EFFICIENT INFERENCE IN GENERAL SEMIPARAMETRIC REGRESSION MODELS

A Dissertation

by

ARNAB MAITY

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

August 2008

Major Subject: Statistics

EFFICIENT INFERENCE IN GENERAL SEMIPARAMETRIC REGRESSION MODELS

A Dissertation

by

ARNAB MAITY

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

Approved by:

Chair of Committee,	Raymond J. Carroll
Committee Members,	Bani K. Mallick
	Soumendra N. Lahiri
	Ursula Mueller-Harknett
	Edward R. Dougherty
Head of Department,	Simon J. Sheather

August 2008

Major Subject: Statistics

ABSTRACT

Efficient Inference in General Semiparametric Regression Models. (August 2008)

Arnab Maity, B.Stat., Indian Statistical Institute;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Raymond J. Carroll

Semiparametric regression has become very popular in the field of Statistics over the years. While on one hand more and more sophisticated models are being developed, on the other hand the resulting theory and estimation process has become more and more involved. The main problems that are addressed in this work are related to efficient inferential procedures in general semiparametric regression problems.

We first discuss efficient estimation of population-level summaries in general semiparametric regression models. Here our focus is on estimating general population-level quantities that combine the parametric and nonparametric parts of the model (e.g., population mean, probabilities, etc.). We place this problem in a general context, provide a general kernel-based methodology, and derive the asymptotic distributions of estimates of these population-level quantities, showing that in many cases the estimates are semiparametric efficient.

Next, motivated from the problem of testing for genetic effects on complex traits in the presence of gene-environment interaction, we consider developing score test in general semiparametric regression problems that involves Tukey style 1 d.f form of interaction between parametrically and non-parametrically modeled covariates. We develop adjusted score statistics which are unbiased and asymptotically efficient and can be performed using standard bandwidth selection methods. In addition, to over-

come the difficulty of solving functional equations, we give easy interpretations of the target functions, which in turn allow us to develop estimation procedures that can be easily implemented using standard computational methods.

Finally, we take up the important problem of estimation in a general semiparametric regression model when covariates are measured with an additive measurement error structure having normally distributed measurement errors. In contrast to methods that require solving integral equation of dimension the size of the covariate measured with error, we propose methodology based on Monte Carlo corrected scores to estimate the model components and investigate the asymptotic behavior of the estimates.

For each of the problems, we present simulation studies to observe the performance of the proposed inferential procedures. In addition, we apply our proposed methodology to analyze nontrivial real life data sets and present the results.

To Bijali R. Maity and Amal K. Maity

ACKNOWLEDGMENTS

The work presented here reflects research goals I have shared with my mentors for the past several years. Thank you to the many fine faculty at Indian Statistical Institute, above all Dr. Debapriya Sengupta for motivating and encouraging me to progress far beyond my expectations.

The five years that I spent at Texas A&M University will be always with me. Dr. Raymond J. Carroll taught me so much more than I expected. I would like to thank him with all my heart for his valuable teachings and excellent mentorship, without which I would not have succeeded in my goal. I would like to thank Dr. Yanyuan Ma for her excellent guidance and advice during my graduate studies. I applaud and thank you for your dedication. I would also like to thank Dr. Bani Mallick, Dr. Soumendra N. Lahiri, Dr. Ursula Mueller-Harknett and Dr. Edward R. Dougherty for their support throughout my studies. To the many fine faculty in the Department of Statistics at Texas A&M University, thank you for fine education that I received at Texas A&M University.

To my parents, thank you for always believing in me and supporting me, even when you are thousands of miles away.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

We consider a wide class of semiparametric regression models in which interest focuses on population-level quantities that combine both the parametric and the nonparametric parts of the model. Special cases in this approach include generalized partially linear models, generalized partially linear single-index models, structural measurement error models, and many others. For estimating the parametric part of the model efficiently, profile likelihood kernel estimation methods are well established in the literature. Here our focus is on estimating general population-level quantities that combine the parametric and nonparametric parts of the model (e.g., population mean, probabilities, etc.). We place this problem in a general context, provide a general kernel-based methodology, and derive the asymptotic distributions of estimates of these population-level quantities, showing that in many cases the estimates are semiparametric efficient. For estimating the population mean with no missing data, we show that the sample mean is semiparametric efficient for canonical exponential families, but not in general. We apply the methods to a problem in nutritional epidemiology, where estimating the distribution of usual intake is of primary interest and semiparametric methods are not available. Extensions to the case of missing response data are also discussed.

Many of the regular semiparametric regression models assume that there is no interaction present between the parametric and the nonparametric components of the

---

This dissertation follows the style of the *Journal of the Royal Statistical Society.*

models. However, this may not be true in many real life situations. Motivated from the problem of testing for genetic effects on complex traits in the presence of gene-environment interaction, we consider developing score test in general semiparametric regression problems that involves Tukey style 1 d.f form of interaction between parametrically and non-parametrically modeled covariates. We find that the score-test in this type of model, as recently developed by Chatterjee et al. (2007) in the fully parametric setting, is biased and requires undersmoothing to be valid in the presence of non-parametric components. Moreover, in the presence of repeated outcomes, the asymptotic distribution of the score test depends on the estimation of functions which are defined as solutions of complex integral equations, making implementation difficult and computationally taxing. We develop adjusted score statistics which are unbiased and asymptotically efficient and can be performed using standard bandwidth selection methods. In addition, to overcome the difficulty of solving functional equations, we give easy interpretations of the target functions, which in turn allow us to develop estimation procedures that can be easily implemented using standard computational methods. We present simulation studies to evaluate type-I error and power of the proposed method compared to a naive test that does not consider interaction. Finally, we illustrate our methodology by analyzing data from a case-control study of colorectal adenoma designed to investigate the association between colorectal adenoma and the candidate gene *NAT2* in relation to smoking history.

Finally, we consider the problem of estimation in a general semiparametric regression model when covariates are measured with an additive measurement error structure having normally distributed measurement errors. The semiparametric part of the model arises with a covariate measured without error being modeled nonparametrically. In contrast to methods that require solving integral equation of dimension

the size of the covariate measured with error, we propose methodology based on Monte Carlo corrected scores to estimate the model components and investigate the asymptotic behavior of the estimates. For example, our method applies to repeated measures data, while integral equation methods are not practical in this context. The resulting methods are functional, i.e., they make no assumptions about the distribution of the error-prone covariates. We investigate the special cases of logistic partially linear and multivariate partially linear measurement error models and compare our results with the existing literature. We also present a simulation study to illustrate the performance of our method. Finally, we demonstrate our method by applying it to Nevada Test Site (NTS) Thyroid Disease Study data.

CHAPTER II

EFFICIENT ESTIMATION OF POPULATION-LEVEL SUMMARIES IN GENERAL SEMIPARAMETRIC REGRESSION MODELS

II.1. Introduction

Often, in semiparametric regression models, one is interested in estimating a population quantity such as the mean, variance, probabilities, etc. The unique feature of the problem is that the quantities of interest are functions of both the parametric and nonparametric parts of the model. We will also allow for partially missing responses, but handling such a modification is relatively easy. The main aim of this chapter is to estimate population quantities that involve both the parametric and nonparametric parts of the model, and to do so efficiently and in considerable generality.

We will construct estimators of these population-level quantities that exploit the semiparametric structure of the problem, derive their limiting distributions, and show in many cases that the methods are semiparametric efficient. The work is motivated by and illustrated with an important problem in nutritional epidemiology, namely estimating the distribution of usual intake for episodically consumed foods such as red meat.

A special simple case of our results is already established in the literature (Wang, Linton and Härdle 2004, and references therein), namely the partially linear model

$$Y_i \quad = \quad X_i^{\mathrm{T}}\beta_0 + \theta_0(Z_i) + \xi_i, \tag{2.1}$$

where $\theta_0(\cdot)$ is an unknown function and $\xi_i = \text{Normal}(0, \sigma_0^2)$. We allow the responses to be partially missing, important in cases that the response is difficult to measure but the predictors are not. Suppose that Y is partially missing, and let $\delta = 1$ indicate that Y is observed, so that the observed data are $(\delta_i Y_i, X_i, Z_i, \delta_i)$. Suppose further that Y is missing at random, so that $\text{pr}(\delta = 1|Y, X, Z) = \text{pr}(\delta = 1|X, Z)$.

Usually, of course, the main interest is in estimating β_0 efficiently. This is not the problem we discuss, because in our example the parameters β_0 are themselves of relatively minor interest. In their work, Wang et al. (2004) estimate the marginal mean $\kappa_0 = E(Y) = E\{X^{\mathrm{T}}\beta_0 + \theta_0(Z)\}$. Note how this combines both the parametric and nonparametric parts of the model. One of the results of Wang et al. is that if one uses only the complete data that Y is observed, then fits the standard profile likelihood estimator to obtain $\widehat{\beta}$ and $\widehat{\theta}(\cdot, \widehat{\beta})$, it transpires that a semiparametric efficient estimator of the population mean κ_0 is $n^{-1}\sum_{i=1}^n \{X_i^{\mathrm{T}}\widehat{\beta} + \widehat{\theta}(Z_i, \widehat{\beta})\}$. If there are no missing data, the sample mean is also semiparametric efficient.

Actually, quite a bit more is true even in this relatively simple Gaussian case. Let $\mathcal{B} = (\beta^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ and let $\widehat{\mathcal{B}}$ and $\widehat{\theta}(\cdot, \widehat{\mathcal{B}})$ be the profile likelihood estimates in the complete data, see for example Severini and Wong (1992) for local constant estimation and Claeskens and Carroll (2007) for local linear estimation. Consider estimating any functional $\kappa_0 = E[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ for some function $\mathcal{F}(\cdot)$ that is thrice continuously differentiable: this of course includes such quantities as population mean, probabilities, etc. Then one very special case of our results is that the semiparametric efficient estimate of κ_0 is just $\widehat{\kappa} = n^{-1}\sum_{i=1}^n \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}$.

In contrast to Wang et al. (2004), we deal with general semiparametric models

and general population-level quantities. Thus, consider a semiparametric problem in which the loglikelihood function given (X, Z) is $\mathcal{L}\{Y, X, \theta(Z), \mathcal{B}\}$. If we define $\mathcal{L}_{\mathcal{B}}(\cdot)$ and $\mathcal{L}_{\theta}(\cdot)$ to be derivatives of the loglikelihood with respect to $\mathcal{B}$ and $\theta(Z)$, we have the properties that $E[\mathcal{L}_{\mathcal{B}}\{Y, X, \theta_0(Z), \mathcal{B}_0\}|X, Z] = 0$ and similarly for $\mathcal{L}_{\theta}(\cdot)$. We use profile likelihood methods computed at the observed data. With missing data, this local linear kernel version of the profile likelihood method of Severini and Wong (1992) works as follows. Let $K(\cdot)$ be a smooth symmetric density function with bounded support, let h be a bandwidth, and let $K_h(z) = h^{-1}K(z/h)$. For any fixed $\mathcal{B}$, let $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ be the local likelihood estimator obtained by maximizing in (α_0, α_1)

$$\sum_{i=1}^{n} \delta_i K_h(Z_i - z)\mathcal{L}\{Y_i, X_i, \alpha_0 + \alpha_1(Z_i - z), \mathcal{B}\}, \tag{2.2}$$

and then setting $\widehat{\theta}(z, \mathcal{B}) = \widehat{\alpha}_0$. The profile likelihood estimator of $\mathcal{B}_0$ modified for missing responses is obtained by maximizing in $\mathcal{B}$

$$\sum_{i=1}^{n} \delta_i \mathcal{L}\{Y_i, X_i, \widehat{\theta}(Z_i, \mathcal{B}), \mathcal{B}\}. \tag{2.3}$$

Our estimator of $\kappa_0 = E[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ is then

$$\widehat{\kappa} = n^{-1} \sum_{i=1}^{n} \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}. \tag{2.4}$$

We emphasize that the possibility of missing response data and finding a semiparametric efficient estimate of $\mathcal{B}_0$ is not the focus of the article. Instead, the focus is on estimating quantities $\kappa_0 = E[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ that depend on both the parametric and nonparametric parts of the model: this is a very different problem than simply estimating $\mathcal{B}_0$. Previous work in the area has considered only the partially linear model and only estimation of the population mean: our work deals with general

semiparametric models and general population-level quantities.

An outline of this chapter is as follows. In Section II.2 we discuss the general semiparametric problem with loglikelihood $\mathcal{L}\{Y, X, \theta(Z), \mathcal{B}\}$ and a general goal of estimating $\kappa_0 = E[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$. We derive the limiting distribution of (2.4) and show that it is semiparametric efficient. We also discuss the general problem where the population quantity κ_0 of interest is the expectation of a function of Y alone, and describe doubly-robust estimators in this context.

In Section II.3, we consider the class of generalized partially linear single index models (Carroll, Fan, Gijbels and Wand 1997). Single index modeling, see Härdle and Stoker (1989) and Härdle, Hall and Ichimura (1993), is an important means of dimension reduction, one that is finding increased use in this age of high-dimensional data. We develop methods for estimating population quantities in the generalized partially linear single index modeling framework, and show that the methods are semiparametric efficient.

Section II.4 describes an example from nutritional epidemiology that motivated this work, namely estimating the distribution of usual intake of episodically consumed foods such as red meat. The model used in this area is far more complex than the simple partially linear Gaussian model (2.1), and while the population mean is of some interest, of considerably more interest is the probability that usual intake exceeds thresholds. We will illustrate why in this context one cannot simply adopt the percentages of the observed responses that exceeding a threshold.

