods. It shows that the bias of the profile-kernel estimator is negligible over the range of bandwidths, but the bias of the backfitting estimator increases sharply as the bandwidth gets larger. This observation implies that backfitting estimator is more sensitive to bandwidth selection, as suggested by our theory. Table 1 also shows that the backfitting estimator has larger empirical standard deviations, about twice the size of the profile-kernel standard deviations. This observation agrees with

Table 1. Simulation results for 500 *clustered datasets.* $\widehat{\beta}_{\mathrm{P}}$ *stands for profile-kernel estimator,* $\widehat{\beta}_{\mathrm{BF}}$ *stands for backfitting estimator.*

		bandwidth					
Estimator		$h=0.1$	$h=0.2$	$h=0.3$	$h=0.4$	$h=0.5$	$h=0.6$
$\widehat{\beta}_{\mathrm{P}}$	bias	-0.0014	-0.0022	-0.0016	-0.0015	-0.0019	-0.0022
	SD	0.1608	0.1563	0.1539	0.1534	0.1533	0.1530
$\widehat{\beta}_{\mathrm{BF}}$	bias	0.0147	0.0641	0.1412	0.2473	0.3801	0.5385
	SD	0.3801	0.3730	0.3671	0.3610	0.3609	0.3625

our general theoretical result in Proposition II.1.

As a contrast, a numerical study was also carried out on independent data, where 500 datasets were generated, each comprising 300 subjects. Variables (T_i, X_i) and Y_i were generated in the same way as in the clustered-data case, except that the responses Y_i are independent of each other. The empirical biases and standard deviations from the two methods are reported in Table 2.

Table 2 shows a similar pattern in bias to that for clustered data, but the backfitting estimator has very similar standard deviations to those of the profile-kernel estimator. This indicates that the two estimators are nearly equally efficient for independent data, which is consistent with the traditional finding.

Another observation from Table 1 and 2 is that, since the working covariance matrix V_i used in the clustered-data simulation does not depend on X, the profile-kernel estimator is actually root-n consistent. This is the situation in Corollary II.1. Thus, it is natural that we observe negligible bias from profile-kernel estimation in both the clustered and independent cases.

Table 2. Simulation results for 500 *independent datasets.* $\widehat{\beta}_{\mathrm{P}}$ *stands for profile-kernel estimator,* $\widehat{\beta}_{\mathrm{BF}}$ *stands for backfitting estimator.*

		bandwidth					
Estimator		$h = 0.1$	$h = 0.2$	$h = 0.3$	$h = 0.4$	$h = 0.5$	$h = 0.6$
hline $\widehat{\beta}_{\mathrm{P}}$	bias	0.0015	0.0078	0.0090	0.0091	0.0096	0.0100
	SD	0.2444	0.2351	0.2328	0.2320	0.2320	0.2316
$\widehat{\beta}_{\mathrm{BF}}$	bias	0.0160	0.0802	0.1695	0.2783	0.4153	0.5802
	SD	0.2710	0.2555	0.2548	0.2603	0.2615	0.2663

2.5 An Application in Ophthalmology

In this section we analyze data from a prospective ophthalmology study on the use of intraocular gas in retinal repair surgeries (Meyers et al. 1992; Song and Tan 2000). Three different volumes of gas were injected into the eye before surgery in a total of 31 patients. The patients were then followed up 3 to 8 times over a 60-day time period, and the volume of the gas left in the eye at the follow-up times was recorded as a percentage of the initial gas level in that eye. The issue was to estimate the kinetics of the disappearance of the gas with respect to time. We let the response variable be the arcsin square root transformed percentage of gas left in the eye. The covariates are the initial level of gas concentration in the eye, denoted by X, and the follow-up observation time T, in the unit of days. We then assume that the transformed response follows the partially linear model (2.1).

Since there seems to exist a positive correlation among responses from the same patient, we need to incorporate a correlation structure into the estimation scheme. From the analysis of the residuals from the initial estimate assuming working independence (Diggle et al. 2002, Ch. 3), we found that the compound symmetry covariance matrix fit the data reasonably well. The estimated correlation is $\rho = 0.5442$, and the estimated variance is $\sigma^2 = 0.0678$.

The bandwidth h was chosen by 'leaving one subject out' cross validation (Rice and Silverman 1991; Härdle et al. 2000, §2.1.3) using the profile-kernel method. The exact procedure and a short justification of the use of this bandwidth selection method are given in the Appendix. We found that estimates with bandwidth ranging from 6 to 7 performed best and that differences among them were negligible. To ensure that the conclusion was not bandwidth dependent, we carried out the estimation for the bandwidth choices 6, 6.5, 7 and 8. We then applied the profile-kernel and the backfitting estimation methods as described in section 2.1 to these data, where the estimated compound symmetry working covariance matrix was assumed in the parametric estimation and the local linear smoother was used for nonparametric estimation. The results are given in Table 3.

Table 3. Ophthalmology example. Estimates and standard errors of the parametric coefficient using profile-kernel and backfitting methods.

		bandwidth			
Estimator		$h = 6.0$	$h = 6.5$	$h = 7.0$	$h = 8.0$
$\widehat{\beta}_{\mathrm{P}}$	estimate	0.1037	0.1024	0.1014	0.1041
	SE	0.0080	0.0072	0.0070	0.0063
$\widehat{\beta}_{\mathrm{BF}}$	estimate	0.0898	0.0890	0.0884	0.0879
	SE	0.0118	0.0119	0.0117	0.0151

Based on the results, we see that the percentage of gas volume left in the eye depends positively on the original gas concentration in the eye. The positive estimated values of β indicate that the percentage of gas volume left in the eye is high when the original level is high. This result is consistent with the findings of Song and Tan (2000). Moreover, both profile-kernel and backfitting estimation show a significant effect from the original gas concentration for all bandwidths considered. Regarding this aspect, the semiparametric model and estimation scheme considered here improve over Song and Tan (2000), where a more

complex model involving the same response and covariates suggests that the effect from the original gas concentration is insignificant. Our graphical diagnosis indicates that modeling the transformed responses with a semiparametric partially linear model provides sufficient flexibility to model the data reasonably well. The assumption violation observed in the parametric model considered in Song and Tan, which motivated their proposed model, no longer exists.

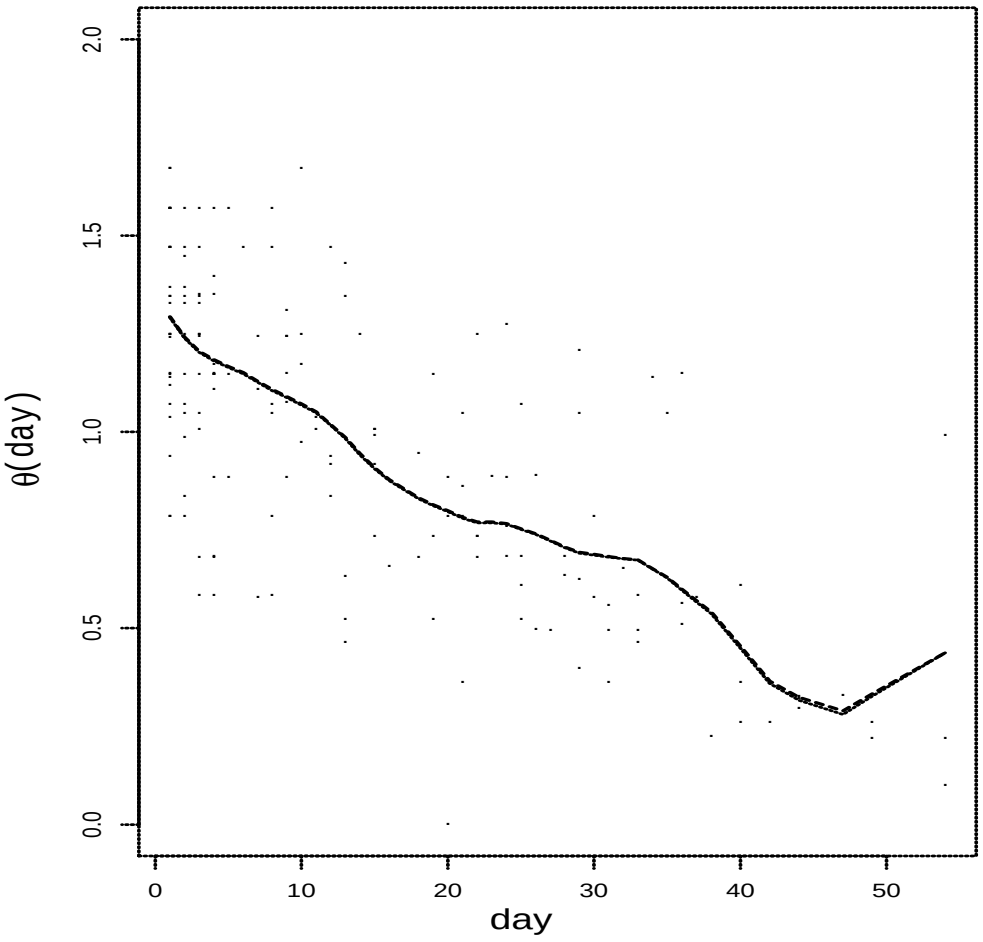

Figure 1. Ophthalmology example. Fitted curve for $\theta(t)$ by profile-kernel and backfitting methods, shown by dotted and dashed lines respectively, at bandwidth $h = 7$.

The time profile of the percentage of gas left in the eye is reflected by $\theta(t)$ in the semiparametric model, and we plot the estimated curve of $\theta(t)$ based on bandwidth $h = 7$ in Figure 1. The plots from profile-kernel and backfitting estimation are almost identical and indicate the same decreasing trend.

Finally, we note that in Table 3, for all bandwidths, the backfitting estimator had larger estimated standard error than the profile-kernel estimator. This observation agrees with the asymptotic properties and the simulation results in section 2.2 and section 2.3. It also suggests that for multivariate data one should no longer use backfitting as a substitute for the profile-kernel method.

2.6 Discussion

For a comment on the use of kernel methods versus penalized spline approaches as a general statistical methodology, and in particular the implementation of penalized splines via variance component model representations. We will let others comment on the somewhat controversial nature of penalized low-order basis splines versus smoothing splines, knot selection methods without penalties and estimation of smoothing parameters, the spline literature being in no agreement on these points.

The advantages and disadvantages of kernel methods and penalized splines using variance component model representations are fairly well known. As made clear by Ruppert et al. (2003, Ch. 1-2), penalized splines have the advantage that they are easily adopted into a wide variety of likelihood-type problems, by incorporating the penalties via a variance components representation.