Section II.5 describes three issues of importance: (a) bandwidth selection (Section

II.5.1); (b) the efficiency and robustness of the sample mean when the population mean is of interest (Section II.5.2); and numerical and theoretical insights into the partially linear model and the nature of our assumptions (Section II.5.3). An interesting special case is of course the partially linear model when κ_0 is the population mean. For this problem, we show in Section II.5.2 that with no missing data, the sample mean is semiparametric efficient for canonical exponential families but not of course in general, thus extending and clarifying the results of Wang et al. (2004) that were specific to the Gaussian case.

All technical results are given in an Appendix.

II.2. Semiparametric Models with a Single Component

II.2.1. Main Results

We benefit from the fact that the limiting expansions for $\widehat{\mathcal{B}}$ and $\widehat{\theta}(\cdot)$ are essentially already well-known, with the minor modification of incorporating the missing response indicators. Let $f(z)$ be the density function of Z, assumed to have bounded support and to be positive on that support. Let $\Omega(z) = f(z)E\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z = z\}$. Let $\mathcal{L}_{i\theta}(\cdot) = \mathcal{L}_{\theta}\{Y_i, X_i, \theta_0(Z_i), \mathcal{B}_0\}$, etc. Then it follows from standard results (see the Appendix for more discussion) that as a minor modification of the work of Severini and Wong (1992),

$$\begin{aligned}\widehat{\theta}(z,\widehat{\mathcal{B}}) - \theta_0(z) &= (h^2/2)\theta_0^{(2)}(z) - n^{-1}\sum_{i=1}^{n}\delta_i K_h(Z_i - z)\mathcal{L}_{i\theta}(\cdot)/\Omega(z) \\ &\quad +\theta_{\mathcal{B}}(z,\mathcal{B}_0)(\widehat{\mathcal{B}} - \mathcal{B}_0) + o_p(n^{-1/2}); \qquad (2.5)\\ \widehat{\mathcal{B}} - \mathcal{B}_0 &= \mathcal{M}_1^{-1}n^{-1}\sum_{i=1}^{n}\delta_i\epsilon_i + o_p(n^{-1/2}), \qquad (2.6)\end{aligned}$$

where

$$\begin{aligned}
\theta_{\mathcal{B}}(z,\mathcal{B}_0) &= -E\{\delta\mathcal{L}_{\mathcal{B}\theta}(\cdot)|Z=z\}/E\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z=z\}; && (2.7)\\
\epsilon_i &= \{\mathcal{L}_{i\mathcal{B}}(\cdot)+\mathcal{L}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i,\mathcal{B}_0)\}; && (2.8)\\
\mathcal{M}_1 &= E(\delta\epsilon\epsilon^{\mathrm{T}}) = -E[\delta\{\mathcal{L}_{\mathcal{B}\mathcal{B}}(\cdot)+\mathcal{L}_{\mathcal{B}\theta}(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z,\mathcal{B}_0)\}],
\end{aligned}$$

and where under regularity conditions, (2.5) is uniform in z. Conditions guaranteeing (2.6) are well-known, see the Appendix.

Define

$$\begin{aligned}
D_i(\cdot) &= -\mathcal{L}_{i\theta}(\cdot)\frac{E\{\mathcal{F}_\theta(\cdot)|Z_i\}}{E\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z_i\}};\\
\mathcal{M}_2 &= E\{\mathcal{F}_{\mathcal{B}}(\cdot)+\mathcal{F}_\theta(\cdot)\theta_{\mathcal{B}}(Z,\mathcal{B}_0)\}.
\end{aligned}$$

In the Appendix, we show the following result.

Result 1 Suppose that $nh^4 \to 0$ and that (2.5)-(2.6) hold, the former uniformly in z. Suppose also that Z has compact support, that its density is bounded away from zero on that support, and that the kernel function also has a finite support. Then the estimator $\widehat{\kappa}$ of $\kappa_0 = E[\mathcal{F}\{X,\theta_0(Z),\mathcal{B}_0\}]$ is semiparametric efficient in the sense of Newey (1990). In addition, as $n\to\infty$,

$$\begin{aligned}
n^{1/2}(\widehat{\kappa}-\kappa_0) &= n^{-1/2}\sum_{i=1}^{n}\Big\{\mathcal{F}_i(\cdot)-\kappa_0+\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\delta_i\epsilon_i+\delta_i D_i(\cdot)\Big\}+o_p(1) && (2.9)\\
&\Rightarrow \ \mathrm{Normal}(0,\mathcal{V}_0), && (2.10)
\end{aligned}$$

where $\mathcal{V}_0 = E\{\mathcal{F}(\cdot)-\kappa_0\}^2+\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\mathcal{M}_2+E\{\delta D^2(\cdot)\}$.

Remark 1 In order to obtain asymptotically correct inference about κ_0, there are two possible routes. The first is to use the bootstrap: while Chen, Linton and Van

Keilegom (2003) only justify the bootstrap for estimating $\mathcal{B}_0$, we conjecture that the bootstrap works for κ_0 as well. More formally, one requires only a consistent estimate of the limiting variance in (2.10). This is a straightforward exercise, although programming-intense: one merely replaces all the expectations by sums in that expression and all the regression functions by kernel estimates.

Remark 2 Our analysis of semiparametric efficiency in the sense of Newey (1990) has this outline. We first assume pathwise differentiability of κ, see Section A for definition. Working with this assumption, we derive the semiparametric efficient score. With this score in hand, we then prove pathwise differentiability. Details are in the Appendix.

Remark 3 With a slight modification using a device introduced to semiparametric methods by Bickel (1982), Theorem 1 also holds for estimated bandwidths. We confine our discussion to bandwidths of order $n^{-1/3}$, see Section II.5.1.2 for a reason. Write such bandwidths as $h_n = cn^{-1/3}$, where following Bickel the values for c are allowed to take values in the set $\mathcal{U} = a\{0, \pm 1, \pm 2, ...\}$, where a is an arbitrary small number. We discretize bandwidths so that they take on values $cn^{-1/3}$ with $c \in \mathcal{U}$. Denote estimators as $\widehat{\kappa}(h_n)$, and note that for an arbitrary c_*, and an arbitrary fixed, deterministic sequence $c_n \to c_0$ for finite c_0, Theorem 1 shows that $n^{1/2}\{\widehat{\kappa}(c_n n^{-1/3}) - \widehat{\kappa}(c_0 n^{-1/3})\} = o_p(1)$, and that $n^{1/2}\{\widehat{\kappa}(c_0 n^{-1/3}) - \widehat{\kappa}(c_* n^{-1/3})\} = o_p(1)$. Hence, it follows from Bickel (1982, p. 653, just after equation 3.7) that if $\widehat{h}_n = \widehat{c}_n n^{-1/3}$, with $\widehat{c} \in \mathcal{U}$, is an estimated bandwidth with the property that $\widehat{h}_n = O_p(n^{-1/3})$, then $n^{1/2}\{\widehat{\kappa}(\widehat{c}_n n^{-1/3}) - \widehat{\kappa}(c_* n^{-1/3})\} = o_p(1)$. Hence, Theorem 1 holds for these estimated bandwidths.

II.2.2. General Functions of the Response and Double-Robustness

It is important to consider estimation in problems where κ_0 can be constructed outside the model. Suppose that $\kappa_0 = E\{\mathcal{G}(Y)\}$, and define $\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\} = E\{\mathcal{G}(Y)|X, Z\}$. We will discuss two estimators with the properties that (a) if there are no missing response data, the semiparametric model is not used and the estimator is consistent, and (b) under certain circumstances, the estimator is consistent if either the semiparametric model is correct or if a model for the missing-data process is correct.

Our motivating example discussed in Section II.4 dose not fall into the category discussed in this section.

The two estimators are based upon different constructions for estimating the missing data process. The first is based upon a nonparametric formulation for estimating $\text{pr}(\delta = 1|Z) = \pi_{\text{marg}}$, where the subscript indicates a marginal estimation of the probability that Y is observed. The second is based upon a parametric formulation for estimating $\text{pr}(\delta = 1|Y, X, Z) = \pi(X, Z, \zeta)$, where ζ is an unknown parameter estimated by standard logistic regression of δ on (X, Z).

The first estimator, similar to one defined by Wang et al. (2004) and efficient in the Gaussian partially linear model, can be constructed as follows. Estimate π_{marg} by local linear logistic regression of δ on Z, leading to the usual asymptotic expansion

$$\widehat{\pi}_{\text{marg}}(z) - \pi_{\text{marg}}(z) = n^{-1} \sum_{j=1}^{n} \{\delta_j - \pi_{\text{marg}}(Z_j)\} K_h(z - Z_j)/f_Z(z) + o_p(n^{-1/2}), \quad (2.11)$$

assuming that $nh^4 \to 0$. Then construct the estimator

$$\widehat{\kappa}_{\text{marg}} = n^{-1}\sum_{i=1}^{n}\left[\frac{\delta_i}{\widehat{\pi}_{\text{marg}}(Z_i)}\mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\widehat{\pi}_{\text{marg}}(Z_i)}\right\}\mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}\right].$$

The estimator has two useful properties: (a) if there are no missing data, it does not depend on the model and is hence consistent for κ_0; and (b) if observation of the response Y depends only on Z, it is consistent even if the semiparametric model is not correct.

In a similar vein, the second estimate, also similar to another estimate of Wang et al. (2004), is given as

$$\widehat{\kappa} = n^{-1}\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi(X_i, Z_i, \widehat{\zeta})}\mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \widehat{\zeta})}\right\}\mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\}\right].$$

This estimator has the double-robustness property that if either the parametric model $\pi(X, Z, \zeta)$ or the underlying semiparametric model for $\{\mathcal{B}, \theta(\cdot)\}$ are correct, then $\widehat{\kappa}$ is consistent and asymptotically normally distributed. Generally, the second terms in both $\widehat{\kappa}_{\text{marg}}$ and $\widehat{\kappa}$ improve efficiency: it is also important for the double robustness property of $\widehat{\kappa}$.

If both models are correct, then the following results obtain as a consequence of (2.5) and (2.6), see the Appendix for a sketch.

Lemma 1 Make the definitions

$$\begin{aligned}
\mathcal{M}_{2,\text{marg}} &= E\left[\left\{1 - \frac{\delta}{\pi_{\text{marg}}(Z)}\right\}\{\mathcal{F}_{\mathcal{B}}(\cdot) + \mathcal{F}_{\theta}(\cdot)\theta_{\mathcal{B}}(Z, \mathcal{B}_0)\}^{\mathrm{T}}\right];\\
D_{i,\text{marg}}(\cdot) &= -\mathcal{L}_{i\theta}(\cdot)E\left[\left\{1 - \frac{\delta_i}{\pi_{\text{marg}}(Z_i)}\right\}\mathcal{F}_{i\theta}(\cdot)|Z_i\right]/E\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z_i\}.
\end{aligned}$$

Then, to terms of order $o_p(1)$,

$$n^{1/2}(\widehat{\kappa}_{\text{marg}} - \kappa_0) \approx n^{-1/2}\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi_{\text{marg}}(Z_i)}\mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\pi_{\text{marg}}(Z_i)}\right\}\mathcal{F}_i(\cdot) - \kappa_0\right] + \mathcal{M}_{2,\text{marg}}\mathcal{M}_1^{-1}n^{-1/2}\sum_{i=1}^{n}\delta_i\epsilon_i + n^{-1/2}\sum_{i=1}^{n}\delta_i D_{i,\text{marg}}(\cdot). \quad (2.12)$$

Lemma 2 Define $\pi_\zeta(X, Z, \zeta) = \partial\pi(X, Z, \zeta)/\partial\zeta$. Assume that

$$n^{1/2}(\widehat{\zeta} - \zeta) = n^{-1/2}\sum_{i=1}^{n}\psi_{i\zeta}(\cdot) + o_p(1)$$

with $E\{\psi_\zeta(\cdot)|X, Z\} = 0$. Then, to terms of order $o_p(1)$,

$$n^{1/2}(\widehat{\kappa} - \kappa_0) \approx n^{-1/2}\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\{\mathcal{G}(Y_i) - \kappa_0\} + \left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\right\}\{\mathcal{F}_i(\cdot) - \kappa_0\}\right]. \quad (2.13)$$

Remark 4 The expansions (2.12) and (2.13) show that $\widehat{\kappa}_{\text{marg}}$ and $\widehat{\kappa}$ are asymptotically normally distributed. One can show that the asymptotic variances are given as

$$\mathcal{V}_{\kappa,\text{marg}} = \text{var}\left[\frac{\delta}{\pi_{\text{marg}}(Z)}\mathcal{G}(Y) + \left\{1 - \frac{\delta}{\pi_{\text{marg}}(Z)}\right\}\mathcal{F}(\cdot) + \mathcal{M}_{2,\text{marg}}\mathcal{M}_1^{-1}\delta\epsilon + \delta D_{\text{marg}}(\cdot)\right]$$

$$\mathcal{V}_\kappa = \text{var}\left[\frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\mathcal{G}(Y_i) + \left\{1 - \frac{\delta_i}{\pi(X_i, Z_i, \zeta)}\right\}\mathcal{F}_i(\cdot)\right],$$

respectively, from which estimates are readily derived.

Finally, we note that Claeskens and Carroll (2007) show that in general likelihood problems, if there is an omitted covariate, then under contiguous alternatives the effect on estimators is to add an asymptotic bias, without changing the asymptotic variance.