However, a variance component model representation of penalized splines may not always make sense, as for example in the marginal generalized partially linear model in Lin and Carroll (2001) when the responses were non-Gaussian. There is no likelihood function for such problems in general, so that the penalized spline method would have to abandon

the variance component representation in favor of ad hoc approaches or alternatives which are known to have non-trivial computation and marginalization problems.

While variance component model representation of penalized splines can have certain advantages over kernels in terms of ease of method development, the opposite is true in terms of theoretical development. It is generally easy to analyze kernel methods, to develop appropriate bandwidths and to estimate these bandwidths in such a way that theoretical properties are ensured. In our Propositions 2 and 3, for example, we see that a standard bandwidth of order $n^{-1/5}$ will not result in $\sqrt{n}$-convergence rates for estimated β in general, while one of order $n^{-1/3}$ will do so. In contrast, the variance component model representation of penalized splines results in an estimated smoothing parameter, but it is generally unknown whether or not that smoothing parameter is estimated at rates that ensure asymptotic properties, especially for example for low-order basis representations where the number of knots is allowed to grow with the sample size.

Other examples of this difference in ease of theoretical development are available, such as in partially linear single-index models; Carroll et al. (1997) develop a semiparametric efficient kernel method for estimating the parameters in the model. We conjecture that the method of penalized low-order basis splines of Yu and Ruppert (2002) is also semiparametric efficient if the number of knots grows at an appropriate rate and if the smoothing parameter is appropriately selected, but deriving these two items in generality may well prove to be extremely challenging.

2.7 Proofs

The purpose of this section is to prove the propositions and results in chapter II.

In the following proofs, $\mathbf{T}$, $\mathbf{X}$, and $\mathbf{Y}$ denote the observations over all the clusters. That is $\mathbf{T} = (T_1^{\mathrm{T}}, \cdots, T_n^{\mathrm{T}})^{\mathrm{T}}$, and similarly for $\mathbf{X}$ and $\mathbf{Y}$. Also, $\mathbf{V}$ and $\mathbf{\Sigma}$ stand for the $nm \times nm$ assumed and true covariance matrices for all data, respectively.

Proof for Proposition II.1

*Proposition II.1*Under the assumption that both the working covariance matrix V and the true covariance matrix Σ depend only on T, the asymptotic variance of the backfitting estimator is at least as large as that of profile-kernel estimator. That is, $V_{\mathrm{BF}} - V_{\mathrm{P}}$ is positive semidefinite.

Proof: For clustered data, the asymptotic variance V_{BF} has its central component generated from $n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon$; as we will show in (A5). Similarly, the central component in the asymptotic variance V_{P} is from $n^{-1}\mathbf{X}^{\mathrm{T}}(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon$, which is $n^{-1}\widetilde{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon$ asymptotically. To compare V_{P} and V_{BF}, it is thus sufficient to compare the variances of the two central terms.

We now show that, under the condition that $\mathbf{V}$ and Σ depend only on $\mathbf{T}$, $\mathrm{cov}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon\} \geq \mathrm{cov}\{\widetilde{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon\}$. For the backfitting estimator,

$$\begin{aligned}
\mathrm{cov}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon\} &= E\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\Sigma(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\} \\
&= E\{m_X^{\mathrm{T}}(\mathbf{T})\mathbf{V}^{-1}(I-\mathbf{S})\Sigma(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}m_X(\mathbf{T})\} \\
&\quad + E[\mathrm{tr}\{\mathbf{V}^{-1}(I-\mathbf{S})\Sigma(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}\mathrm{cov}(\mathbf{X}|\mathbf{T})\}]. \qquad \text{(A.1)}
\end{aligned}$$

In this expression, $m_X(\mathbf{T})$ is generally nonzero and the first term is positive semidefinite because $\mathbf{V}^{-1}(I-\mathbf{S})\Sigma(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}$ is positive semidefinite. Also,

$$\begin{aligned}
\mathrm{cov}(\widetilde{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon) &= E\{\widetilde{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\Sigma(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}\widetilde{X}\} \\
&= E\{E(\widetilde{X}|\mathbf{T})^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\Sigma(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}E(\widetilde{X}|\mathbf{T})\} \\
&\quad + E[\mathrm{tr}\{\mathbf{V}^{-1}(I-\mathbf{S})\Sigma(I-\mathbf{S})^{\mathrm{T}}\mathbf{V}^{-1}\mathrm{cov}(\widetilde{X}|\mathbf{T})\}]. \qquad \text{(A.2)}
\end{aligned}$$

Note that

$$\begin{aligned}
E(\widetilde{X}_i|T_i) &= E\{X_i - m_X(T_i)|T_i\} = 0, \qquad \text{(A.3)} \\
\mathrm{cov}(\widetilde{X}_i|T_i) &= \mathrm{cov}\{X_i - m_X(T_i)|T_i\} = \mathrm{cov}(X_i|T_i).
\end{aligned}$$

Therefore, the first term in (A.2) is 0, and the second terms in (A.1) and (A.2) are identical. It follows that $\mathrm{cov}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon\} \geq \mathrm{cov}\{\widetilde{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon)\}$, and consequently $V_{\mathrm{BF}} \geq V_{\mathrm{P}}$.

Proof for Proposition II.2

Proposition II.2 Suppose that $h \propto n^{-\alpha}$, $1/5 \leq \alpha \leq 1/3$ and $n \to \infty$ and define

$$\widetilde{X} \;=\; X + \lim_{n\to\infty} \partial\widehat{\theta}(T;\beta)/\partial\beta.$$

Then $\widehat{\beta}_{\mathrm{P}}$ converges in distribution: $\sqrt{n}\{\widehat{\beta}_{\mathrm{P}} - \beta + h^2 b_{\mathrm{P}}(\beta,\theta)/2\} \longrightarrow \mathrm{N}(0, V_{\mathrm{P}})$, where

$$\begin{aligned} b_{\mathrm{P}}(\beta,\theta) &= E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1}E\{\widetilde{X}^{\mathrm{T}}V^{-1}\theta^{(2)}(T)\}, \\ V_{\mathrm{P}} &= E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1}E\{(Z_1-Z_2)^{\mathrm{T}}\Sigma(Z_1-Z_2)\}E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1}. \end{aligned}$$

Here $\widetilde{X} = \{X - m_X(T)\}$, $\Sigma = \mathrm{var}(Y|X,T), Z_1 = V^{-1}\widetilde{X}$, $Z_2 = (Z_2^1,\cdots,Z_2^m)^{\mathrm{T}}$, with $Z_2^j = \{\sum_{k=1}^m \sum_{l=1}^m E(\widetilde{X}^k V^{kl}|T^l = T^j)\}\, f_j(T^j)/\sum_{l=1}^m f_l(T^j)$, and V^{kl} denotes the (k,l) entry of V^{-1}.
proof: See Lin and Carroll (2001).

Proof for Proposition II.3

Proposition II.3 Under the same conditions as those of Proposition II.2, the backfitting estimator $\widehat{\beta}_{\mathrm{BF}}$ converges in distribution: $\sqrt{n}\{\widehat{\beta}_{\mathrm{BF}} - \beta + h^2 b_{\mathrm{BF}}(\beta,\theta)/2\} \longrightarrow \mathrm{N}(0, V_{\mathrm{BF}})$, where

$$\begin{aligned} b_{\mathrm{BF}}(\beta,\theta) &= E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1}E\{X^{\mathrm{T}}V^{-1}\theta^{(2)}(T)\}, \\ V_{\mathrm{BF}} &= E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1}E\{(Z_1^*-Z_2^*)^T\Sigma(Z_1^*-Z_2^*)\}E(\widetilde{X}^{\mathrm{T}}V^{-1}\widetilde{X})^{-1}, \end{aligned}$$

and $Z_1^* = V^{-1}X$, $Z_2^* = (Z_2^{*1},\cdots,Z_2^{*m})^{\mathrm{T}}$, with $Z_2^{*j} = \{\sum_{k=1}^m \sum_{l=1}^m E(X^k V^{kl}|T^l = T^j)\}\, f_j(T^j)$ $/\sum_{l=1}^m f_l(T^j)$.

proof: For the backfitting estimator, based on expression (2.3),

$$\widehat{\beta}_{\mathrm{BF}} - \beta = \{n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\mathbf{X}\}^{-1}\{n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})(\theta(\mathbf{T})+\varepsilon)\} \qquad \text{(A.4)}$$

In the first term of (A.4), with probability 1,

$$\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\mathbf{X} \to E[X_i^{\mathrm{T}}V_i^{-1}\{X_i - m_X(T_i)\}]$$

where $E[X_i^{\mathrm{T}}V_i^{-1}\{X_i - m_X(T_i)\}] = E[\{X_i - m_X(T_i)\}^{\mathrm{T}}V_i^{-1}\{X_i - m_X(T_i)\}] = E(\widetilde{X}_i^{\mathrm{T}}V_i^{-1}\widetilde{X}_i)$. In the second term of (A.4),

$$\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})(\theta(\mathbf{T})+\varepsilon) = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\theta(\mathbf{T}) + \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\varepsilon, \tag{A.5}$$

where the first term determines the bias of the backfitting estimator in Proposition 3:

$$\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}(I-\mathbf{S})\theta(\mathbf{T}) = -\frac{h^2}{2}E\left\{X_i^{\mathrm{T}}V_i^{-1}\theta^{(2)}(T_i)\right\} + o_{\mathrm{P}}(h^2),$$

see Opsomer and Ruppert (1997). The second term in (A.5) determines the 'centred' asymptotic distribution of the backfitting estimator and can be written as

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{\mathrm{T}}V_i^{-1}\varepsilon_i - \frac{1}{n}\sum_{i=1}^{n} X_i^{\mathrm{T}}V_i^{-1}\{\widehat{m}_\varepsilon(T_i) - m_\varepsilon(T_i)\},$$

where $\widehat{m}_\varepsilon(t)$ is the nonparametric smooth of ε at t and $m_\varepsilon(t)$ is its expectation.

Recalling that $K_h(s) = h^{-1}K(s/h)$, where K is a kernel function in nonparametric estimation, we have

$$\widehat{m}_\varepsilon(t;\beta) - m_\varepsilon(t) = w_2^{-1}(t)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} K_h(T_{ij}-t)\varepsilon_{ij} + o_{\mathrm{P}}(n^{-1/2}),$$

where $w_2(t) = \sum_{l=1}^{m} f_l(t)$. Proposition 3 follows by substituting this expression back into (A.4) and carrying out the expectation calculation.

Leave one subject out cross validation

This is to prove the validity of using "Leave one subject out cross validation" for choosing bandwidth in partially linear model estimation. This bandwidth selection is applied in section 2.4.

Proof: Let $\widehat{\beta}_{P[i]}$ and $\widehat{\theta}_{h[i]}(t) = \widehat{\theta}_{h[i]}(t, \widehat{\beta}_{P[i]})$ be the profile-kernel estimators of β and $\theta(T)$ without observations from subject i. We let $CV(h)$ be $n^{-1}\sum_i \left\{Y_i - X_i\widehat{\beta}_{P[i]} - \widehat{\theta}_{h[i]}(T_i)\right\}^{\otimes 2}$, where $v^{\otimes 2} = v^T v$, and consider the following decomposition:

$$CV(h) = n^{-1}\left(\sum_{i=1}^{n}\varepsilon_i^{\otimes 2} + \sum_{i=1}^{n}\left\{X_i(\widehat{\beta}_{P[i]} - \beta)\right\}^{\otimes 2} + \sum_{i=1}^{n}\left\{\widehat{\theta}_{h[i]}(T_i) - \theta(T_i)\right\}^{\otimes 2}\right.$$

$$- \sum_{i=1}^{n} \left[\varepsilon_i^T \left\{ X_i(\widehat{\beta}_{P[i]} - \beta) + \widehat{\theta}_{h[i]}(T_i) - \theta(T_i) \right\} + \left\{ X_i(\widehat{\beta}_{P[i]} - \beta) + \widehat{\theta}_{h[i]}(T_i) - \theta(T_i) \right\}^{\mathrm{T}} \varepsilon_i \right]$$

$$+ \sum_{i=1}^{n} \left[\left\{ X_i(\widehat{\beta}_{P[i]} - \beta) \right\}^{\mathrm{T}} \left\{ \widehat{\theta}_{h[i]}(T_i) - \theta(T_i) \right\} + \left\{ \widehat{\theta}_{h[i]}(T_i) - \theta(T_i) \right\}^{\mathrm{T}} \left\{ X_i(\widehat{\beta}_{P[i]} - \beta) \right\} \right] \Bigg) \quad \text{(A.6)}$$

We select the bandwidth to be h^* which minimizes $CV(h)$ in an interval of $[b_1 n^{-1/5}, b_2 n^{-1/5}]$, where $0 < b_1 < b_2 < \infty$. The first term in the right-hand side of (A.6) does not depend on h, while, under the conditions of Proposition 1 and for $h = O_p(n^{-1/5})$, the second term is negligible when compared to the third term. Direct derivations also show that, for $h = O_p(n^{-1/5})$, all other terms in (A.6) converge to 0 faster than the third term; that is, the bandwidth selection criterion that minimises $CV(h)$ is asymptotically equivalent to the criterion that minimises

$$n^{-1} \sum_{i=1}^{n} \left\{ \widehat{\theta}_{h[i]}(T_i) - \theta(T_i) \right\}^{\otimes 2} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ \widehat{\theta}_{h[i]}(T_{ij}) - \theta(T_{ij}) \right\}^2.$$

The asymptotic bias and variance structures in Lin and Carroll (2001) and Wang (2003) can be used to show that the selected optimal h is of order $n^{-1/5}$, as in the independent case.

CHAPTER III

SEMIPARAMETRIC APPROACH FOR LATENT COVARIATES IN MIXED EFFECTS MODELS

3.1 Introduction

Colon cancer is the second leading cause of death from cancer in the United States. There are strong epidemiological and clinical indications that a high proportion of the deaths could be prevented through appropriate diet (AICR 1997). Until recently, colon cancer development was thought to occur primarily due to increasing cell proliferation. This emphasis has now been shifted and there is considerable interest in linking colon tumor development to inhibition of apoptosis (cell death; see Heemels et al. 2000). When affected by carcinogen, apoptosis causes the termination of the cells with irreparable genetic damages that have the potential to progress into cancer cells. That is, by getting rid of damaged cells, apoptosis prevents them from proliferating to cancer cells.

There is a family of oncogenes that encode products adversely affecting apoptosis. An oncogene closely linked to apoptosis is *bcl-2*. Over-expression of the *bcl-2* gene leads to the suppression of apoptosis, thus allows tumor cells alive and proliferating. During the initial stage of colon carcinogenesis (e.g., the first 12 hours post exposure to a carcinogen), few apoptotic cells are formed and the main information is carried by an apoptosis-related gene (e.g., *bcl-2*). Therefore, in this study, we focus on investigating the relationship between *bcl-2* gene expression and the amount of DNA damage during this initial stage of colon cancer. Our primary interest is how the diet affects this relationship at different time after exposure to carcinogen. In the laboratory, the amount of DNA damage is measured by the DNA adduct level.

We now briefly describe the experiment. Thirty rats were divided evenly into two

groups. Each group was fed with one of the two diets: a fish oil supplemented or a corn oil supplemented for two weeks. After this, all 30 rats were injected with azoxymethane (AOM), a carcinogen that induces colon cancer. Three rats from each diet group were then euthanized at 0, 3, 6, 9, and 12 hours post exposure to carcinogen to measure the DNA adduct level and *bcl-2* gene expression in colonic cells. For each rat, 20 crypts were selected to measure *bcl-2*, and another group of 14 to 25 crypts were selected to measure the adduct level. These two measurements were taken at each cell within the selected crypts: about 14 to 25 cells in the crypts for *bcl-2* measurement, and 14 to 56 cells in the crypts for DNA adduct measurement.

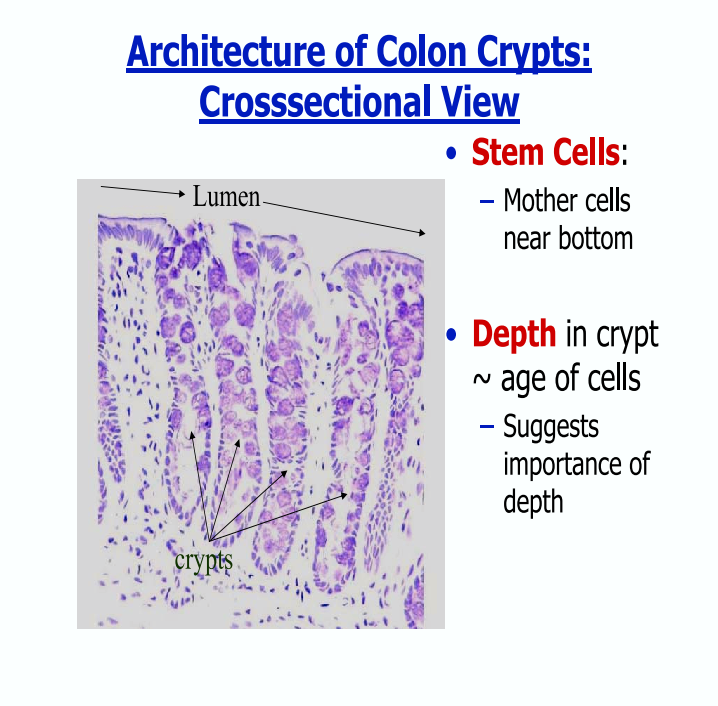

Figure 2. Structure of colon crypts

Colon crypts are discrete units where colonic cells replicate. Within each crypt, there are stationary, permanent cells called stem cells that generate all of the cells within that crypt. Daughter cells are formed at the crypt depth where the stem cells are located. As

more cells are created, they move up the crypt unit and exfoliate into the intestinal lumen. Thus, a cell's relative position within a crypt is an indicator of its age: cells at the bottom are younger, while cells near the top are older. In this data, the relative cell positions in a crypt are recorded ranging from 0 at the bottom to 1 at the top. Figure 2 shows the structure of the crypts in a colon.

Our goal is to understand the relationship between the two measurements, cell DNA adduct level and the *bcl-2* gene expression, as well as the effect of diet. More precisely, we want to investigate, in comparison to the corn oil supplemented diet, whether the fish oil supplemented diet helps reduce *bcl-2* gene expression when DNA damage increases. We need a mixed effects model for this relationship to accommodate the diet and time treatment effects, and also the random effects from 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 selected to take DNA adduct measurement, this same crypt could not be used again to measure *bcl-2*. Instead, a different crypt from the same rat was used. Since the number of cells varies from crypt to crypt, cells within different crypts have different relative cell positions. Consequently, the two measurements, *bcl-2* gene expression and DNA adduct level, were observed at different cell positions. It is a problem of misaligned measurements. Conventional regression methods are not appropriate here.

For the misaligned measurements problem, we propose a semiparametric statistical methodology. When the covariate values are unavailable, the traditionally common practice are the nearest neighbor (NN, Pielou 1961) method or the last observation carry-forward (LOCF, Mallinckrodt et al. 2003) method. These two methods, when applied to the colon carcinogenesis study, are to use the DNA adduct values that are observed at cell positions nearest to or immediately in front of the *bcl-2* measuring positions as the DNA adduct values corresponding to the *bcl-2* measurement. Our semiparametric approach is to assume

a latent process that relates DNA adduct to the relative cell position at rat level, and use this latent process for the latent covariate - DNA adduct - in the mixed effects model estimation. The rat level latent process can be estimated nonparametrically from the group of crypts selected for measuring DNA adduct. We refer to this practice of incorporating nonparametric estimates in parametric model estimation as the semiparametric approach. This semiparametric approach and the NN, LOCF methods, are all based on the fact that *bcl-2* and the unobserved corresponding DNA adduct are related through the cell position. However, the semiparametric approach takes into account that the two measurements are not only misaligned, but more importantly from different crypts. Another possible approach for this misaligned measurements problem is the EM (estimation maximization) method. However, due to the complexity of the colon carcinogenesis data, EM method is not as applicable as the semiparametric approach.

This semiparametric approach can be considered as an extension of Carroll and Wand (1991) and Pepe and Fleming (1991) in that a nonparametric estimation method is used to obtain the estimates of the unobserved covariate. The major differences are two fold: first, the previous two papers actually partially observe the true covariates while we do not. Secondly, the DNA adduct measurement forms a nonparametric mixed effect model with the marginal mean as a function of the relative cell position. That is, the observed surrogates are correlated while the previous work focus on independent responses. While the method developed in this paper is motivated by and applied to the colon cancer data, the proposed method has more general applications. In biological studies, it is common that true covariates are not directly observable and can only be postulated as coefficients or functional of another regression model (see Wang and Wang 2001, for a parametric example).