II.3. Single Index Models

One means of dimension reduction is single index modeling. Single index models can be viewed as a generalized version of projection pursuit, in that only the most influential direction is retained to keep the model tractable and to reduce dimension. Since its introduction in Härdle and Stoker (1989), it has been widely studied and used. A comprehensive summary of the model is given in Härdle, Müller, Sperlich and Werwatz (2004). Let $Z = (R, S^{\mathrm{T}})^{\mathrm{T}}$ where R is a scalar. We consider here the generalized partially linear single index model (GPLSIM) of Carroll et al. (1997), namely the exponential family (2.20) with $\eta(X, Z) = X^{\mathrm{T}}\beta_0 + \theta_0(Z^{\mathrm{T}}\alpha_0)$, where $\theta_0(\cdot)$ is an unknown function, and for identifiability purposes $\|\alpha_0\| = 1$. Since identifiability requires that one of the components of Z be a non-trivial predictor of Y, for convenience we will make the very small modification that one component of Z, what we call R, is a known non-trivial predictor of Y. The reason for making this modification can be seen in Theorem 4 of Carroll et al. (1997) where the final limit distribution of the estimate of α_0 has a singular covariance matrix. In addition, their main asymptotic expansion, given in their equation (A.12), is about the nonsingular transformation $(I - \alpha_0\alpha_0^{\mathrm{T}})(\widehat{\alpha} - \alpha_0)$.

With this modification, we write the model as

$$E(Y|X, Z) = \mathcal{C}^{(1)}[c\{\eta(X, Z)\}] = \mu\{X^{\mathrm{T}}\beta_0 + \theta_0(R + S^{\mathrm{T}}\gamma_0)\}, \tag{2.14}$$

where γ_0 is unrestricted.

Carroll et al. (1997) use profile likelihood to estimate $\mathcal{B}_0 = (\gamma_0, \beta_0)$ and $\theta_0(\cdot)$, although they present no results concerning the estimate of ϕ_0, their interest largely being in

logistic regression where $\phi_0 = 1$ is known. Rewrite the likelihood function (2.20) as $L\{Y, X, \beta, \theta(R + S^{\mathrm{T}}\gamma), \phi\}$. Then, given $\mathcal{B} = (\gamma^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$, they form $U(\gamma) = R + S^{\mathrm{T}}\gamma$ and then compute the estimate $\widehat{\theta}\{u(\gamma), \mathcal{B}\}$ by local likelihood of Y on $\{X, U(\gamma)\}$ as in Severini and Staniswalis (1994), using the data with $\delta = 1$. Then they maximize $\sum_{i=1}^{n} \delta_i \log[L\{Y_i, X_i, \beta, \widehat{\theta}(R_i + S_i^{\mathrm{T}}\gamma, \mathcal{B}), \phi\}]$ in $\mathcal{B}$ and ϕ.

Our goal is to estimate $\kappa_{\mathrm{SI}} = E[\mathcal{F}\{X, \theta_0(R + S^{\mathrm{T}}\gamma_0), \beta_0, \phi_0\}]$. Our proposed estimate is $\widehat{\kappa}_{\mathrm{SI}} = n^{-1}\sum_{i=1}^{n} \mathcal{F}\{X_i, \widehat{\theta}(R_i + S_i^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}}), \widehat{\beta}, \widehat{\phi}\}$.

Our main result is as follows. First define $U = R + S^{\mathrm{T}}\gamma_0$, and

$$\mathcal{G} = \mathcal{D}_{\phi}(Y, \phi_0) - [Yc\{X^{\mathrm{T}}\beta_0 + \theta_0(U)\} - \mathcal{C}\{c(\cdot)\}]/\phi_0^2.$$

Make the further definitions $\Lambda = \{S^{\mathrm{T}}\theta_0^{(1)}(U), X^{\mathrm{T}}\}^{\mathrm{T}}$, $\rho_{\ell}(\cdot) = \{\mu^{(1)}(\cdot)\}^{\ell}/V(\cdot)$ and $\epsilon = [Y - \mu\{X^{\mathrm{T}}\beta_0 + \theta_0(U)\}]\rho_1\{X^{\mathrm{T}}\beta_0 + \theta_0(U)\}$. Define

$$\mathcal{N}_i = \Lambda_i - [E\{\delta\rho_2(\cdot)|U_i\}]^{-1}E\{\delta_i\Lambda_i\rho_2(\cdot)|U_i\}$$

and $\mathcal{Q} = E\{\delta\mathcal{N}\mathcal{N}^{\mathrm{T}}\rho_2(\cdot)\}$. Make further definitions $\mathcal{F}_{\beta}(\cdot) = \partial\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\}/\partial\beta_0$, $\mathcal{F}_{\phi}(\cdot) = \partial\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\}/\partial\phi_0$ and $\mathcal{F}_{\theta}(\cdot) = \partial\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\}/\partial\theta_0(U)$. Also define

$$\begin{aligned} J(U) &= [E\{\delta\rho_2(\cdot)|U\}]^{-1}E\{\mathcal{F}_{\theta}(\cdot)|U\}; \\ D &= \begin{bmatrix} E\{\mathcal{F}_{\theta}(\cdot)\theta^{(1)}(U)S\} - E\Big(\mathcal{F}_{\theta}(\cdot)[E\{\delta\rho_2(\cdot)|U\}]^{-1}\theta^{(1)}(U)E\{\delta S\rho_2(\cdot)|U\}\Big) \\ E\{\mathcal{F}_{\beta}(\cdot)\} - E\Big(\mathcal{F}_{\theta}(\cdot)[E\{\delta\rho_2(\cdot)|U\}]^{-1}E\{\delta X\rho_2(\cdot)|U\}\Big) \end{bmatrix}. \end{aligned}$$

Then we have the following result regarding the asymptotic distribution of $\widehat{\kappa}_{\mathrm{SI}}$:

Result 2 Assume that $(Y_i, \delta_i, X_i, Z_i), i = 1, 2, ..., n$ are i.i.d and that the conditions

in Carroll et al. (1997) hold, in particular that $nh^4 \to 0$. Then

$$\begin{aligned} & n^{1/2}(\widehat{\kappa}_{\mathrm{SI}} - \kappa_{\mathrm{SI}}) \\ & \quad = n^{-1/2} \sum_{i=1}^{n} \Big[\mathcal{F}\{X_i, \theta_0(U_i), \beta_0, \phi_0\} - \kappa_{\mathrm{SI}} + D^{\mathrm{T}} \mathcal{Q}^{-1} \delta_i \mathcal{N}_i \epsilon_i + \delta_i J(U_i) \epsilon_i \\ & \qquad\qquad + \delta_i \mathcal{G}_i E\{\mathcal{F}_\phi(\cdot)\} / E(\delta \mathcal{G}^2) \Big] + o_p(1) \qquad (2.15) \\ & \quad \Rightarrow \mathrm{Normal}(0, \mathcal{V}), \end{aligned}$$

where

$$\begin{aligned} \mathcal{V} \;\; = \;\; & E[\mathcal{F}\{X, \theta_0(U), \beta_0, \phi_0\} - \kappa_{\mathrm{SI}}]^2 + D^{\mathrm{T}} \mathcal{Q}^{-1} D + \mathrm{var}\{\delta J(U) \epsilon\} \\ & + E(\delta \mathcal{G}^2)[E\{\mathcal{F}_\phi(\cdot)\}]^2 / \{E(\delta \mathcal{G}^2)\}^2. \end{aligned}$$

Further, $\widehat{\kappa}_{\mathrm{SI}}$ is semiparametric efficient.

II.4. Motivating Example

II.4.1. Introduction

There is considerable interest in understanding the distribution of dietary intake in various populations. For example, as obesity rates continue to rise in the United States (Flegal, Carroll, Ogden and Johnson 2002), the demand for information about diet and nutrition is increasing. Information on dietary intake has implications for establishing population norms, research, and making public policy decisions (Woteki 2003).

We wish to emphasize that there are no missing response data in this example. We also emphasize that the problem is vastly different from simply estimating the population mean using a Gaussian partially linear model. The strength of our approach is

that once one has proposed a semiparametric model, then our methodology, asymptotics and semiparametric efficiency results are readily employed.

This work was motivated by the analysis of the Eating at America's Table (EATS) study (Subar et al. 2001), where estimating the distribution of the consumption of episodically consumed foods is of interest. The data consist of 4 24hr recalls over the course of a year as well as the National Cancer Institute's (NCI) dietary history questionnaire (DHQ), a particular version of a food frequency questionnaire (FFQ, see Willett et al. 1985 and Block et al. 1986). The goal is to estimate the distribution of usual intake, defined as the average daily intake of a dietary component by an individual in a fixed time period, a year in the case of EATS. There were $n = 886$ individuals in the data set.

When the responses are continuous random variables, this is a classical problem of measurement error, with a large literature. However, little of the literature is relevant to episodically consumed foods, as we now describe. Consider, for example, consumption of red meat, dark green vegetables and deep yellow vegetables, all of interest in nutritional surveillance. In the EATS data, 45% of the 24-hour recalls reported no red meat consumption. In addition, 5.5% of the individuals reported no red meat consumption on any of the four separate 24-hour recalls: for deep yellow vegetables these numbers are 63% and 20%, respectively, while for dark green vegetables the numbers are 78% and 46%, respectively. Clearly, methods aimed at understanding usual intakes for continuous data are inappropriate for episodically consumed foods with so many zero-reported intakes.

II.4.2. Model

To handle episodically consumed foods, two-part models have been developed (Tooze, Grunwald and Jones 2002). These are basically zero-inflated repeated measures examples. Our methods are applicable to such problems when the covariate Z is evaluated only once for each subject, as it is in our example.

We describe here a simplification of this approach, used to illustrate our methodology. On each individual, we measure age and gender, the collection being what we call R. We also observe energy (calories) as measured by the DHQ, the logarithm of which we call Z. The reader should note that Z is evaluated only once per individual, and hence while there are repeated measures on the responses, there are no repeated measures on Z: $\theta_0(Z)$ occurs only once in the likelihood function, and our methodology applies.

Let $X = (R, Z)$. The response data for an individual i consists of four 24-hour recalls of red meat consumption. Let $\Delta_{ij} = 1$ if red meat is reported consumed on the j^{th} 24-hour recall for $j = 1, ..., 4$. Let $\mathcal{Y}_{ij}$ be the product of Δ_{ij} and the logarithm of reported red meat consumption, with the convention that $0 \log(0) = 0$. Then the response data are $Y_i = (\Delta_{ij}, \mathcal{Y}_{ij})_{j=1}^4$.

II.4.2.1. Modeling the Probability of Zero Response

The first part of the model is whether the subject reports red meat consumption. We model this as a repeated measures logistic regression, so that

$$\text{pr}(\Delta_{ij} = 1 | R_i, Z_i, U_{i1}) = H(\beta_0 + X_i^{\text{T}} \beta_1 + U_{i1}), \tag{2.16}$$

where $H(\cdot)$ is the logistic distribution function and $U_{i1} = \text{Normal}(0, \sigma^2_{u1})$ is a person-specific random effect. Note that for simplicity we have modeled the effect of energy consumption as linear, since in the data there is little hint of nonlinearity.

II.4.2.2. Modeling Positive Responses

The second part of the model consists of a distribution of the logarithm of red meat consumption on days when consumption is reported, namely

$$[\mathcal{Y}_{ij}|\Delta_{ij} = 1, R_i, Z_i, U_{i2}] = \text{Normal}\{R_i^{\mathrm{T}}\beta_2 + \theta(Z_i) + U_{i2}, \sigma^2\}, \tag{2.17}$$

where $U_{i2} = \text{Normal}(0, \sigma^2_{u2})$ is a person-specific random effect which we take to be independent of U_{i1}. Note that (2.17) means that the non-zero $\mathcal{Y}$-data within an individual marginally have the same mean $R_i^{\mathrm{T}}\beta_2 + \theta(Z_i)$, variance $\sigma^2 + \sigma^2_{u2}$ and common covariance σ^2_{u2}.

II.4.2.3. Likelihood Function

The collection of parameters is $\mathcal{B}$, consisting of β_0, β_1, β_2, σ^2_{u1}, σ^2_{u2}, and σ^2. The loglikelihood function $\mathcal{L}(\cdot)$ is readily computed with numerical integration, as follows:

$$\begin{aligned}\exp\{\mathcal{L}(\cdot)\} \;=\; & \frac{1}{\sigma_{u1}} \int \phi(u_1/\sigma_{u1}) \prod_{j=1}^{4} [\{H(\beta_0 + X^{\mathrm{T}}\beta_1 + u_1)\}^{\Delta_{ij}} \\ & \qquad \times \{1 - H(\beta_0 + X^{\mathrm{T}}\beta_1 + u_1)\}^{1-\Delta_{ij}}]\, du_1 \\ & \times \frac{1}{\sigma_{u2}\sigma^{\Delta_{i\cdot}}} \int \phi\Big(\frac{u_2}{\sigma_{u2}}\Big) \prod_{j=1}^{4} \bigg(\phi\Big[\frac{\mathcal{Y}_{ij} - \{R_i^{\mathrm{T}}\beta_2 + \theta(Z_i) + u_2\}}{\sigma}\Big]\bigg)^{\Delta_{ij}} du_2,\end{aligned}$$

where $\Delta_{i\cdot} = \sum_j \Delta_{ij}$. Of course, the second numerical integral is not necessary, since the integration can be done analytically.