This chapter is organized as following. Section 3.2 formulates the mixed effects models for the colon cancer data, and describes the proposed semiparametric method. Section

3.3 develops the asymptotic properties of the semiparametric estimators. Section 3.4 gives a simulation study. Section 3.5 presents the application of our method to the colon cancer data. Finally the concluding remarks are in section 3.6, and the proofs of the results in this chapter are provided in section 3.7.

3.2 The Model and the Method

3.2.1 Model Specification

Due to the fact that the cell DNA adduct measurement is unavailable for the crypts where *bcl-2* gene expression is taken, we assume a rat level latent adduct process $X_i(t)$ for rat *i* at relative cell position *t*, $t \in (0,1)$. Here cell position, or the relative cell position, refers to the relative position of each cell within the selected crypt.

The following mixed effects model describes the relationship between *bcl-2* and the rat level latent covariate:

$$Y_{ijk}^{\text{tr}} = H(X_i^{\text{tr}}(t_{ijk}), \beta^{\text{tr}}) + Z_{ij}^{\text{tr}} b_{ij}^{\text{tr}} + \varepsilon_{ijk}^{\text{tr}}. \tag{3.1}$$

In this model, *i* is the index of rat, *j* is the index of crypt selected to measure *bcl-2*, *k* is the index of the cells in the selected crypt, and the sup-index "tr" is the treatment indicator for the diet and time group. The cell level *bcl-2* gene expression, Y_{ijk}^{tr}, is linked to the rat level DNA adduct covariate, X_i^{tr}, through the relative cell position t_{ijk}. β is the unknown fixed effect parameter vector and *H* is the known link function. The random effect, b^{tr}, coupling with rat and crypt level observed covariate, Z^{tr}, lay out the hierarchical rat and crypt-level dependency in model (3.1). Finally, we let γ^{tr} denote the unknown parameters in the error distribution of b^{tr} and the additive cell level error ε^{tr}. Hereafter, to ease the notation, we 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 conditional mean at cell position *t* . That is, we can link the observed cell DNA

adduct measurement at rat i, crypt j' and cell k', which is denoted as $W_{ij'k'}$, to $X_i(\cdot)$ through the following model:

$$W_{ij'k'} = X_i(t_{ij'k'}) + d_{ij'}(t_{ij'k'}) + e_{ij'k'}, \tag{3.2}$$

where $d_{ij'}$ denotes the crypt level variation and $e_{ij'k'}$ denotes the cell level additive error. Conditional on $X_i^{\mathrm{tr}}(t)$, we assume that measurements from different crypts are independent of each other. Model (3.2) is equivalent to the nonparametric model considered in Morris, et al. (2001). Note that, j' is the index of the crypts selected for DNA adduct measure, and k' is the index of the cell within that crypt. Due to the nature of the experiment, in no situation, $j' = j$ in (3.1) and (3.2).

Since crypts are randomly selected from the same rat to measure *bcl-2* and DNA adduct, biologically, the two groups of crypts should have similar properties. Therefore, it is reasonable to assume that the latent process for adduct X_i is the same for the two groups of crypts. This suggests that we can estimate the latent covariate X_i in model (3.1) from the nonparametric model (3.2).

3.2.2 Method Description

Since the latent covariate $X_i(t)$ can be considered as the rat-level conditional mean at cell position t, one way to recover this unobserved covariate is to estimate it nonparametrically. Our semiparametric method to estimate parameters, β and γ in model (3.1) can be described by two steps. Step 1: nonparametrically estimate the latent process $X_i(.)$ for each rat. That is to estimate $X_i(t)$ at cell position t based on model (3.2). Step 2: Use the estimated $X_i(t_{ijk})$ to replace the true $X_i(t_{ijk})$ in mixed effects model (3.1) and then estimate β and γ.

For estimation of the latent adduct process, we estimate within each rat separately, due to the fact that the rats are independent. To estimate $X_i(t)$, we use the local linear smoothing and assume the working independence correlation structure (Lin and Carroll 2000). That

is, we will ignore the correlation among observed DNA adduct from a common crypt in the process of nonparametric estimation of X_i.

For estimation of the parametric part, we use the generalized estimating equation (GEE). We will focus on two special cases of model (3.1) in the study of semiparametric estimation and its application in colon carcinogenesis. That is, the link function is taken as quadratic or generalized linear function.

$$Y_{ijk} = \beta_0 + \beta_1 X_i(t_{ijk}) + \beta_2 X_i(t_{ijk})^2 + Z_{ij} b_{ij}^{\mathrm{tr}} + \varepsilon_{ijk} \tag{3.3}$$

or

$$Y_{ijk} = H(X_i(t_{ijk})\beta) + Z_{ij} b_{ij}^{\mathrm{tr}} + \varepsilon_{ijk}. \tag{3.4}$$

For the mixed effects quadratic model (3.3), semiparametric estimator for β is,

$$\hat{\beta}^* = \{n^{-1}\sum_{i=1}^{n}\left[\underline{1}, \widetilde{X}_i(\underline{T_i}), \widetilde{X}_i^2(\underline{T_i})\right]^T \hat{\Sigma}_i^{-1}\left[\underline{1}, \widetilde{X}_i(\underline{T_i}), \widetilde{X}_i^2(\underline{T_i})\right]\}^{-1}\{n^{-1}\sum_{i=1}^{n}\left[\underline{1}, \widetilde{X}_i(\underline{T_i}), \widetilde{X}_i^2(\underline{T_i})\right]^T \hat{\Sigma}_i^{-1}\underline{Y_i}\} \tag{3.5}$$

Where n is the number of rats, $\underline{T_i}$ is the vector of observation cell positions for *bcl-2* in rat i, $\underline{Y_i}$ is the vector of observed *bcl-2*, $\underline{X_i}$ is the realization of $X_i(\cdot)$ at $\underline{T_i}$, $\widetilde{X}_i(\underline{T_i})$ is the nonparametrically estimated latent process X_i at $\underline{T_i}$. $\hat{\Sigma}_i$ is the estimated covariance matrix for the *bcl-2* measurements in rat i.

For the mixed effects generalized linear model, the semiparametric estimate of β can be calculated using scoring method using $\widetilde{X}_i(\underline{T_i})$. The estimator has the following asymptotic expression:

$$\begin{aligned}\hat{\beta}^* - \beta &= \{n^{-1}\sum_{i=1}^{n}\widetilde{X}_i(\underline{T_i})^T \triangle_i \hat{\Sigma}_i^{-1} \triangle_i \widetilde{X}_i(\underline{T_i})\}^{-1} \\ &\quad [n^{-1}\sum_{i=1}^{n}\widetilde{X}_i(\underline{T_i})^T \triangle_i \hat{\Sigma}_i^{-1}\{\underline{Y_i} - H(\widetilde{X}_i(\underline{T_i})\beta)\}]\{1 + o_P(1)\}.\end{aligned} \tag{3.6}$$

Here, $\triangle_i = H^{(1)}(\widetilde{X}_i(\underline{T_i})\beta)$ is the first order derivative of link function H.

Accounting for the nested experimental design in the colon carcinogenesis study: cells within a crypt and crypts within a rat, we consider the following simple structure of the covariance matrix in model (3.1, 3.3, 3.4):

$$\Sigma_i = \sigma_a^2 \mathbf{J}_{N_i} + \sigma_b^2 \text{diag}(\mathbf{J}_{K_{i,1}}, \cdots, \mathbf{J}_{K_{i,J_i}}) + \sigma_c^2 \mathbf{I}_{N_i}, \tag{3.7}$$

where σ_a^2 and σ_b^2 are the variance components for the random effects from rat and crypt respectively, and σ_c^2 is that for the random error. Thus $\gamma = (\sigma_a^2, \sigma_b^2, \sigma_c^2)$. $\mathbf{J}$ is matrix of entry 1; $\mathbf{I}$ is the identity matrix. N_i is the total number of *bcl-2* observations in rat i; J_i is the number of crypts for *bcl-2* observation in rat i; $K_{i,j}$ is the number of cells for *bcl-2* observation in crypt j of rat i. $\hat{\Sigma}_i$ is calculated by replacing γ by $\hat{\gamma}$.

When the goal is to construct consistent variance component estimator to be used in the estimated covariance in say (3.5) or in the asymptotic inference procedure, we replace X in the design matrix by $\tilde{X}$ and use the traditional maximum likelihood (ML) or restricted maximum likelihood (REML) estimators when true X were observed. To ease the presentation of our investigation on conditions that allow such an replacement and still result consistency outcomes, we focus our presentation of this subject on the use of an "fitting-of-constants" method (Henderson, 1953) in quadratic models. The outcomes are given in Section 3.3. This "fitting-of-constants" method was studied extensively by Fuller and Battese (1973) for the nested design. We choose this particular estimator for two reasons. First, its construction is particularly suitable for nested design so we use it in our data analysis. Secondly, the role played by the latent covariate can be clearly described. This feature simplifies the task of presenting the basic rationale behind the consistency of the estimated variance components and the conditions required for the consistency. The basic idea behind the estimator is to use simple regression analysis on transformed response and covariates to ease the task of obtaining estimated variance components. The estimation procedure and the exact form of estimators with observed covariates are given in Fuller and Battese

(1973) and summarized in section 3.7.

3.3 Asymptotic Properties of the Semiparametric Estimator

We develop the asymptotic properties of the semiparametric estimators based on using local linear smoother. The nonparametric estimate of the latent process $X_i(.)$ is obtained by local linear smoothing rat by rat. Hereafter, the sub-index for rat i is suppressed. Routine derivations give the following asymptotic expression which is used throughout the section. To ease the presentation, we assume that all crypts within a rat have the same number of cells.

$$\widetilde{X}(t) = X(t) + W_2^{-1}(t)\frac{1}{J'}\sum_{j'=1}^{J'}\sum_{k'=1}^{K'}\mathbf{K}_h(T_{j'k'}-t)\eta_{j'k'} + D^2X(t)h^2/2 + o_p\{(J')^{-1/2}\}, \quad (3.8)$$

where J' is the number of crypts for adduct observation in a specified rat, K' is the number of cells per crypt. Further, $D^2X(t)$ denotes the second derivative of $X(t)$ and $\mathbf{K}_h(v) = h^{-1}\mathbf{K}(v/h)$ with $\mathbf{K}$ being a symmetric, variance 1 kernel density function. $\eta_{j'k'} = d_{j'} + e_{j'k'}$ is the random error in the DNA adduct model. Let $f_T(t)$ be the marginal density of relative cell counts at cell position t. $W_2(t) = \sum_{k'=1}^{K'} f_T(t)$. For the mixed effects quadratic model, we obtain the following properties:

Proposition III.1. With 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

$$B_\beta = B^{-1}A(\beta)h^2 + o_p\{(J')^{-1/2}\} + O\{(J'K'h)^{-1}\}, \quad (3.9)$$

$$V_\beta = B^{-1} + (J'K'h)^{-1}B^{-1}C(\beta)B^{-1}, \quad (3.10)$$

where

$$B = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[\underline{1}, \underline{X_i}, \underline{X_i}^2]^T\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i}^2],$$

$$A(\beta) = h^2\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[\underline{0}, \underline{X_i}/2, \underline{X_i} * D^2\underline{X_i}]\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i}^2]\beta,$$

$$C(\beta) \quad = \quad \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} P_i^T \Sigma_i^w P_i.$$

$P_i = [\underline{0}, p_i, p_i * 2\underline{X_i}]$ with $p_i = \Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i}^2]\beta$, and * standing for the element-wise vector/matrix product. $(J'K'h)^{-1}\Sigma_i^w$ is the covariance matrix of the nonparametric local linear estimate in (3.8) evaluated at $\underline{T_i}$. Σ_i^w is a $(JK)\times(JK)$ matrix, with the diagonal entry $(\Sigma_i^w)_{l,l} = \gamma_K(0)(\sigma_d^2+\sigma_e^2)\ / f_T(T_{il})$, and the off-diagonal entry $(\Sigma_i^w)_{l_1,l_2} = h\sigma_d^2 W_{12}(T_{il_1}, T_{il_2})$ / $\{f_T(T_{il_1})W_2(T_{il_2})\}$, where $W_{12}(t_1,t_2) = \sum_{k_1'\neq k_2'} f_{(T_1,T_2)}(t_1,t_2)$, $\gamma_{\mathbf{K}}(0) = \int K^2(s)ds$, and $f_{(T_1,T_2)}(t_1,t_2)$ is the bivariate density of relative frequency of having cells at positions t_1, t_2 within the same crypt.

When $nh^4 \to 0$, $\hat{\beta}^*$ is $\sqrt{n}$-consistent. A sketch of proof of the proposition III.1 is in appendix.

Remarks:

1. Both (3.9) and the second term in (3.10) goes to 0 as $1/(J'K'h)$ and h go to 0 and J' goes to ∞. Thus the bandwidth selection is not determined by the crypt number alone, but the number of observations of DNA adduct from all crypts within a rat. Even though we do not need to assume $K' \to \infty$, we carefully keep track of the role of K' in the asymptotic distribution. In the colon carcinogenesis study, though the crypt number J' is around twenty, $J'K'$ is much bigger, ranging from several hundreds to one thousand.

2. Covariance of semiparametric estimator $\hat{\beta}^*$ has two parts. B^{-1} is the asymptotic covariance matrix when X is observed. The second term in (3.10) is from the nonparametric estimation of the latent covariate X_i's. As $J'K'h \to \infty$, the second part will diminish, and variance of $\hat{\beta}^*$ is mainly from B^{-1}. Also, when matrix B is diagonal, the variance of intercept $\hat{\beta}_0$ and its covariance with $\hat{\beta}_1$ and $\hat{\beta}_2$ are nearly not affected by the nonparametric estimation, due to the fact that the first row and first column in $C(\beta)$ are zero, see (B.5) in section 3.7.

3. Estimation on the covariance of $\hat{\beta}^*$ can be obtained by,

$$\widehat{var}(\hat{\beta}^*) = \hat{B}^{-1} + \hat{B}^{-1}(\sum_{i=1}^{n} \underline{\zeta_i^*}\underline{\zeta_i^{*T}})\hat{B}^{-1}, \tag{3.11}$$

based on the proof in the Appendix, where

$$\begin{aligned} \hat{B} &= \sum_{i=1}^{n} [\underline{1}, \underline{\widetilde{X}_i}, \underline{\widetilde{X}_i^2}]^T \hat{\Sigma}_i^{-1} [\underline{1}, \underline{\widetilde{X}_i}, \underline{\widetilde{X}_i^2}], \\ \underline{\zeta_i^*} &= [\underline{0}, \underline{\mathcal{W}_i'}, 2\underline{\widetilde{X}_i} * \underline{\mathcal{W}_i'}]^T \hat{\Sigma}_i^{-1} [\underline{1}, \underline{\widetilde{X}_i}, \underline{\widetilde{X}_i^2}] \hat{\beta}^*. \end{aligned}$$

and $\underline{\mathcal{W}_i'}$ is the random error in the nonparametric estimation of latent process $\underline{X_i}$, as defined in the appendix.

4. When there are more than one treatment groups, indicator variables can be used to enlarge the design matrix. Because of the block diagonal nature of the setup, the extension is straightforward.

For the estimates of the variance components in (3.7), we obtain the following consistency property.

Proposition III.2. Estimates of variance components σ_a^2, σ_b^2 and σ_c^2 in the mixed effect model (3.1), with the nonparametrically estimated X_i, are consistent as $h \to 0$, $J' \to \infty$ and $J'K'h \to \infty$.

A sketch of the proof of Proposition III.2 is given in section 3.7. Fuller and Battese (1973) have shown that estimation of the variance components does not affect the asymptotic properties of their weighted least square estimator. Following their derivations, we can show that when Proposition III.2 holds, the asymptotic property for the semiparametric estimated β remains the same when we replace Σ by $\widehat{\Sigma}$. We just need to obtain the following properties of semiparametric estimator of β for Σ.

For the mixed effects general linear model (3.4), the semiparametric estimator has the similar asymptotic property.

Proposition III.3. Under the same condition of Proposition 1, the semiparametric estimator (3.6) is consistent and asymptotically normally distributed, with

$$\begin{aligned} B_\beta &= B^{-1}A(\beta)h^2 + o_p\{(J')^{-1/2}\} + O\{(J'K'h)^{-1}\}, \\ V_\beta &= B^{-1} + (J'K'h)^{-1}B^{-1}C(\beta)B^{-1}, \end{aligned}$$

where

$$\begin{aligned} B &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \underline{X_i}^T \triangle_i \Sigma_i^{-1} \triangle_i \underline{X_i}, \\ A(\beta) &= h^2 \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \underline{X_i}^T \triangle_i \Sigma_i^{-1} \triangle_i D^2 \underline{X_i}\beta/2, \\ C(\beta) &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \underline{X_i}^T \triangle_i \Sigma_i^{-1} \triangle_i \Sigma_i^{w} \triangle_i \Sigma_i^{-1} \triangle_i \underline{X_i}\beta^2 \end{aligned}$$

A sketch proof of this proposition is given in section 3.7..

3.4 Simulation Study

To study the numerical performance of the proposed semiparametric approach, we conduct a small simulation study.

Fifty ($n = 50$) subjects are generated. For each subject, we generate $K = 40$ response (Y) within each of the $J = 20$ crypts. Also, in that same subject, another $J' = 30$ crypts and $K' = 50$ covariate (X) within each crypt are generated. The cell-positions for observing X are evenly spaced, and those for observing Y are randomly uniformly distributed, both in $[0,1]$.

The 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 $r_{i2} \sim N(0,1)$. The observed covariate, W, is generated by model (3.2), with $d_{ij'}(t) \equiv d_{ij'}$, $d_{ij'}$ and $e_{ij'k'}$ are independent of each other and normally distributed with means 0 and variance σ_d^2 and σ_e^2, respectively. In this simulation, the variance components are chosen as $\sigma_d = 0.3$, $\sigma_e = 0.7$. The observed response Y is generated by mixed quadratic model

(3.3) with covariance structure as specified in (3.7). The parametric values are chosen to be $\beta_0 = 1$, $\beta_1 = -2$, $\beta_2 = 1$, and variance components $\sigma_a = 1$, $\sigma_b = 1$, $\sigma_c = 3$. There are 300 replications in the simulation. We carry out the estimation of the mixed quadratic model with latent covariate by four methods: (1). GEE with the true covariate values (True), (2) the nearest neighbor method (NN); (3) the last observation carry-forward method (LOCF), and (4) the semiparametric method (Semip). For the semiparametric method, estimates are computed over several bandwidths. We report In table 4 the Monte-Carlo mean and the Monte-Carlo standard deviation of the estimated quadratic coefficients. Also, for the semiparametric estimates, we report the estimated standard error based on their asymptotic distribution in Proposition III.1.

Table 4. Simulation results: Mean is the Monte-Carlo mean of the estimates, SD is the Monte-Carlo standard deviation, and ESE is the estimated standard error from the asymptotic distribution.

Method		$\beta_0 = 1$	$\beta_1 = -2$	$\beta_2 = 1$
True	mean	0.998	-1.995	0.999
	SD	0.166	0.023	0.005
NN	mean	0.779	-0.929	0.677
	SD	0.185	0.068	0.017
LOCF	mean	0.808	-0.927	0.667
	SD	0.185	0.070	0.018
Semip				
$h = 0.03$	mean	0.980	-1.983	0.994
	SD(ESE)	0.167 (0.157)	0.031(0.027)	0.007 (0.006)
$h = 0.04$	mean	0.987	-1.994	0.996
	SD(ESE)	0.167 (0.157)	0.030 (0.027)	0.007 (0.006)
$h = 0.05$	mean	9.995	-2.004	0.998
	SD(ESE)	0.166 (0.156)	0.030 (0.026)	0.007 (0.006)
$h = 0.06$	mean	1.002	-2.015	1.000
	SD(ESE)	0.166 (0.156)	0.029 (0.026)	0.007 (0.006)
$h = 0.07$	mean	1.011	-2.028	1.002
	SD(ESE)	0.165 (0.156)	0.029 (0.025)	0.007 (0.006)

In table 4, we see that the LOCF and the NN estimates are biased toward null findings,

due to the effect of attenuation. However, the semiparametric method yields much better results than the other two methods. In addition, the estimated standard errors from the asymptotic distribution of the semiparametric estimator are quite close to the Monte-Carlo standard deviation. Here we see that $\hat{\beta}^*$ from semiparametric estimation has more deviation than that from the regular GEE by using the true covariate values (which is unattainable in practice for misaligned measurements problem). This additional deviation from nonparametric estimation on X_i's decreases as $J'K'h$ gets big.