II.4.2.4. Defining Usual Intake at the Individual Level

Noting from (2.17) that reported intake on days of consumption follows a lognormal distribution, the usual intake for an individual is defined as

$$G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\} = H(\beta_0 + X_i^{\mathrm{T}}\beta_1 + U_1)\exp\{R^{\mathrm{T}}\beta_2 + \theta(Z) + U_2 + \sigma^2/2\}. \quad (2.18)$$

The goal is to understand the distribution of $G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\}$ across a population. In particular, for arbitrary c we wish to estimate $\mathrm{pr}[G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\} > c]$. Define $\mathcal{F}\{X, \mathcal{B}, \theta(Z)\} = \mathrm{pr}[G\{X, U_1, U_2, \mathcal{B}, \theta(Z)\} > c|X, Z]$, a quantity that can be computed by numerical integration. Then $\kappa_0 = E[\mathcal{F}\{X, \mathcal{B}, \theta(Z)\}]$ is the percentage of the population whose long-term reported daily average consumption of red meat exceeds c.

II.4.3. Bias in Naive Estimates, and a Simulation Study

We emphasize that the distribution of mean intake cannot be estimated consistently by the simple device of computing the sample percentage of the observed 24-hour recalls that exceed c, and, as a consequence, going through the model fitting process is actually necessary. To see this, suppose only one 24-hour recall were computed and the percentage of these 24-hour recalls exceeding c is computed. In large samples, this percentage converges to

$$\kappa_{24\mathrm{hr}} = E\Big(H(\beta_0 + X^{\mathrm{T}}\beta_1 + U_1)\Phi\Big[\{R^{\mathrm{T}}\beta_2 + \theta(Z) - \log(c)\}/(\sigma^2 + \sigma_2^2)^{1/2}\Big]\Big).$$

In contrast, for $\sigma_2 > 0$,

$$\kappa_0 = E\Big\{\Phi\Big([R^{\mathrm{T}}\beta_2 + \theta(Z) + \sigma^2/2 - \log\{c/H(\beta_0 + X^{\mathrm{T}}\beta_1 + U_1)\}]/\sigma_2\Big)\Big\}.$$

As the number of replicates m of the 24-hour recall $\to \infty$, the percentage $\kappa_{m,24\text{hr}}$ of the means of the 24-hour recalls that exceed $c \to \kappa_0$, so we would expect that the fewer the replicates, the less our estimate agrees with the sample version of $\kappa_{m,24\text{hr}}$, a phenomenon observed in our data, see below.

To see this numerically, we ran the following simulation study. Gender, age and the DHQ were kept the same as in the EATS Study. The parameters $(\beta_0, \beta_1, \beta_2, \sigma^2, \sigma_1^2, \sigma_2^2)$ were the same as our estimated values, see below. The function $\theta(\cdot)$ was roughly in accord with our estimated function, for simplicity being quadratic in the logarithm of the DHQ, standardized to have minimum 0.0 and maximum 1.0, with intercept, slope and quadratic parameters being 0.50, 1.50 and -0.75, respectively. The true survival function, i.e., 1 - the cdf, was computed analytically, while the survival functions for the mean of two 24-hour recalls and the mean of four 24-hour recalls were computed by $1,000$ simulated data sets. The results are given in Figure 1, where the bias from not using a model is evident.

We used our methods with a nonparametrically estimated function, a bandwidth $h = 0.30$ and the Epanechnikov kernel function. We generated 300 data sets, with results displayed in Figure 2. The mean over the simulation was almost exactly the correct function, not surprising given that the sample size is large ($n = 886$). In Figure 2 we also display a 90% confidence range from the simulated data sets, indicating that in the EATS data at least, the results of our approach are relatively accurate.

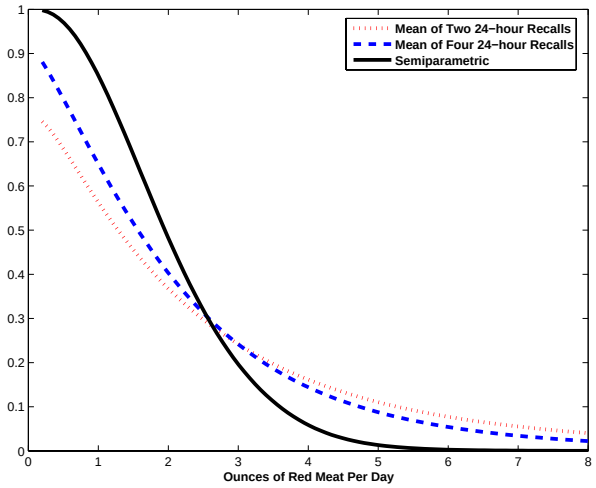

Fig. 1 Results of the simulation study meant to mimic the EATS Study. All results are averages over $1,000$ simulated data sets. Solid line: the mean of the semiparametric estimator of the survival curve, which is almost identical to the true survival curve. Dotted line: the empirical survival function of the mean of two 24-hour recalls from 1,000 simulated data sets. Dashed line: the empirical survival function of the mean of four 24-hour recalls from 1,000 simulated data sets.

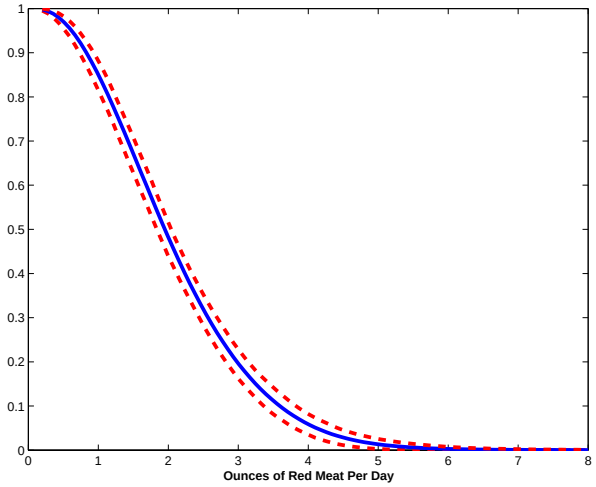

Fig. 2 Results of the simulation study meant to mimic the EATS Study. Plotted is the mean survival function for 300 simulated data sets, along with the 90% pointwise confidence intervals. The mean fitted function is almost exact.

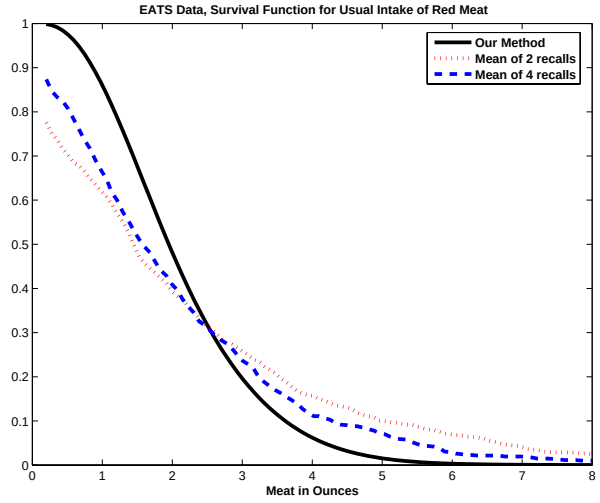

Fig. 3 Results from the EATS Example. Plotted are estimates of the survival function (1 - the cdf) of usual intake of red meat. The solid line is the semiparametric method described in Section II.4. The dotted line is the empirical survival function of the mean of the first two 24-hour recalls per person, while the dashed line is survival function of the mean of all the 24-hour recalls per person.

II.4.4. Data Analysis

We standardized age to have mean zero and variance one. In the logistic part of the model, the intercept was estimated as -8.15, with the coefficients for (gender,age,DHQ) $= (0.13, 0.14, 1.09)$. The random effect variance was estimated as $\widehat{\sigma}_1^2 = 0.66$. In the continuous part of the model, we used bandwidths ranging from 0.05 to 0.40, with little change in any of the estimates, as described in more detail in Section II.5.1. With a bandwidth $h = 0.30$, our estimates were $\widehat{\sigma}^2 = 0.76$, $\widehat{\sigma}_2^2 = 0.043$, and the coefficients for gender and age were -0.25 and 0.02, respectively. The coefficient for the person specific random effect σ_2^2 appears intrinsic to the data: we used other methods such as mixed models with polynomial fits and obtained roughly the same answers.

We display the computed survival function in Figure 3. Displayed there are our method, along with the empirical survival functions for the mean of the first two 24-hour recalls and the mean of all four 24-hour recalls. While these are biased, it is interesting to note that using the mean of only two 24-hour recalls is more different from our method than using the mean of four 24-hour recalls, which is expected as described above. The similarity of Figures 1 and 3 is striking, mainly indicating that naive approaches, such as using the mean of two 24-hour recalls, can result in badly biased estimates of κ_0.

II.5. Bandwidth Selection, the Partially Linear Model, and the Sample Mean

II.5.1. Bandwidth Selection

II.5.1.1. Background

We have used a standard first-order kernel density function, i.e., one with mean zero and positive variance. With this choice, in Theorem 1 we have assumed that the bandwidth satisfies $nh^4 \to 0$: for estimation of the population mean in the partially linear model. In contrast, if one were interested only in $\mathcal{B}_0$, then it is well-known that by using profile likelihood the usual bandwidth order $h \sim n^{-1/5}$ is acceptable, and off-the-shelf bandwidth selection techniques yield an asymptotically normal limit distribution.

The reason for the technical need for undersmoothing is the inclusion of $\theta_0(\cdot)$ in κ_0. For example, suppose that $\kappa_0 = E\{\theta_0(Z)\}$. Then it follows from (2.5) that $\widehat{\kappa} - \kappa_0 = O_p(h^2 + n^{-1/2})$. Thus, in order for $n^{1/2}(\widehat{\kappa} - \kappa_0) = O_p(1)$, we require that $nh^4 = O_p(1)$. The additional restriction that $nh^4 \to 0$ merely removes the bias term

entirely.

Note that κ_0 is not a parameter in the model, being a mixture of the parametric part $\mathcal{B}_0$, the nonparametric part $\theta_0(\cdot)$, and the joint distribution of (X, Z). Thus, it does not appear that κ_0 can be estimated by profiling ideas.

II.5.1.2. Optimal Estimation

As seen in Theorem 1, the asymptotic distribution of $n^{1/2}(\widehat{\kappa} - \kappa_0)$ is unaffected by the bandwidth, at least to first order. In Section II.5.1.3 below we give intuitive and numerical evidence of the lack of sensitivity to the bandwidth choice, see also Section II.5.3 for further numerical evidence. In Section II.5.1.4 we describe three different, simple practical methods for bandwidth selection in this problem, all of which work quite well in our simulations and example.

Since first-order calculations do not get squarely at the choice of bandwidth, other than to suggest that it is not particularly crucial, an alternative theoretical device is to do second order calculations. Define $\eta(n, h) = n^{1/2}h^2 + (n^{1/2}h)^{-1}$. In a problem similar to ours, Sepanski, Knickerbocker and Carroll (1994) show that the variance of linear combinations of the estimate of $\mathcal{B}_0$ has a second order expansion as follows. Suppose we want to estimate $\xi^{\mathrm{T}}\mathcal{B}_0$. Then, for constants (a_1, a_2),

$$\begin{aligned} n^{1/2}(\xi^{\mathrm{T}}\widehat{\mathcal{B}} - \xi^{\mathrm{T}}\mathcal{B}_0) &= V_n + o_p\{\eta(n, h)\}; \\ \mathrm{cov}(V_n) &= \text{constant} + \left\{a_1 n^{1/2}h^2 + a_2(hn^{1/2})^{-1}\right\}^2. \end{aligned}$$

This means that the optimal bandwidth is of the order $h = cn^{-1/3}$ for a constant c depending on (a_1, a_2), which in turn depend on the problem, i.e., on the distribution

of (Y, X, Z) as well as $\mathcal{B}_0$ and $\theta_0(\cdot)$. In their practical implementation, translated from the Gaussian kernel function to our Epanechnikov kernel function, Sepanski et al. (1994) suggest the following device, namely that if the optimal bandwidth for estimating $\theta_0(\cdot)$ is $h_o = cn^{-1/5}$, then they use the correct-order bandwidth $h = cn^{-1/3}$. They also did sensitivity analysis, e.g., $h = (1/2)cn^{-1/3}$, but found little change in their simulations. One of our three methods of practical bandwidth selection is exactly this one.

A problem not within our framework but carrying a similar flavor was considered by Powell and Stoker (1996) and Newey, Hsieh and Robins (2004), namely the estimation of the weighted average derivative $\kappa_{\text{AD}} = E\{Y\theta_0^{(1)}(Z)\}$. As done by Sepanski et al. (1994), Powell and Stoker (1996) show that the optimal bandwidth constructed from second-order calculations is an undersmoothed bandwidth. Newey et al. (2004) suggest that a simple device of choosing the bandwidth is to choose something optimal when using a standard second-order kernel function but to then undersmooth, in effect, by using a higher-order kernel such as the twicing kernel. This is our second bandwidth selection method described in Section II.5.1.4. Like the first, it appears to be an effective means of eliminating the bias term.

In our problem, the paper by Sepanski et al. (1994) is more relevant. Preliminary calculations based upon the basic tools in that paper suggest that for our problem, the optimal bandwidth is also of order $n^{-1/3}$. We intend to pursue these very calculations in another paper.