For the estimation of the variance components in the mixed quadratic model, based on semiparametric approach as in (B.6, B.7, B.8), the estimated variance components at $h = 0.05$ are: $\tilde{\sigma}_a = 0.999$, $\tilde{\sigma}_b = 1.022$, and $\tilde{\sigma}_c = 3.003$, with the Monte Carlo SD as 0.064, 0.030, and 0.019 respectively.

3.5 Analysis of Colon Carcinogenesis Data

In this section, we summarize the procedures and outcomes of the analysis to the colon carcinogenesis data introduced in Section 3.1. The goal of this study is to investigate whether increase in DNA adduct level induces an increment or a decrement in *bcl-2* gene expression, also whether the increment slopes vary with diet within a individual rat. Recall that the response, *bcl-2*, and the covariate, DNA adduct, were not observed from the same crypts within a rat. Since the relative cell positions for observing these two measurements differed from crypt to crypt, it is a problem with misaligned measurements. We assume each rat's DNA adduct level follows a specific process X_i. We then postulate the relationship between *bcl-2* and DNA adduct by the semiparametric approach.

For this colon carcinogenesis study, these are several features to be noticed. First, as introduced earlier, the relative cell position actually indicates the age of the cell. To detect any effect from the cell age, we carry out analysis within each portion of the crypt separately: the bottom $1/3$ section, the middle $1/3$ section, and the top $1/3$ section. This has

been a common practice in the field of animal studies of colon carcinogenesis and the simple models provide directly interpretable outcomes for easy communication. Secondly, we focus on the analysis using the mixed effects linear model, that is to use linear link function H in the general mixed effects model (3.1). For the model checking purpose, we carry out regression using mixed effects quadratic models, and find out that for most treatment groups, the quadratic coefficients are not significantly different from zero. Therefore, we choose to use a mixed effects linear model instead of the quadratic model. The properties we developed for quadratic models apply here, and the linear model allows easier interpretation for the diet effect on the *bcl-2* vs. DNA adduct relationship. Finally, we use the centered regression. That is to regress on the centered DNA adduct. By centered DNA adduct, we mean the DNA adduct values centered around their rat level mean within each section: bottom, middle, and top. The reason behind "centered" regression is as following. Due to subject to subject variation, different rats could have different range of adduct values even within the same treatment group. The analysis using the centered adduct captures the common structure of rat specific pattern. In fact, it models the trend between *bcl-2* and DNA adduct within each rat and then summarizes the trends over all the rats within a same treatment group. On the contrary, the regression using uncentered adduct practically models the trend between the rat level averages of *bcl-2* and DNA adduct cross different rats.

In summary, we study the colon cancer data by the linear mixed effects model on the centered adduct values, at each of the three sections of crypts. Based on how the values of adduct — the latent covariate in primary model (3.1) are obtained, we consider and compare three methods: the proposed semiparametric method, the NN method, and the LOCF method. For the semiparametric approach, bandwidth selection is from the leave one subject out cross-validation (Rice and Silverman 1991) with the selected bandwidth $h = 0.05$. It is worth noting that the values of the generalized cross validation function

changes little over a range of bandwidths around 0.05 and the outcomes vary little using bandwidths in that neighborhood.

Analyses are performed in all three sections of the crypts. Here, we focus on reporting the results for the top section. Due to the research by Hong et al. (2000), it is the location where the proportions of apoptosis differ between fish oil enhanced and corn oil enhanced diets in the later stage of carcinogenesis. The results for the other two sections are either non-significant or similar to the findings in this top section. In table 5, we list the estimated coefficients from the three methods: semiparametric (Semip), last observation carry-forward (LOCF), and the nearest neighbor (NN). For the semiparametric method, we report the estimated intercepts and slopes for the 10 treatment groups. Also reported are the standard errors of the estimates and the *p*-values for the contrast between the two diets within each of the 5 time groups. While the estimates of intercepts and slopes are from point estimation on the colon carcinogenesis data, estimates of the standard error and *p*-value are from parametric bootstrap, based on the semiparametric regression results. As comparison, we also report the estimated slopes from LOCF and NN methods. We see that these estimated slopes are shrunk toward zero. However, they lead to non-contradicting conclusions as the semiparametric estimates, in the sense that the contrasts between the two diet groups are of the similar pattern, though of much lower significance.

From table 5, we can see that during the initial stage of colon cancer development (first 12 hours after exposure to carcinogen), except for time group 0, the fish oil fed rats have significantly smaller slopes than the corn oil fed rats. More specifically, as DNA damage increases, for the fish oil fed rats, the *bcl-2* gene expression either decreases as at time 3, 9, and 12 hours after injection of carcinogen, or increases at a much lower rate than the corn oil fed rats as at time 6.

As we know, *bcl-2* is an oncogene that prohibits apoptosis. Under-expression of *bcl-2* gene expression enhances higher activity of apoptosis, and consequently more active self-

Table 5. Estimates for the linear mixed effects model of bcl-2 versus DNA adduct: SE is the standard error, p-val is the p value for the comparison between the two diets within each time group.

time	diet	semip estimates		semip P-val		LOCF	NN
		intercept (SE)	slope (SE)	diff int	diff slope	slope	slope
0	fish	33.54 (2.79)	2.32 (0.53)	0.38	< 0.01	0.052	0.069
	corn	37.08 (2.94)	0.83 (0.43)			-0.081	-0.031
3	fish	25.13 (2.79)	-0.79 (0.28)	0.04	< 0.01	-0.092	-0.108
	corn	33.08 (2.80)	0.35 (0.28)			0.050	-0.034
6	fish	25.57 (2.81)	0.18 (0.25)	0.43	< 0.01	0.084	0.014
	corn	28.51 (2.81)	2.25 (0.37)			0.118	0.065
9	fish	19.48 (2.88)	-1.28 (0.27)	0.54	< 0.01	-0.021	-0.115
	corn	22.38 (2.89)	0.92 (0.36)			-0.023	-0.041
12	fish	24.99 (2.72)	-0.52 (0.30)	0.72	< 0.01	0.065	-0.022
	corn	26.42 (3.04)	0.42 (0.27)			0.093	0.044

termination of the cancer-prone damaged cells. Therefore, our findings in table 5 suggest that during initial stage of colon carcinogenesis, in comparison to corn oil diet, fish oil diet suppresses the increment in the gene expression of *bcl-2* when the DNA damage increases and thus potentially has a better chance in promoting apoptosis. Plots of the regression curves from semiparametric approach are shown in Figure 3.

3.6 Summary

For the colon carcinogenesis study, the objective is to find the relationship between the cell DNA damage (measured by DNA adduct) and the *bcl-2* gene expression, during the initial stage of colon cancer. Also, we are interested in how the diet (fish oil versus corn oil) affects this relationship at different times post the exposure to carcinogen. A mixed effects model is appropriate for this study to incorporate the diet, time effects, and the random effects from rat and crypt.

The two measurements — DNA adduct level and *bcl-2* gene expression were mea-

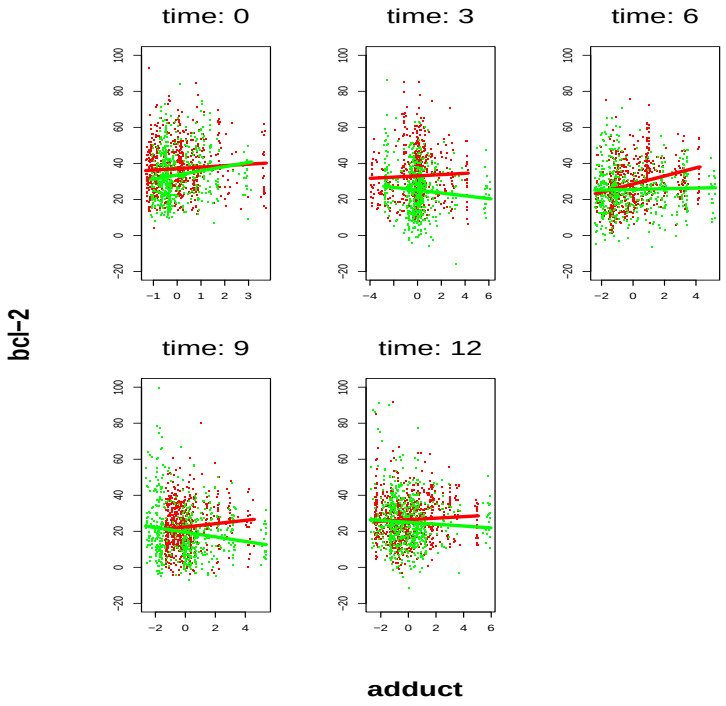

Figure 3. Fitted regression curves for bcl-2 vs. DNA adduct at each time points from semiparametric approach: light points and lines are for the fish oil diet group, dark points and lines are for the corn oil group, bandwidth $h = 0.05$.

sured from different crypts though in the same rats. Since different crypts inside the colon have different cell numbers, the two measurements were not measured at the same cell positions. It is a problem of misaligned measurements. Consequently, there is the latent covariate (DNA adduct level measured at the cell positions of *bcl-2*) in the mixed effects model. We propose the semiparametric approach for this misaligned measurements problem. Based on the theoretical investigation and the simulation results, our semiparametric approach produces estimates with nice properties. Compared with the two traditional methods, nearest neighbor approach and the last observation carry-forward approach, our

semiparametric approach has better performance.

Biologically, the semiparametric results support that fish oil helps to reduce the risk of developing colon cancer at the initiation stage. During this early stage, fish oil can lower the rate of increase in *bcl-2* gene expression when the DNA damage increases. Since over-expression of *bcl-2* gene inhibits apoptosis, reduced rate of increase in *bcl-2* leads to more active functioning of apoptosis, thus reduces the danger of colon cancer by getting rid of more damaged cells.

3.7 Proofs

This section is to prove the results in chapter III.