II.5.1.3. Lack of Sensitivity to Bandwidth

We have used the term *technical need for undersmoothing* because that is what it really is. In practice, as Theorem 1 states, the asymptotic distribution of $\widehat{\kappa}$ is unaffected by bandwidth choice for very broad ranges of bandwidths. This is totally different from what happens with estimation of the function $\theta_0(\cdot)$, where bandwidth selection is typically critical in practice, and this is seen in theory through the usual bias-variance tradeoff.

In practice, we expect little effect of the bandwidth selection on estimation of $\mathcal{B}_0$, and even less effect on estimation of κ_0. The reason is that broad ranges of bandwidths lead to no asymptotic effect on the distribution of $\widehat{\mathcal{B}}$. The extra amount of smoothing inherent in the summation in (2.4) should mean that $\widehat{\kappa}$ will be even less sensitive to the bandwidth, the so-called *double-smoothing* phenomenon.

To see this issue, consider the simulation in Wang et al. (2004). They set X and Z to be independent, with $X = \text{Normal}(1, 1)$ and $Z = \text{Uniform}[0, 1]$. In the partially linear model, they set $\mathcal{B}_0 = 1.5$, $\epsilon = \text{Normal}(0, 1)$ and $\theta_0(z) = 3.2z^2 - 1$. They used the kernel function $(15/16)(1 - z^2)^2 I(|z| \leq 1)$, and they fixed the bandwidth to be $h = n^{-2/3}$, which at least asymptotically is very great undersmoothing, since $h \sim n^{-1/3}$ is already acceptable and typically something like $nh^2/\log(n) \to \infty$ is usually required. In their Case 3, they used effective sample sizes for complete data of 18, 36 and 60, with corresponding bandwidths 0.146, 0.092 and 0.065, respectively.

We reran the simulation of Wang et al. (2004), with complete response data and $n = 60$. We used bandwidths 0.02, 0.06, 0.10, 0.14, ranging from a very small band-

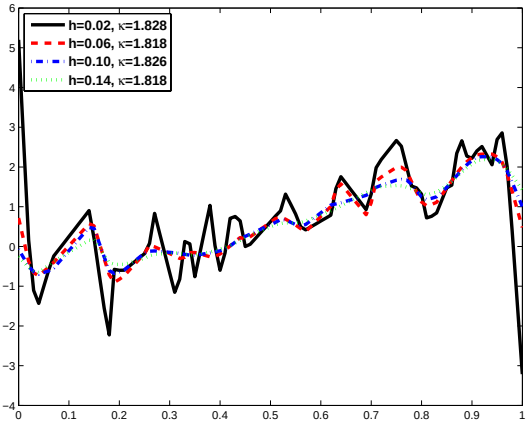

Fig. 4 Results for a single data set in a simulation as in Wang et al. (2004), the partially linear model with $n = 60$, complete response data, and when $\kappa_0 = E(Y)$. Various bandwidths are used, and the estimates of the function $\theta_0(\cdot)$ are displayed. In the legend, the actual estimates of κ_0 are displayed. Note how the bandwidth has a major impact on the function estimate with the bandwidth is too small ($h = 0.02$), but very little effect on the estimate of κ_0.

width, less than 1/3 that used by Wang et al. (2004), to a larger bandwidth, more than double that used. As another perspective, if one sets $h = \sigma_z n^{-c}$, where σ_z is the standard deviation of Z, then the bandwidths used are equivalent to $c = 0.73$, 0.46, 0.34 and 0.26, In other words, a bandwidth here of $h = 0.02$ is very great under-smoothing, while even $h = 0.14$ satisfies the theoretical constraint on the bandwidth.

In Figure 4, we plot the results for a single data set, where, as in Wang et al. (2004), interest lies in estimating $\kappa_0 = E(Y)$. As is obvious from this figure, the bandwidth choice is very important for estimation of the function, but trivially unimportant for estimation of κ_0, the estimate of which ranged from 1.818 to 1.828.

In Figure 5, we plot the mean estimated functions from 100 simulated data sets. Again the bandwidth matters a great deal for estimating the function $\theta_0(\cdot)$. Again

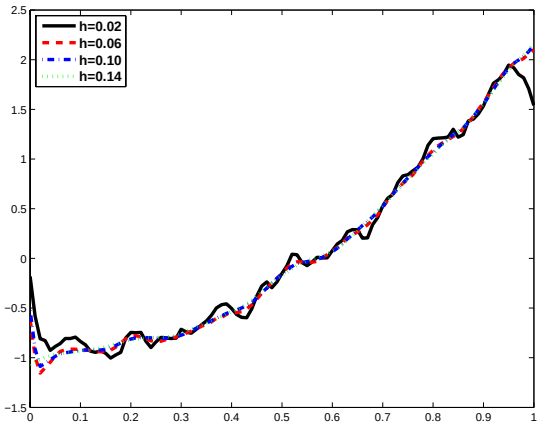

Fig. 5 Results for 100 simulated data sets in a simulation as in Wang et al. (2004), the partially linear model with $n = 60$ and complete response data. Various bandwidths are used, and the mean estimates of the function $\theta_0(\cdot)$ are displayed. Note how even over these simulations, the bandwidth has a clear impact on the function estimate: there is almost no impact on estimates of the population mean and variance.

too, the bandwidth matters hardly at all for estimating κ_0. Thus, for estimating κ_0, the mean estimates across the bandwidths range from 1.513 to 1.526, and the standard deviations of the estimates range from 0.249 to 0.252. There is somewhat more effect of bandwidth on the estimate of $\mathcal{B}_0$: for $h \geq 0.06$, there is almost no effect, but choosing $h = 0.02$ results in a 50% increase in standard deviation.

In other words, as expected by theory and intuition, bandwidth selection has little effect on the estimate of $\mathcal{B}_0$ except when the bandwidth is much too small, and very little effect on the estimation of $\kappa_0 = E(Y)$. Similar remarks occur when one looks at the variance of the errors as the parameter, and κ_0 is the population variance.

II.5.1.4. Bandwidth Selection

As described above in Section II.5.1.3, bandwidth selection is not a vital issue for estimating κ_0: of course, it is vital for estimating $\theta_0(\cdot)$. Effectively, what this means is that the real need is simply to get bandwidths that satisfy the technical assumption of undersmoothing but are not too ridiculously small: a precise target is often unnecessary. In addition, because the asymptotic distribution of $\widehat{\kappa}$ does not depend on the bandwidth, simple first-order methods of the type that are used in bandwidth selection for function estimation are not possible. Thus, in our example, we used three different methods, all of which gave answers that were as nearly identical as in the simulation of Wang et al. (2004).

All the methods are based on a so-called "typical device" to get an optimal bandwidth for estimating θ_0, of the form $h_{\text{opt}} = c\sigma_z n^{-1/5}$. In practice, this can be accomplished by constructing a finite grid of bandwidths of the form $h_{\text{grid}} = c_{\text{grid}}\sigma_z n^{-1/5}$: we use a grid from 0.20 to 5.0. After estimating $\mathcal{B}_0$ by $\widehat{\mathcal{B}}(h_{\text{grid}})$, this value is fixed, and then a loglikelihood cross-validation score obtained. The maximizer of the loglikelihood crossvalidation score is selected as h_{opt}.

- If $h_{\text{opt}} = c\sigma_z n^{-1/5}$, an extremely simple device is simply to set $h = h_{\text{opt}} n^{-2/15} = c\sigma_z n^{-1/3}$, which satisfies the technical condition of undersmoothing without becoming ridiculously too small. This device may seem terribly ad hoc, but the theory, the simulation of Wang et al. (2004), the discussion in Section II.5.1.3, and our own work suggests that this method actually works reasonably well. Note too that in Section II.5.1.2 we give evidence that this bandwidth rate is most likely optimal.

- A second approach is taken by Newey et al. (2004), and is also an effective

practical device. The technical need for undersmoothing comes from the fact that the bias term in a first-order local likelihood kernel regression is of order $O(h^2)$. One can use higher-order kernels to get the bias to be of order $O(h^{2s})$ for $s \geq 2$, but this does not really help in that the variance remains of order $O\{(nh)^{-1}\}$, so that the optimal mean squared error kernel estimator has $h = O\{n^{-1/(4s+1)}\}$, and thus undersmoothing to estimate κ_0 is still required. However, as Newey et al. (2004) point out, if one uses the optimal bandwidth $h_{\text{opt}} = c\sigma_z n^{-1/5}$, but then does the estimation procedure replacing the first-order kernel by a higher order kernel, then the bias is $O(h_{\text{opt}}^{2s}) = o(n^{-1/2})$ if $s \geq 2$. A convenient higher-order kernel is the second-order twicing kernel $K_{\text{tw}}(u) = 2K(u) - \int K(u-v)K(v)dv$, where $K(\cdot)$ is a first-order kernel.

- One can also use loglikelihood crossvalidation, but with the grid of values being of the form $h_{\text{grid}} = c_{\text{grid}}\sigma_z n^{-1/3}$. Because crossvalidation scores often have multiple modes, this is not the same as optimal smoothing.

It may be worth pointing out again that Wang et al. (2004) set $h = n^{-2/3}$, and even then, with too much undersmoothing (asymptotically), the performance of the method is rather good.

II.5.2. Efficiency and Robustness of the Sample Mean

In general problems with complete data, with no assumptions about the response Y other than that it has a second moment, the sample mean $\overline{Y}$ is semiparametric efficient for estimating the population mean $\kappa_0 = E(Y)$, see for example Newey (1990). Somewhat remarkably, Wang et al. (2004) show that in the partially linear model with Gaussian errors, with complete data the sample mean is still semiparametric ef-

ficient. This fact is crucial of course in establishing that with missing response data, their estimators are still semiparametric efficient.

It is clear that with complete data, the sample mean will not be semiparametric efficient for all semiparametric likelihood models. Simple counter-examples abound, e.g., the partially linear model for Laplace or t-errors. More complex examples can be constructed, e.g., the partially linear model in the Gamma family with loglinear mean $\exp\{X^{\mathrm{T}}\mathcal{B}_0 + \theta_0(Z)\}$: details follow from Lemma 4 below.

The model-robustness of the sample mean for estimating the population mean in complete data is nonetheless a powerful feature. It is therefore of considerable interest to know whether there are cases of semiparametric likelihood problems where the sample mean is still semiparametric efficient, and thus would be used because of its model-robustness. It turns out that such cases exist. In particular, the sample mean for complete response data is semiparametric efficient in canonical exponential families with partially linear form.

Lemma 3 Recall that ϵ is defined in (2.8). If there are no missing data, the sample mean is a semiparametric efficient estimator of the population mean only if

$$Y - E(Y|X,Z) = E(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\epsilon + \mathcal{L}_\theta(\cdot)\frac{E\{Y\mathcal{L}_\theta(\cdot)|Z\}}{E\{\mathcal{L}_\theta^2(\cdot)|Z\}}. \tag{2.19}$$

It is interesting to consider (2.19) in the special case of exponential families with likelihood function

$$f(y|x,z) \;=\; \exp\Big[\frac{yc\{\eta(x,z)\} - \mathcal{C}[c\{\eta(x,z)\}]}{\phi} + \mathcal{D}(y,\phi)\Big], \tag{2.20}$$

where $\eta(x,z) = x^{\mathrm{T}}\beta_0 + \theta_0(z)$, so that $E(Y|X,Z) = \mathcal{C}^{(1)}[c\{\eta(X,Z)\}] = \mu\{\eta(X,Z)\} =$

$\mu(X, Z)$ and $\text{var}(Y|X, Z) = \phi\mathcal{C}^{(2)}[c\{\eta(X, Z)\}] = \phi V[\mu\{\eta(X, Z)\}]$.

As it turns out, effectively, (2.19) holds and the sample mean is semiparametric efficient only in the canonical exponential family for which $c(t) = t$. More precisely, we show in Appendix A the following result.

Lemma 4 If there are no missing data, under the exponential model (2.20), the sample mean is a semiparametric efficient estimate of the population mean if $\partial c\{X^{\mathrm{T}}\beta + \theta(Z)\}/\partial\theta(Z)$ is a function only of Z for all β, e.g., the canonical exponential family. Otherwise, the sample mean is generally not semiparametric efficient: the precise condition is given in equation (A.29) in the appendix. In particular, outside the canonical exponential family, the only possibility for the sample mean to be semiparametric efficient is that if for some known (a, b), $c\{x^{\mathrm{T}}\beta + \theta(z)\} = a + b\log\{x^{\mathrm{T}}\beta + \theta(z)\}$.

Remark 5 We consider Lemmas 3-4 to be positive results, although an earlier version of the work had a misplaced emphasis. Effectively, we have characterized the cases, with complete data, that the sample mean is both model-free and semiparametric efficient. In these cases, one would use the sample mean, or perhaps a robust version of it, rather than fit a potentially complex semiparametric model that can do no better, and if that model is incorrect can incur non-trivial bias.

II.5.3. Numerical Experience and Theoretical Insights in the Partially Linear Model, and Some Tentative Conclusions

In responding to a referee about the estimation of the population mean in the partially linear model (2.1), we collect here a few remarks based upon our numerical experience. Since the problem of estimating the population mean is the problem focused upon by Chen et al. (2004), we focus on the simulation set-up in their paper, although

some of the conclusions we reach may be supportable in general cases. To remind the reader, in their simulation, X and Z are independent, with $X = \text{Normal}(0, 1)$, $Z = \text{Uniform}[0, 1]$, $\beta = 1.5$, $\theta(z) = 3.2z^2 - 1$ and $\epsilon = \text{Normal}(0, 1)$.

II.5.3.1. Can Semiparametric Methods Improve Upon the Sample Mean?

When there are missing response data, the simulations in Wang et al. (2004) show conclusively that substantial gains in efficiency can be made over using the sample mean of the observed responses alone. In addition, if missingness depends on (X, Z), the sample mean of the observed responses will be biased.

This leaves the issue of what happens when there are no missing data. Obviously, if one thought that ϵ were normally distributed, it would be delusional to use anything other than the sample mean, it being efficient.

Theoretically, some insight can be gained by the following considerations. Suppose that X and Z are independent. Suppose also that ϵ has a symmetric density function known up to a scale parameter. Let σ_ϵ^2 be the variance of ϵ, and let $\zeta \leq \sigma_\epsilon^2$ be the inverse of the Fisher information for estimating the mean in the model $Y = \mu + \epsilon$. Then, it can be shown that $E\{\mathcal{F}_{\mathcal{B}}(\cdot)\} = 0$, that $\theta_{\mathcal{B}}(z, \mathcal{B}) = 0$, and that the asymptotic mean squared error efficiency (MSE) of the semiparametric efficient estimate of the population mean compared to the sample mean is

$$\text{MSE Efficiency of Sample Mean} = \frac{\beta^2 \text{var}(X) + \text{var}\{\theta(Z)\} + \zeta}{\beta^2 \text{var}(X) + \text{var}\{\theta(Z)\} + \sigma_\epsilon^2} \leq 1.$$

Note that there are cases that ζ/σ_ϵ^2 may be quite small, especially when ϵ is heavy tailed, so that if $\beta = 0$ and $\theta(\cdot)$ is approximately constant, the MSE efficiency of

the sample mean would be ζ/σ_ϵ^2, and then substantial gains in efficiency would be gained. However, the usual motivation for fitting semiparametric models is that the regression function is not constant, in which case the MSE efficiency gain will be attenuated towards 1.0, often dramatically.

We conclude then that with no missing data, in the partially linear model, substantial improvements upon the sample mean will be realized mainly when the regression errors are heavy-tailed and the regression signal is slight.

We point out that in the example that motivated this work (Section II.4), there is no simple analogue to the sample mean, one that could avoid fitting models to the data.

II.5.3.2. How Critical Are Our Assumptions on Z?

We have made two assumptions on Z: it has a compact support and its density function is positive on that support. We have indicated in Section A that all general papers in the semiparametric kernel-based literature make this assumption, and that it appears to be critical for deriving asymptotic results for problems such as our example in Section II.4. It is certainly well beyond our capabilities to weaken this assumption as it applies to problems such as our motivating example.

The condition that the density of Z be bounded away from zero warns users that the method will deteriorate if there are a few sparsely observed outlying Z-values, see below for numerical evidence of this phenomenon.

Estimation in sub-populations formed by compact subsets of Z can also be of consid-

erable interest in practice, and these compact subsets can be chosen to avoid density spareness and meet our assumptions. A simple example might be where Z is age, and one might be interested in population summaries for those in the 40-60 year age range.

The partially linear model is a special case, however, because all estimates are explicit and what few Taylor expansions are necessary simplify tremendously. That is, the estimates are simple functions of sums of random variables. Cheng (1994) considers the different problem where there is no X and where local constant estimation of the nonparametric function is used, rather than local linear estimation, so that $\widehat{\theta}(z_0) = \sum_{i=1}^{n} K_h(Z_i - z_0)Y_i / \sum_{i=1}^{n} K_h(Z_i - z_0)$. He indicates that the essential condition for this case is that the tails of the density of Z decay exponentially fast.

We tested this numerically in the normal-based simulation of Wang et al. (2004) with the sample size of $n = 500$: similar results were found with $n = 100$. We use the Epanechnikov kernel and estimated the bandwidth using the following methods. First, we regressed Y and X separately on Z, using the DPI bandwidth selection method of Ruppert, Sheather and Wand (1995) to form different estimated bandwidths on each. We then calculated the residuals from these fits, and regressed the residual in Y on the residual in X to get a preliminary estimate $\widehat{\beta}_{\text{start}}$ of β. Following this, we regressed $Y - X^{\mathrm{T}}\widehat{\beta}_{\text{start}}$ on Z to get a common bandwidth, then undersmoothed it by multiplication by $n^{-2/15}$ to get a bandwidth of order $n^{-1/3}$ to eliminate bias, and then reestimated β and $\theta(\cdot)$.

We found that for various Beta distributions on Z, e.g., the Beta(2,1) that violates our assumptions, the sample mean and the semiparametric efficient method were equally efficient. The same occurs for the case that Z is normally distributed. However, when

Z has a t-distribution with 9 degrees of freedom, the sample mean greatly outperforms the undersmoothed estimator (MSE efficiency ≈ 2.0), which in turn out-performed the method that did not employ undersmoothing (MSE efficiency ≈ 2.5). An interesting quote from Ma, Chiou and Wang (2006) is relevant here: also operating in a partial linear model, they state "*This condition enables us to simplify asymptotic expression of certain sums of functions of variables .... also excludes pathological cases where the number of observations in a window defined by the bandwidth may not increase to infinity when* $n \to \infty$".

We conclude that if the design density in Z is at all heavy tailed, then the semiparametric methods will be badly affected. If such a phenomenon happens in the simple case of the partially linear model, it is likely to hold in most other cases. Otherwise, in practice at least, as long as their are no design "stragglers", the assumption is likely to be one required by the technicalities of the problem. How well this generalizes to complex nonlinear problems is unknown.

CHAPTER III

TESTING IN SEMIPARAMETRIC MODELS WITH INTERACTION

III.1. Introduction

Modern genetic association studies often focus on discovery of susceptibility loci, i.e., identification of genetic variants that are associated with the trait under study. The risks of multi-factorial traits, such as cancer, however, are determined by complex interactions among genetic and environmental exposures and the chance for discovery of the underlying susceptibility genes can be substantially reduced if the possibility of heterogeneity in genetic effects due to interactions is ignored. Thus, in recent years, there has been increasing attention in omnibus testing of genetic main effects and gene-environment/gene-gene interactions for detection of susceptibility genes for complex traits. Clearly, tests of association incorporating interactions require larger degrees-of-freedom than those based only on main effects. When the extra degrees-of-freedom required is relatively small, recent studies have shown that the omnibus tests can be a robust and powerful approach for detecting genetic association irrespective of certain specific forms of interactions are present or not (Chatterjee et al., 2006; Kraft et al., 2007). However, if the required degrees-of-freedom is large, then the omnibus tests can have poor power. Thus parsimonious modeling of gene-gene and gene-environment interactions should be considered for construction of powerful omnibus tests.

Chatterjee et al. (2006) proposed the use of Tukey style 1 degree-of-freedom model for interaction for testing the genetic association of a disease with a set of genetic

variants, such as tagging Single Nucleotide Polymorphisms (SNPs) in a candidate gene, that may potentially interact with another set of genetic variants or/and with one or more environmental exposures. SNPs represent a natural genetic variability at high density in the human genome. A genetic locus corresponding to a SNP has two possible alleles (states), namely the normal and the variant. The SNP-genotype data for a subject can have three possible values and are often coded numerically as the number of variant alleles the subject carries on the pair of homologous chromosomes inherited from his/her parents.

In this chapter, we will consider extending the work of Chatterjee et al focussing on the problem of gene-environment interaction. Thus, for example, if D denotes the binary indicator of a disease outcome, X denotes a "design matrix" associated with a set genetic variants G, Z denotes the design matrix associated with an environmental exposure of interest and S denotes a set of additional co-factors, such as age and sex, then the risk of the disease can be modeled using Tukey's form of gene-environment interaction as

$$\text{pr}(D = 1|X, S, Z, \gamma) = H(X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + Z^T\theta_0 + \gamma X^{\mathrm{T}}\beta_0 Z^T\theta_0), \tag{3.1}$$

where $H(\cdot)$ is the logistic distribution function. Notice, unlike in the standard logistic regression model where potentially a separate interaction parameter is allowed between each pair of design elements of the genetic and environmental factors, in model (3.1), a single parameter (γ) is used to capture interactions. Moreover, in model (3.1), the omnibus null hypothesis of interest can be simply stated as $\beta_0 = 0$ under which both genetic main effects and gene-environment interactions disappear from the model. A complication, however, is that under $\beta_0 = 0$, the parameter γ also disappears from the model and hence is not identifiable from the data. Nevertheless,

Chatterjee et al. noticed that for each fixed value of γ, the model (3.1) can be used to construct a valid score-test for $\beta_0 = 0$. They proposed to use maximal of such score-statistics over a range of the parameter γ as the final test statistics to be used for testing $\beta_0 = 0$. They observed that the score-test has particular computational advantages, because under the null hypothesis the model (3.1) reduces to a standard logistic regression model involving only main effects of Z and S.

In this chapter, we extend the work by Chatterjee et al. (2006) in two novel ways. First, we consider modeling complex effects of continuous environmental exposures using nonparametric regression models. The problem is particularly motivated by the fact that modern molecular epidemiologic studies often involve measurement of environmental exposures through continuous biomarkers, the relationships of which with the disease can be highly complex and nonlinear. Thus for example in the logistic context, one might consider the model

$$\text{pr}(D = 1|X, S, Z, \gamma) = H\{X^{\text{T}}\beta_0 + S^{\text{T}}\eta_0 + \theta_0(Z) + \gamma X^{\text{T}}\beta_0\theta_0(Z)\}, \tag{3.2}$$

where $\theta_0(\cdot)$ is an unknown function. Second, we consider very general semiparametric models with possible repeated measures (Lin and Carroll, 2006), where the effects are given through terms roughly of the form on the right-hand-side of (3.2). In particular, we assume that for each subject or cluster i, there are $j = 1, ..., J$ observations $(Y_{ij}, X_{ij}, S_{ij}, Z_{ij})$. We write $\widetilde{Y}_i = (Y_{i1}, ..., Y_{iJ})$, and work with a criterion function

$$\mathcal{L}\{\widetilde{Y}, \nu_1, ..., \nu_J, \zeta_0\}, \text{ with } \nu_j = X_j^{\text{T}}\beta_0\{1 + \gamma\theta_0(Z_j)\} + S_j^{\text{T}}\eta_0, \tag{3.3}$$

where a *criterion function* could mean either a proper likelihood function, a composite likelihood function, i.e., one that is a likelihood function for a reduced set of data, or a proper working likelihood function. In particular, criterion functions have scores in

the parameters $(\beta_0, \eta_0, \zeta_0, \theta_0)$ that have mean zero given appropriate subcomponents of $(X_j, S_j, Z_j)_{j=1}^J$. The case of no repeated measures as in (3.1) occurs when $J = 1$.

Our interest in is in testing for the hypothesis of the form $H_0 : \beta_0 = 0$. As in the Chatterjee, et al. (2006), it is natural to use a score-testing approach to this problem so as to avoid numerical difficulty associated with parameter estimation under general models of the form (3.1) and (3.2). In particular, we note that estimation of γ in these models can be numerically unstable because of lack of identifiability of this parameter under $\beta_0 = 0$. Following Chatterjee et al, we propose to perform score-type tests for each value of γ and then maximize these tests over an interval of γ-values, and use numerical devices to create significance levels. It is possible to create the score statistic directly, and to apply the asymptotic expansions developed by Lin and Carroll (2006) to analyze these statistics. However, two problems arise.

- The first problem is that the direct score statistic requires undersmoothing for the nonparametric estimation of $\theta_0(\cdot)$ in (3.3). By modifying the directly calculated score statistic in a suitable manner, we will show how to create test statistics that lose no local power yet allow regular smoothing, such as crossvalidation.

- The second problem to overcome is that in the repeated measures case that $J > 1$, the distribution of the direct or modified score statistics depend on random variables that are formed as solutions to integral equations. Rather than directly solving the integral equations, we show that the crucial terms can be estimated using nothing more than the Gaussian repeated measures algorithm of Wang (2003), see also Lin, et al. (2004) for a non-iterative solution and Huggins (2006) for another simple computational device.

Thus, we will develop a test statistic that is straightforward to compute, does not require undersmoothing but allows it, and the method also allows a simple implementation when the score test is maximized over a range for γ.

Our methodology is easiest to understand in the non-repeated measures case that $J = 1$, and we take this up in Section III.3, after a discussion in Section III.2 of the difficulties with likelihood ratio testing in this context. The repeated measures case is described in Section III.4. Section III.5 gives the results of a simulation study. Here we find that our maximized tests lose little power when there is no interaction, and can gain great power advantages over a main effects test when there are interactions. Section III.6 illustrate an application of the proposed method for omnibus testing of the effects genetic variants in *NAT2* and their interactions with the number of years since stopping smoking on the risk of colorectal adenoma using a case-control study conducted with the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial (Hayes et al, 2000).

III.2. Identifiability and the Likelihood Ratio Test

The models we study, for example model (3.1), is an example of a problem where γ is a nuisance parameter, and under the null hypothesis (3.5) that $\beta_0 = 0$, the nuisance parameter is unidentified. In other words, the null hypothesis involves a change in the number of parameters from the alternative hypothesis greater than the number of parameters in the null hypothesis. Problems such as this arise in other contexts, see for example Davies (1987).