We derive the asymptotic results based on local linear smoothing for nonparametric estimation of latent process $X_i(\cdot)$. Suppressing the rat index i:

$$\widetilde{X}(t) = X(t) + W_2^{-1}(t)\frac{1}{J'}\sum_{j'=1}^{J'}\sum_{k'=1}^{K'} K_h(T_{j'k'}-t)\eta_{j'k'} + D^2X(t)2h^2/2 + o_p\{(J')^{-1/2}\} \quad \text{(B.1)}$$

and,

$$\begin{aligned}\widetilde{X}^2(t) &= \Bigg[X^2(t) + \frac{2}{J'}X(t)W_2^{-1}(t)\sum_{j'=1}^{J'}\sum_{k'=1}^{K'} K_h(T_{j'k'}-t)\eta_{j'k'} + X(t)D^2X(t)h^2 \\ &+ \{\frac{1}{J'}W_2^{-1}(t)\sum_{j'=1}^{J'}\sum_{k'=1}^{K'} K_h(T_{j'k'}-t)\eta_{j'k'}\}^2 + o_p\{(J')^{-1/2}\}\Bigg]\{1+o_p(1)\}.\end{aligned}$$

In the following, denote $\underline{X_i}$ as $X_i(\underline{T_i})$, which is the realization of latent process X_i at the *bcl-2* measuring positions $\underline{T_i}$ in rat i, and $\underline{\widetilde{X}_i}$ for the nonparametric estimate of $\underline{X_i}$. Similarly, we define the vectors $\underline{\mathcal{W}_i'}$ whose entry corresponding to cell position T_{ijk} is

$W_2^{-1}(T_{ijk})\frac{1}{J'}\sum_{j'=1}^{J'}\sum_{k'=1}^{K'} K_h(T_{ij'k'}-T_{ijk})\eta_{ij'k'}$. That is, $\underline{\mathcal{W}_i'}$ contains the random errors in the local linear smoothing estimate of $\underline{X_i}$. Each entry in $\underline{\mathcal{W}_i'}$ is $O_p\{(J'K'h)^{-1/2}\}$, and each entry in $\underline{\mathcal{W}_i'}^2$ is $O\{(J'K'h)^{-1}\}$.

In the asymptotics study, the number of crypts and the number of cells within a crypt for observing the response Y in each rat are assumed as fixed, and denoted as J_i and K_{ij} respectively. So, $\underline{T_i}$, $\underline{X_i}$, and $\underline{\mathcal{W}_i'}$ are of fixed dimension $\sum_{j=1}^{J_i} K_{ij}$.

Proof of Proposition III.1

For the semiparametric estimator $\hat{\beta}^*$ in (3.5), $\hat{\beta}^* = A_1^{-1}A_2$, where

$$\begin{aligned} A_1 &= \frac{1}{n}\sum_{i=1}^{n}[\underline{1},\underline{\widetilde{X}_i},\underline{\widetilde{X}_i}^2]^T\Sigma_i^{-1}[\underline{1},\underline{\widetilde{X}_i},\underline{\widetilde{X}_i}^2], \\ A_2 &= \frac{1}{n}\sum_{i=1}^{n}[\underline{1},\underline{\widetilde{X}_i},\underline{\widetilde{X}_i}^2]^T\Sigma_i^{-1}Y_i. \end{aligned}$$

In A_1, denote

$$\widetilde{A}^i = [\underline{1},\underline{\widetilde{X}_i},\underline{\widetilde{X}_i}^2]^T\Sigma_i^{-1}[\underline{1},\underline{\widetilde{X}_i},\underline{\widetilde{X}_i}^2],$$

so, $\widetilde{A}^i$ is the matrix with entry $\widetilde{A}^i{}_{r,s} = (\underline{\widetilde{X}_i}^{r-1})^T\Sigma_i\underline{\widetilde{X}_i}^{s-1}$,for $r,s = 1,2,3$.

Due to the fact that $\underline{\widetilde{X}_i} = \underline{X_i} + O(h^2) + O_p\{(J'K'h)^{-1/2}\} + o_p\{(J')^{-1/2})\}$ and $\underline{\widetilde{X}_i}^2 = \underline{X_i}^2 + O(h^2) + O\{(J'K'h)^{-1}\} + o_p\{(J')^{-1/2})\}$,

$$\widetilde{A}^i{}_{r,s} \to (\underline{X_i}^{r-1})^T\Sigma_i^{-1}\underline{X_i}^{s-1}, \qquad \text{for} \quad i = 1,\cdots,n. \tag{B.2}$$

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

Write

$$\begin{aligned} A_2 &= A_{21} + A_{22} + A_{23} + A_{24} + A_{25}, \quad \text{where} \\ A_{21} &= \frac{1}{n}\sum_{i=1}^{n}[\underline{1},\underline{X_i},\underline{X_i^2}]^T\Sigma_i^{-1}[\underline{1},\underline{X_i},\underline{X_i^2}]\beta + \frac{1}{n}\sum_{i=1}^{n}[\underline{1},\underline{X_i},\underline{X_i^2}]^T\Sigma_i^{-1}\underline{\varepsilon_i}, \\ A_{22} &= \frac{1}{n}\sum_{i=1}^{n}[\underline{0},\underline{\mathcal{W}_i'},2\underline{X_i}*\underline{\mathcal{W}_i'} + \underline{\mathcal{W}_i'}^2]^T\Sigma_i^{-1}[\underline{1},\underline{X_i},\underline{X_i^2}]\beta, \\ A_{23} &= \frac{1}{n}\sum_{i=1}^{n}[\underline{0},D^2\underline{X_i}h^2/2 + o_p\{(J')^{-1/2}\},\underline{X_i}*D^2\underline{X_i}h^2 + o_p\{(J')^{-1/2}\}]\Sigma_i^{-1}[\underline{1},\underline{X_i},\underline{X_i^2}]\beta, \\ A_{24} &= \frac{1}{n}\sum_{i=1}^{n}[\underline{0},\underline{\mathcal{W}_i'},2\underline{X_i}*\underline{\mathcal{W}_i'} + \underline{\mathcal{W}_i'}^2]^T\Sigma_i^{-1}\underline{\varepsilon_i}, \end{aligned} \tag{B.3}$$

$$A_{25} = \frac{1}{n}\sum_{i=1}^{n}[\underline{0}, D^2\underline{X_i}h^2/2 + o_p\{(J')^{-1/2}\}, \underline{X_i} * D^2\underline{X_i}h^2 + o_p\{(J')^{-1/2}\}]\Sigma_i^{-1}\underline{\varepsilon_i}. \quad \text{(B.4)}$$

Note that A_{21} corresponds to the mean and variance terms in the quadratic regression if X were observed. The first order bias of $\hat{\beta}^*$ originates from A_{22} and A_{23}, the leading extra variance is also from A_{22}. For the bias,

$$\begin{aligned} A_{23} &= \begin{bmatrix} 0 \\ \frac{1}{2}(\beta_0 a_{00} + \beta_1 a_{01} + \beta_2 a_{02}) \\ \beta_0 a_{10} + \beta_1 a_{11} + \beta_2 a_{12} \end{bmatrix} h^2 + o_p\{(J')^{-1/2}\} \\ &= A(\beta)h^2 + o_p\{(J')^{-1/2}\} \end{aligned}$$

where $a_{rs} = \frac{1}{n}\sum_{i=1}^{n}(\underline{X_i}^r)^T\Sigma_i^{-1}D^2\underline{X_i} * \underline{X_i}^s$, for $r = 0, 1$, and $s = 0, 1, 2$, are finite.

$$\begin{aligned} E(A_{22}) &= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} 0 \\ 0 \\ E(\underline{\mathcal{W}_i^2})^T\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i}^2]\beta \end{bmatrix} \\ &= \frac{\gamma_{\mathbf{K}}(0)(\sigma_d^2 + \sigma_e^2)}{J'K'h} \cdot \frac{1}{n}\begin{bmatrix} 0 \\ 0 \\ \mathcal{D}^{-1}(\underline{T_i})^T\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i}^2] \end{bmatrix} \end{aligned}$$

thus, $E(A_{22}) = O\{(J'K'h)^{-1}\}$ provided that $\frac{1}{n}\sum_{i=1}^{n}\mathcal{D}(\underline{T_i})^T\Sigma_i^{-1}[\underline{1}, \underline{X_i}, \underline{X_i^2}] < \infty$, with $\mathcal{D}(\underline{T_i}) = [\cdots, f^{-1}(T_{ijk}), \cdots]$.

For the covariance, $cov(A_{21}) = \frac{1}{n}B$,

$$cov(A_{22}) = \frac{1}{n^2}\sum_{i=1}^{n}\begin{bmatrix} \underline{0}^T \\ p_i^T \\ 2(\underline{X_i} * p_i)^T \end{bmatrix} cov(\underline{W_i})[\underline{0}, p_i, 2\underline{X_i} * p_i]$$

$$
\begin{aligned}
&= \frac{1}{n}(J'K'h)^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & c_{00} & 2c_{01} \\ 0 & 2c_{10} & 4c_{11} \end{bmatrix} \\
&= \frac{1}{n}(J'K'h)^{-1}C(\beta) \qquad \text{(B.5)}
\end{aligned}
$$

with $c_{rs} = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\underline{X_i}^s * p_i)^T\Sigma_i^w(p_i * \underline{X_i}^t)$, for $r = 0,1$ and $s = 0,1$. $(J'K'h)^{-1}\Sigma_i^w = cov(\underline{\mathcal{W}_i'})$ is the covariance of the estimated $\underline{X_i}$.

Proposition III.1 follows that A_{21} and A_{22} are independent given $\underline{X}$.

Proof of Proposition III.2

Fuller and Battese (1973) gave the variance components estimators for nested design, and shown that they are unbiased. For estimator of $\gamma = (\sigma_a^2, \sigma_b^2, \sigma_c^2)$ in the mixed model, they have the following expressions:

$$
\begin{aligned}
\hat{\sigma}_c^2 &= \hat{\tau}^T\hat{\tau}/(N_2 - N_1 - p + \lambda_{12}) && \text{(B.6)} \\
\hat{\sigma}_b^2 &= \frac{\hat{u}^T\hat{u} - (N_2 - n - p + \lambda_1)\hat{\sigma}_c^2}{N_2 - \mathrm{tr}(H_b)} && \text{(B.7)} \\
\hat{\sigma}_a^2 &= \frac{\hat{v}^T\hat{v} - (N_2 - p)\hat{\sigma}_c^2 - \{N_2 - \mathrm{tr}(H_{a_1})\}\hat{\sigma}_b^2}{N_2 - \mathrm{tr}(H_{a_2})} && \text{(B.8)}
\end{aligned}
$$

where τ is the vector of residuals from the centered regression of $y_{ijk} - \bar{y}_{ij.}$ on $x_{ijkm} - \bar{x}_{ij.m}$, $m = 1, \cdots, p$; u is the residual of $y_{ijk} - \bar{y}_{i..}$ on $x_{ijkm} - \bar{x}_{i..m}$; v is that of y_{ijk} on x_{ijkm}. N_1 is the total number of sub-units (crypts for observing *bcl-2*) $\sum_{i=1}^{n} J_i$. N_2 is the total number of sub-sub-units(cells for observing *bcl-2*) $\sum_{i=1}^{n}\sum_{j=1}^{J_i} K_{ij}$. n is the number of subjects (rats). λ_1 and λ_{12} are the number of x-variables that have constant values for the sub-units and the sub-sub units respectively. p is dimension of β. For the mixed quadratic model, $p = 3$ and $\lambda_1 = \lambda_{12} = 1$. X is the design matrix at the true value of the covariates. H_b, H_{a_1}, and H_{a_2} are the hat matrices in the estimation of variance components. $H_b = (\mathbf{X} - \overline{\mathbf{X}}_{(1..)})^T(\mathbf{X} - \overline{\mathbf{X}}_{(1..)})\sum_{i=1}^{n}\sum_{j=1}^{J_i} K_i^2(\overline{\mathbf{X}}_{ij.} - \overline{\mathbf{X}}_{i..})^T(\overline{\mathbf{X}}_{ij.} - \overline{\mathbf{X}}_{i..})$;