The mixture problem is a famous case of this phenomenon. Suppose that the null

hypothesis is that the data come from Normal(μ, σ^2), while in the alternative the data come from a mixture of $j = 1, 2$ Normal(μ_j, σ_j^2), with mixing probability π. The null hypothesis can be framed as $h_0 : \pi = 1$, a change of one parameter, but the actual number of parameters changes from 5 to 2. It is well know that the null asymptotic distribution of the likelihood ratio test is not chisquared (Titterington, et al., 1985).

In the change point problem, the idea is that parameters change at a change-point. Thus, for example, in the alternative, one might has a change point η and a shift δ in the mean at the change point. Under the null hypothesis that $\delta = 0$, there is no change point and η is not identified. Again, the distribution of the likelihood ratio test statistic at the null is not chisquared (Brown, et al., 1975).

The model (3.1) is of course reminiscent of Tukey's 1-degree of freedom test for interaction (Tukey, 1949). However, unlike in that context, in our problem the parameter γ is a nuisance parameter and is not of primary interest. The method of Chatterjee, et al. (2006) is more closely akin to the basic suggestion in Davies (1987), namely to fix the nuisance parameter, compute an appropriate test statistic, and then maximize that test statistic over a range of values for the nuisance parameter. Thus, one way to think about our testing procedure is as the appropriate, efficient (both computationally and in terms of power) way of implementing the basic approach of Davies in our context, while taking care to eliminate the concerns of undersmoothing and solution of integral equations that arise from a less targeted approach.

It is interesting to note that the nuisance parameter γ cannot be consistently estimated at the null hypothesis, because it is not identified. This also means that γ cannot be consistently estimated at contiguous alternatives. In practice, even in fully

parametric models, this lack of identifiability means that estimating γ is numerically instable, leading to non-convergence if its range is not restricted.

III.3. Testing Without Repeated Measures

III.3.1. Data and Notation

The data consist of a response Y, parametrically modeled covariates S and X, the latter possibly interacting with a nonparametrically modeled covariate Z. We consider a general loglikelihood or criterion function

$$\mathcal{L}\left[Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z) + X^{\mathrm{T}}\beta_0\{1 + \gamma\theta_0(Z)\}, \zeta_0\right], \tag{3.4}$$

where β_0 and η_0 are the main effects, $\theta_0(\cdot)$ is an unknown function, γ is the interaction effect and ζ_0 are nuisance parameters. In this section, we are interested in testing the parametric hypothesis

$$H_0 : \beta_0 = 0. \tag{3.5}$$

As described in the introduction, Chatterjee et. al. (2006) addressed a similar problem for a fully parametric model where Z is also modeled parametrically. They used a score based testing procedure to test H_0. We generalize their idea for the general semiparametric model given in (3.4). We describe below the major steps to derive the test statistic for testing (3.5).

In what follows, we use a simple subscripting convention for derivatives of the loglikelihood. Thus, with $(\cdot) = [Y, S^{\mathrm{T}}\eta + \theta(Z) + X^{\mathrm{T}}\beta\{1 + \gamma\theta(Z)\}, \zeta]$, we set

$$\mathcal{L}_\theta(\cdot) \;=\; (\partial/\partial v)\mathcal{L}\{Y, S^{\mathrm{T}}\eta + v + X^{\mathrm{T}}\beta(1 + \gamma v), \zeta\}|_{v=\theta(Z)};$$

$$\begin{aligned}
\mathcal{L}_{\theta\theta}(\cdot) &= (\partial^2/\partial v^2)\mathcal{L}\{Y, S^{\mathrm{T}}\eta + v + X^{\mathrm{T}}\beta(1+\gamma v), \zeta\}|_{v=\theta(Z)};\\
\mathcal{L}_{\zeta}(\cdot) &= (\partial/\partial\zeta)\mathcal{L}\{Y, S^{\mathrm{T}}\eta + v + X^{\mathrm{T}}\beta(1+\gamma v), \zeta\}|_{v=\theta(Z)};\\
\mathcal{L}_{\theta\zeta}(\cdot) &= (\partial/\partial\zeta)\mathcal{L}_{\theta}\{Y, S^{\mathrm{T}}\eta + v + X^{\mathrm{T}}\beta(1+\gamma v), \zeta\}|_{v=\theta(Z)},
\end{aligned}$$

etc. Thus, in abuse of notation we do not indicate in the notation that these partial derivatives do not depend on the parameters and covariates only via $S^T\eta + \theta(Z) + X^T\beta\{1+\gamma\theta(Z)\}$.

III.3.2. Estimation of Parameters Under the Null Hypothesis

Here we show how to estimate the parameters and the function at the null hypothesis. The strength of score tests is that one fits the model under the null hypothesis. Under the null hypothesis, the loglikelihood or criterion function for the model is written as $\mathcal{L}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}$, a standard form that is easy to handle. The loglikelihood under the alternative is much harder to deal numerically because of the interaction.

By definition of a loglikelihood or criterion function, at the null hypothesis,

$$0 = E\left[\mathcal{L}_{\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|X, S, Z\right]. \tag{3.6}$$

The first step of the process is to estimate the function $\theta_0(\cdot)$ for any fixed value of $\delta = \delta^* = (\eta^*, \zeta^*)$. We will use kernel methods because of their convenient theory, but this step can be modified in practice using any smoother. The resulting estimate is denoted as $\widehat{\theta}(\cdot, \delta^*)$. Let $K(\cdot)$ be a smooth symmetric density function with bounded support, let h be a bandwidth, and let $K_h(z) = h^{-1}K(z/h)$. Define $\phi_k = \int z^k K(z)dz$ and $G_h(z) = (1, z/h)^{\mathrm{T}}$. We follow Lin and Carroll (2006) to estimate the parameters under H_0: for any fixed value of $\delta = \delta^*$, estimate $\theta_0(z)$ by solving the local likelihood

equation

$$0 \;=\; n^{-1}\textstyle\sum_{i=1}^{n} K_h(Z_i - z)G_h(Z_i - z)\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta^* + \alpha_0 + \alpha_1(Z_i - z), \zeta^*\},$$

for $\widehat{\alpha}_0$ and set $\widehat{\theta}(z, \delta^*) = \widehat{\alpha}_0$.

The second step in the process is now smoothing-method independent. To estimate $\delta_0 = (\eta_0, \zeta_0)$, maximize in δ the function

$$n^{-1}\textstyle\sum_{i=1}^{n} \mathcal{L}\{Y_i, S_i^{\mathrm{T}}\eta + \widehat{\theta}(Z_i, \delta), \zeta\},$$

the so-called profile method, which solves

$$\begin{aligned} 0 &= n^{-1}\textstyle\sum_{i=1}^{n} \{S_i + \widehat{\theta}_\eta(Z_i, \delta)\}\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta + \widehat{\theta}(Z_i, \delta), \zeta\}; \\ 0 &= n^{-1}\textstyle\sum_{i=1}^{n} [\mathcal{L}_\zeta\{Y_i, S_i^{\mathrm{T}}\eta + \widehat{\theta}(Z_i, \delta), \zeta\} + \widehat{\theta}_\zeta(Z_i, \delta)\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta + \widehat{\theta}(Z_i, \delta), \zeta\}], \end{aligned}$$

where $\widehat{\theta}_\eta(Z_i, \delta)$ and $\widehat{\theta}_\zeta(Z_i, \delta)$ is the derivative of $\widehat{\theta}(Z_i, \delta)$ with respect to η or ζ, respectively. all the resulting estimate $\widehat{\delta}$.

III.3.3. The Score Function and Asymptotic Theory

III.3.3.1. Derivation

One approach to developing a score statistic is to fix the function $\theta(\cdot)$, derive the score statistic, and then plug-in estimates of nuisance parameters and the function $\theta(\cdot)$. This does not work well because the function estimate itself needs profiling, and indeed this approach requires undersmoothing for its validity.

In contrast, our test statistic is a particular implementation of the profiled loglikeli-

hood/criterion function, derived as follows. In general, the loglikelihood function for an observation is $\mathcal{L}\{Y, S^{\mathrm{T}}\eta + X^{\mathrm{T}}\beta + \theta(Z) + \gamma X^{\mathrm{T}}\beta\theta(Z), \zeta\}$. Recall that $\delta = (\eta, \zeta)$. For given (β, δ), let $\theta(Z, \beta, \delta)$ be the profile function that solves

$$E\left[\mathcal{L}_\theta\{Y, S^{\mathrm{T}}\eta + X^{\mathrm{T}}\beta + \theta(Z, \beta, \delta) + \gamma X^{\mathrm{T}}\beta\theta(Z, \beta, \delta), \zeta\}|Z\right] = 0. \tag{3.7}$$

Define $\widetilde{X}_{\mathrm{pro}} = X\{1 + \gamma\theta(Z, 0, \delta)\} + \theta_\beta(Z, 0, \delta)$, where $\theta_\beta(Z, \beta, \delta) = (\partial/\partial\beta)\theta(Z, \beta, \delta)$. The profiled loglikelihood is $\mathcal{L}\{Y, S^{\mathrm{T}}\eta + X^{\mathrm{T}}\beta + \theta(Z, \beta, \delta) + \gamma X^{\mathrm{T}}\beta\theta(Z, \beta, \delta), \zeta\}$. Differentiating it with respect to β and evaluating at the null hypothesis $\beta = 0$, the profiled (efficient) score is easily seen to be $\widetilde{X}_{\mathrm{pro}}\mathcal{L}_\theta\{Y, S^{\mathrm{T}}\eta + \theta(Z, 0, \delta), \zeta\}$.

In addition, differentiating (3.7) with respect to β and evaluating it at $\beta = 0$ and $\delta = \delta_0$ shows that $\widetilde{X}_{\mathrm{pro}} = \{1 + \gamma\theta_0(Z)\}\widetilde{X}$, where

$$\widetilde{X} = X - E[X\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z]/E[\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z].$$

We thus propose the following profiled score statistic for β_0:

$$\mathcal{T}_{n,\mathrm{pro}}(\gamma) \;=\; n^{-1/2}\textstyle\sum_{i=1}^n\{1 + \gamma\widehat{\theta}(Z_i, \widehat{\delta})\}\widetilde{X}_{i,\mathrm{est}}\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\widehat{\eta} + \widehat{\theta}(Z_i, \widehat{\eta}), \widehat{\zeta}\}, \tag{3.8}$$

where $\widetilde{X}_{i,\mathrm{est}}$ is an estimated version of $\widetilde{X}_i$, with the terms to be estimated in $\widetilde{X}$ obtained by separate nonparametric regressions in the numerator and denominator. The normalization by $n^{-1/2}$ is convenient for the asymptotic theory.

III.3.3.2. Theoretical Result

Let $\delta_0 = (\eta_0^{\mathrm{T}}, \zeta_0^{\mathrm{T}})^{\mathrm{T}}$ and make the definitions

$$\theta_\delta(z_0, \delta_0) \;=\; -\frac{E[\mathcal{L}_{\theta\delta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z = z_0]}{E[\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z = z_0]};$$

$$
\begin{aligned}
\epsilon &= \mathcal{L}_\delta\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\} + \theta_\delta(Z, \delta_0)\mathcal{L}_\theta\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\};\\
\mathcal{M} &= -E(\epsilon\epsilon^{\mathrm{T}});\\
\mathcal{N} &= E\Big(X\{1+\gamma\theta_0(Z)\}[\mathcal{L}_{\theta\delta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}\\
&\qquad\qquad +\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}\theta_\delta(Z, \delta_0)]^{\mathrm{T}}\Big);\\
\Psi(\gamma) &= \{1+\gamma\theta_0(Z)\}\widetilde{X}\mathcal{L}_\theta\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\} - \mathcal{N}\mathcal{M}^{-1}\epsilon.
\end{aligned}
$$

The main result of this section justifying our methodology is stated below. Technically, a precise argument requires little more than that the linear expansions for the parametric and nonparametric parts given in Lin and Carroll (2006) hold to order $o_p(n^{-1/2})$, the latter uniformly.

Result 3 Suppose that we are testing for $H_0 : \beta_0 = 0$. Assume that $h \propto n^{-\alpha}$ with $1/3 \le \alpha \le 1/5$. Then, for any fixed γ, the score function for β_0 can be written as

$$
\mathcal{T}_{n,\mathrm{pro}}(\gamma) = n^{-1/2}\textstyle\sum_{i=1}^{n}\Psi_i(\gamma) + o_p(1).
$$

In addition, assume that for any γ_1 and γ_2, $\mathcal{V}(\gamma_1, \gamma_2) = E\{\Psi(\gamma_1)\Psi^{\mathrm{T}}(\gamma_2)\}$ is finite. Then, under the hypothesis that $\beta_0 = 0$, $\mathcal{T}_{n,\mathrm{pro}}(\gamma)$ as a function of $\gamma \in [L, R]$ converges weakly to a Gaussian process $\mathcal{W}(\gamma)$ with mean zero and covariance function $\mathcal{V}(\gamma_1, \gamma_2)$.

Remark 6 There are two methods that can be used to estimate the covariance matrix of the estimated score.