$H_{a_1} = (\mathbf{X}^T\mathbf{X})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i} K_i^2 \overline{\mathbf{X}}_{ij.}^T \overline{\mathbf{X}}_{ij.}$; $H_{a_2} = (\mathbf{X}^T\mathbf{X})^{-1}\sum_{i=1}^{n} J_i^2 K_i^2 \overline{\mathbf{X}}_{i..}^T \overline{\mathbf{X}}_{i..}$. K_i is the number of sub-sub-units within each sub-unit of subject i. To ease the presentation, we assume this number is the same for all the sub-units with a subject. That is, K_i is the number of cells in each crypt for observing bcl-2 at rat i. Here the notations related to design matrix X are the same as in Fuller and Battese (1973).

The semiparametric variance component estimators $\tilde{\sigma_a}^2$, $\tilde{\sigma_b}^2$, and $\tilde{\sigma_c}^2$ are of the same expression as in (B.8), (B.7), and (B.6), 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 on the covariates, which are contained in the following terms:

$$\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} \tag{B.9}$$

$$\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T \tag{B.10}$$

where (B.9) determines the terms in hat matrices H_b, H_{a1}, and H_{a2}; (B.10) determines the estimated sum of squared errors $\hat{\tau}^T\hat{\tau}$, $\hat{u}^T\hat{u}$, and $\hat{v}^T\hat{v}$.

For mixed effects quadratic model (3.3),

$$\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \sum_{i=1}^{n}[\underline{1}, \underline{\widetilde{X}_i}, \underline{\widetilde{X}_i}^2]^T[\underline{1}, \underline{\widetilde{X}_i}, \underline{\widetilde{X}_i}^2]$$

By refering to (B.2), $\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} \to \frac{1}{n}\mathbf{X}^T\mathbf{X}$ in probability as $J' \to \infty$, $h \to 0$ and $J'K'h \to \infty$. Similarly, $\frac{1}{n}\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T \to \frac{1}{n}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ in probability. Thus, the semiparametric variance components estimators ($\tilde{\sigma_a}^2$, $\tilde{\sigma_b}^2$, $\tilde{\sigma_c}^2$) converge to $(\hat{\sigma_a}^2, \hat{\sigma_b}^2, \hat{\sigma_c}^2)$ in probability. Since $(\hat{\sigma_a}^2, \hat{\sigma_b}^2, \hat{\sigma_c}^2)$ are unbiased, the semiparametric variance components estimators are thus consistent.

Proof of Proposition III.3

Estimate $\hat{\beta}^*$ is solution to,

$$\sum_{i=1}^{n} \underline{\widetilde{X}_i} \triangle_i \Sigma_i \{\underline{Y_i} - H(\underline{\widetilde{X}_i}\beta)\} = 0$$

So,

$$\hat{\beta}^* - \beta = (\frac{1}{n}\sum_{i=1}^{n} \underline{\widetilde{X}_i}^T \triangle_i \Sigma_i \triangle_i \underline{\widetilde{X}_i})^{-1} \{\frac{1}{n}\sum_{i=1}^{n} \underline{\widetilde{X}_i}^T \triangle_i \Sigma_i [\underline{Y_i} - H(\underline{\widetilde{X}_i}\beta)]\}\{1 + o_p(1)\}.$$

The rest of this proof can be done following the structure that proves proposition III.1.

CHAPTER IV

CONCLUSION

4.1 Study of the Partially Linear Models

Our study on profile-kernel and backfitting methods for the partially linear model concludes that the two methods are not equivalent for correlated data. When the data is longitudinal/clustered, the backfitting method is more sensitive to the choice of bandwidth, and it generally has larger variation than the profile-kernel method. Though the asymptotic normality of the backfitting estimator is formulated following the common estimation scheme of Zeger and Diggle (1994), the general result on the asymptotic efficiency of the two methods apply to any estimation setup and nonparametric smoothers. The simulation results show that, for both independent and correlated cases, the backfitting estimator is more sensitive to the bandwidth selection. The bias of the backfitting estimator can be large when the selected bandwidth is big. However, when it comes to the standard deviation, backfitting is similar to the profile-kernel for independent data, but its deviation is much larger than profile-kernel for correlated data.

The ophthalmology example indicates that the partially linear model yields more efficient results than a completely parametric model. In the parametric model for the disappearance of intraocular gas in retinal repair surgeries, complicated transformations were adopted to model the unknown time effect. The logistic transformation applied to the mean function causes the effect of initial dosage to be not significant in the inference. However, our partially linear model uses the nonparametric term for the unknown time effect. Not only is the model much simpler, it also detects that the initial dosage of intraocular gas significantly affects the time profile of disappearance of this gas.

4.2 Study of the Semiparametric Approach for Colon Carcinogenesis Study

For the study on colon carcinogenesis, our semiparametric approach for the misaligned measurements problem produces results with good properties. This is demonstrated by the asymptotic study, and also the simulation outcomes. Our semiparametric approach makes use of the latent process for the unobservable covariate corresponding to the response. Compared with the last observation carry-forward and nearest neighbor methods, a semiparametric approach can be consistent under reasonable conditions. In addition, it can reach the estimation efficiency of a regular likelihood estimator as $J'K'h \to \infty$.

Based on semiparametric outcomes for the colon carcinogenesis data, we conclude that the fish oil lowers the rate of increase in *bcl-2* gene expression, when the DNA damage increases in cells. Therefore, fish oil appears advantagous relative to corn oil in preventing colon cancer. During the initial stage of colon cancer, fish oil promotes more active functioning of apoptosis, and thus makes it possible for the body to get rid of more defective cells.

REFERENCES

American Institute for Cancer Research (AICR), World Research Fund in Association with the American Dietetic Association (1997), *Food, Nutrition, and the Prevention of Cancer: a Global Perspective,* Washington, D.C., 85-96.

Buja, A., Hastie,T.J. and Tibshirani, R. J. (1989), " Linear smoothers and additive models (with Discussion)," *Annals of Statistics*,17, 453-555.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M.P. (1997), "Generalized linear single-index models," *Journal of American Statistical Association*, 92, 477-489.

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

Diggle, P. J., Heagerty, P. J., Liang, K.-Y. and Zeger, S. L. (2002), *The Analysis of Longitudinal Data*, 2nd ed, Oxford: Oxford University Press.

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

Härdle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models,* Heidelberg, Germany: Physica-Verlag.

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

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 and Prevention*, 9, 819-826.

Karp, G. (2002), *Cell and Molecular Biology: Concepts and Experiments*, 3rd ed, New York: Wiley.

Lin, X. and Carroll, R. J. (2000)," Nonparametric function estimation for clustered data when the predictor is measured without/with error," *Journal of American Statistical Association*, 95, 520-534.

———— (2001), " Semiparametric regression for clustered data using generalized estimating equations," *Journal of American Statistical Association*, 96, 1045-1056.

Mallinckrodt, C.H., Clark, S.W., Carroll, R.J., and Molenberghs, G. (2003), "Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations," *Journal of Biopharmaceutical Statistics*, 13, 179-190.

Meyers, S. M., Ambler, J. S., Tan, M., Werner, J. C., and Huang, S. S. (1992), "Variation of perfluorpropane disappearance after vitrectomy," *Retina*, 12, 359-363.

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 American Statistical Association*, 96, 816-827.

Opsomer, J. D. and Ruppert, D. (1997), " Fitting a bivariate additive model by local polynomial regression," *Annals of Statistics*, 25, 186-211.

———— (1999), "A root-n consistent backfitting estimator for semiparametric additive modeling," *Journal of Computational Graphical Statistics*, 8, 715-32.

Pepe, M. S. and Couper, D. (1997), "Modeling partly conditional means with longitudinal data," *Journal of American Statistical Association*, 92, 991-998.

Pepe, M.S. and Fleming, T.R. (1991), " A general nonparametric method for dealing with errors in missing or surrogate data," *Journal of 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. (1986), "Convergence rates for partially splined models," *Statistics and Probability Letter*, 4, 204-208.

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

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Severini, T. A. and Staniswalis, J. G. (1994), " Quasi-likelihood estimation in semiparametric models," *Journal of American Statistical Association*, 89, 501-512.

Song X.-K. and Tan, M. (2000), " Marginal models for longitudinal continuous proportional data," *Biometrics*, 56, 496-502.

Speckman, P. E. (1988), "Regression analysis for partially linear models," *Journal of Royal Statistical Society Series B*, 50, 413-436.

Wang, N., Carroll, J.R., and Lin, X. (2004), "Efficient semiparametric marginal estimation for longitudinal/clustered data," *Journal of American Statistical Association*, to appear.

Wang, J. L. and Wang, W. (2001), " Comment on 'Semiparametric and nonparametric regression analysis of longitudinal data'," *Journal of American Statistical Association*, 96, 119-123.

Wang, N. (2003), " Marginal nonparametric kernel regression accounting for within-subject correlation," *Biometrika*, 90, 43-52.

Yu, Y. and Ruppert, D. (2002), " Penalized spline estimation for partially linear single index models," *Journal of American Statistical Association*, 97, 1042-1054.

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.

VITA

Zonghui Hu was born in Dalian, China. She received a Bachelor of Science degree in mathematics from Dalian University of Technology, a Master of Science degree in mathematics from Dalian Univeristy of Technology, and a Master of Science degree in mathematics from Texas A&M University. In August 2000, she was admitted to the Ph.D. program in the Department of Statistics at Texas A&M University. She received her Ph.D. degree in December 2004. Her permanent address is

Xinggongnan street, 54-402

Dalian, Liaoning Provence, 116021

People's Republic of China