- First, suppose as in logistic regression that there are no nuisance parameters ζ_0, and that $\mathcal{L}(\cdot)$ is a loglikelihood function and not a general criterion function. Then we can write $\Psi_i(\gamma) = \Psi_i^*(\gamma)\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i)\}$ with $\Psi_i^*(\gamma) = \{1 + \gamma\theta_0(Z_i)\}\widetilde{X}_i - \mathcal{N}\mathcal{M}^{-1}\widetilde{S}_i$, where

$$
\widetilde{S} = S - E[S\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z]/E[\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z].
$$

Let $\widehat{\Psi}_i^*(\gamma)$ be the estimated version of $\Psi_i^*(\gamma)$. This estimated version requires through definition of $\widetilde{X}_i$, $\widetilde{S}_i$ and additional nonparametric regressions, which are easily accomplished via kernel or spline methods. Further, let $\mathcal{I}_{\theta,\text{null}}\{S_i^{\text{T}}\eta_0 + \theta_0(Z_i), \zeta_0\}$ be the conditional information matrix for θ under the null model. Then we estimate the covariance matrix of $\mathcal{T}_n(\gamma)$

$$\mathcal{I}_{\beta_0,n}(\gamma) \;=\; n^{-1}\textstyle\sum_{i=1}^n \mathcal{I}_{\theta,\text{null}}\{S_i^{\text{T}}\widehat{\eta} + \widehat{\theta}(Z_i,\widehat{\eta})\}\widehat{\Psi}_i^*(\gamma)\{\widehat{\Psi}_i^*(\gamma)\}^{\text{T}}.$$

- In general, $\mathcal{I}_{\beta_0,n}(\gamma)$ can be estimated as the sample covariance matrix of the terms $\widehat{\Psi}_i(\gamma)$, the estimated version of $\Psi_i(\gamma)$. In likelihood problems, simplifications arise because one can compute the covariance matrix of $\Psi(\cdot)$ given (X, Z, S) using Fisher Information calculations.

Remark 7 The validity and unbiasedness of the profiled score statistic primarily depends on the use of $\widetilde{X}$. In simpler models, such as the Gaussian model, $\widetilde{X} = X - E(X|Z)$ is simply the residual of a nonparametric Gaussian regression of each component of X on Z. In general, $\widetilde{X}$ can be thought of the residual of a weighted nonparametric Gaussian regression of each component of X on Z, where the error variance for weighting is taken to be $-1/\mathcal{L}_{\theta\theta}(\cdot)$. This interpretation enables us to construct estimates of $\widetilde{X}$ with considerable ease in many cases, especially in the presence of repeated measurements, see Section III.4 for details.

III.3.4. The Test Statistic and Its Implementation

Here we define our test statistic and show how to implement it in practice to compute critical values.

The score test statistic, for a fixed value of γ, is then given by

$$\mathcal{T}_{n,\text{pro}}(\gamma)^{\text{T}}\mathcal{I}^{-1}_{\beta_0,n}(\gamma)\mathcal{T}_{n,\text{pro}}(\gamma).$$

We compute the final test statistic as

$$\mathcal{T}^*_n \;=\; \max_{L\le\gamma\le R} \mathcal{T}^{\text{T}}_{n,\text{pro}}(\gamma)\mathcal{I}^{-1}_{\beta_0,n}(\gamma)\mathcal{T}_{n,\text{pro}}(\gamma),$$

where L and R are pre-specified lower and upper bound of γ. Our approach is also related to adaptive tests that have been developed for nonparametric alternatives of functions with unknown smoothness, compare e.g. Horowitz and Spokoiny (2001).

To implement the test, we need to simulate the null distribution of $\mathcal{T}^*_n$ and obtain the desired critical values. Our method avoids the need to determine critical values for the maximum of a function of a Gaussian process. Using Result 3 we can generate realizations from the limiting distribution of the score statistic as

$$T_0(\gamma) \;=\; n^{-1/2}\textstyle\sum_{i=1}^n \widehat{\Psi}_i(\gamma)\mathcal{Z}_i,$$

where $\widehat{\Psi}(\gamma)$ is $\Psi(\gamma)$ evaluated at $\widehat{\delta}$ and $\widehat{\theta}(z,\widehat{\delta})$, and $\mathcal{Z}_1,\ldots,\mathcal{Z}_n$ are standard normal random variates which are drawn independent of the data. The null distribution of $\mathcal{T}^*_n$ is then simulated by generating $\mathcal{T}^*_0 = \max_{L\le\gamma\le R} T_0(\gamma)^{\text{T}}\mathcal{I}^{-1}_{\beta_0,n}(\gamma)T_0(\gamma)$ repeatedly. This method is the semiparametric version of a method discussed by Lin and Zhou (2004) and Chatterjee, et al. (2006).

III.4. General Interaction Model with Repeated Measures

III.4.1. Data and Notation

In this section we generalize the ideas presented earlier to the case when repeated measures are present in the data. Repeated measures models can arise from vari-

ous fields of research, e.g., matched case-control studies, finance, epidemiology and many others. The key feature of these models is that the nonparametric function is evaluated for each of the repeated measurements. Lin and Carroll (2006) developed kernel-based estimation procedures and investigated asymptotic properties of the estimators in general semiparametric regression problems. We will use their results and methodology in our context.

In this section we set out the notation to be used. For simplicity only, we suppose that there are J repeated measurements for each individual. Only obvious notational changes are required for the more general case. Specifically, we consider a loglikelihood or criterion function

$$\mathcal{L}\{\widetilde{Y}, \nu_1(\beta_0, \theta_0, \eta_0), \ldots, \nu_J(\beta_0, \theta_0, \eta_0), \zeta_0\},$$

where $\nu_j(\beta_0, \theta_0, \eta_0) = X_j^{\mathrm{T}}\beta_0\{1+\gamma\theta_0(Z_j)\}+\theta_0(Z_j)+S_j^{\mathrm{T}}\eta_0$, γ is the common interaction parameter for each of the repeated measurements and ζ_0 is the collection of all the nuisance parameters. Then, with a slight abuse of notation in the first formula below,

$$E[\partial\mathcal{L}\{\widetilde{Y}, \nu_1(\beta_0, \theta_0, \eta_0), \ldots, \nu_J(\beta_0, \theta_0, \eta_0), \zeta_0\}/\partial\{\theta_0(Z_k)\}|(X_j, Z_j, S_j)_{j=1}^J] = 0;$$
$$E[\partial\mathcal{L}\{\widetilde{Y}, \nu_1(\beta_0, \theta_0, \eta_0), \ldots, \nu_J(\beta_0, \theta_0, \eta_0), \zeta_0\}/\partial(\beta, \eta, \zeta)|(X_j, Z_j, S_j)_{j=1}^J] = 0,$$

see Lin and Carroll (2006) for more discussion. In Section III.4.6, we describe methods for the partially linear model when working independence among the errors is used, and hence weaker conditioning assumptions are required.

Letting

$$\cdot = \{\widetilde{Y}, \nu_1(\beta, \theta, \eta), \ldots, \nu_J(\beta, \theta, \eta), \zeta\},$$

we define terms $\mathcal{L}_{j\theta}(\cdot)$, $\mathcal{L}_{jk\theta}(\cdot)$, $\mathcal{L}_{\zeta}(\cdot)$, $\mathcal{L}_{j\theta\zeta}(\cdot)$ in the same was as described in Section III.3.1. Thus, for example,

$$\begin{aligned}
\mathcal{L}_{j\theta}(\cdot) &= \frac{\partial}{\partial v_j}\mathcal{L}\Big[\widetilde{Y}, S_1^{\mathrm{T}}\eta + \theta(Z_1) + X_1^{\mathrm{T}}\beta\{1+\gamma\theta(Z_1)\}, ..., S_j^{\mathrm{T}}\eta + v_j + X_j^{\mathrm{T}}\beta(1+\gamma v_j), \\
&\qquad\qquad ..., S_J^{\mathrm{T}}\eta + \theta(Z_J) + X_J^{\mathrm{T}}\beta\{1+\gamma\theta(Z_J)\}, \zeta\Big]_{v_j=\theta(Z_j)}; \\
\mathcal{L}_{jk\theta}(\cdot) &= \frac{\partial^2}{\partial v_j \partial v_k}\mathcal{L}\Big[\widetilde{Y}, S_1^{\mathrm{T}}\eta + \theta(Z_1) + X_1^{\mathrm{T}}\beta\{1+\gamma\theta(Z_1)\}, ... \\
&\qquad\qquad S_j^{\mathrm{T}}\eta + v_j + X_j^{\mathrm{T}}\beta(1+\gamma v_j), ... \\
&\qquad\qquad S_k^{\mathrm{T}}\eta + v_k + X_k^{\mathrm{T}}\beta(1+\gamma v_k), ... \\
&\qquad\qquad S_J^{\mathrm{T}}\eta + \theta(Z_J) + X_J^{\mathrm{T}}\beta\{1+\gamma\theta(Z_J)\}, \zeta\Big]_{v_j=\theta(Z_j), v_k=\theta(Z_k)},
\end{aligned}$$

etc.

III.4.2. Estimation Under the Null Model

In this section, we display the method for estimation of parameters and the function $\theta(\cdot)$, at the null hypothesis.

Under the null hypothesis, the criterion function is given by

$$\mathcal{L}\{\widetilde{Y}, \theta_0(Z_1) + S_1^{\mathrm{T}}\eta_0, \ldots, \theta_0(Z_J) + S_J^{\mathrm{T}}\eta_0, \zeta_0\}.$$

Let $\delta = (\eta, \zeta)$. We estimate $\theta_0(\cdot)$ and δ_0 under the null model using methodology proposed in Lin and Carroll (2006): for any fixed $\delta = \delta^* = (\eta^*, \zeta^*)$, estimate $\theta_0(z)$ by solving for (α_0, α_1)

$$\begin{aligned}
0 &= \textstyle\sum_{i=1}^{n}\sum_{j=1}^{J} K_h(Z_{ij}-z)G(Z_{ij}-z) \\
&\quad \times \mathcal{L}_{j\theta}\Big\{\widetilde{Y}_i, \widehat{\theta}(Z_{i1}, \delta^*) + S_{i1}^{\mathrm{T}}\eta^*, \ldots, \alpha_0 + \alpha_1(Z_{ij}-z)/h + S_{ij}^{\mathrm{T}}\eta^*, \ldots,
\end{aligned}$$

$$\widehat{\theta}(Z_{iJ}, \delta^*) + S_{iJ}^{\mathrm{T}}\eta^*, \zeta^*\Big\},$$

and setting $\widehat{\theta}(z, \delta^*) = \widehat{\alpha}_0$. Next, estimate δ by maximizing

$$\textstyle\sum_{i=1}^{n} \mathcal{L}\{\widetilde{Y}_i, \widehat{\theta}(Z_{i1}, \delta) + S_{i1}^{\mathrm{T}}\eta, \ldots, \widehat{\theta}(Z_{iJ}, \delta) + S_{iJ}^{\mathrm{T}}\eta, \zeta\}$$

with respect to δ. This can be accomplished by implementing a profiling algorithm as in Lin and Carroll (2006).

III.4.3. The Score Function and Asymptotic Theory

III.4.3.1. Derivation of the Profile Score

As we have seen in Section III.3.3, our test statistic will be based upon the score function of a profiled loglikelihood. In this section, we derive the profiled loglikelihood and the score function, but here the repeated measures aspect makes the calculations less transparent and indeed leads to real issues of implementation. Let $f_j(z)$ be the marginal density of Z_j Again, for any (β, δ), we define $\theta(z, \beta, \delta)$ by the repeated measures version of (3.7), namely the solution to the equation

$$\begin{aligned} 0 \;=\; & \sum_{j=1}^{J} f_j(z) E\Big[\mathcal{L}_{j\theta}\{\widetilde{Y}, X_1^{\mathrm{T}}\beta\{1 + \gamma\theta(Z_1, \beta, \delta)\} + \theta(Z_1, \beta, \delta) + S_1^{\mathrm{T}}\eta, ..., \\ & X_J^{\mathrm{T}}\beta\{1 + \gamma\theta(Z_J, \beta, \delta)\} + \theta(Z_J, \beta, \delta) + S_J^{\mathrm{T}}\eta, \zeta\} | Z_j = z\Big]. \quad (3.9) \end{aligned}$$

Defining $\omega_j(\beta, \theta, \delta) = X_j^{\mathrm{T}}\beta\{1 + \gamma\theta(Z_j, \beta, \delta)\} + \theta(Z_j, \beta, \delta) + S_j^{\mathrm{T}}\eta$, the profiled loglikelihood function is $\mathcal{L}\{\widetilde{Y}, \omega_1(\beta, \theta, \delta), ..., \omega_J(\beta, \theta, \delta), \zeta\}$.

Let $\mathcal{L}_{j\theta\beta}\{\widetilde{Y}, \omega_1(\beta, \theta, \delta), ..., \omega_J(\beta, \theta, \delta), \zeta\}$ and $\mathcal{L}_{jk\theta}\{\widetilde{Y}, \omega_1(\beta, \theta, \delta), ..., \omega_J(\beta, \theta, \delta), \zeta\}$ be the derivatives of $\mathcal{L}_{j\theta}\{\widetilde{Y}, \omega_1(\beta, \theta, \delta), ..., \omega_J(\beta, \theta, \delta), \zeta\}$ with respect to β and $\theta(Z_k, \beta, \delta)$,

respectively. Differentiating and setting $\beta = 0$, the profiled score becomes

$$\sum_{j=1}^{J} [\{1 + \gamma\theta(Z_j, 0, \delta)\}X_j + \theta_\beta(Z_j, 0, \delta, \gamma)]\, \mathcal{L}_{j\theta}\{\widetilde{Y}, \omega_1(0, \theta, \delta), ..., \omega_J(0, \theta, \delta), \zeta\},$$

where by differentiating (3.9) with respect to β and solving, $\theta_\beta(z, \beta, \delta, \gamma)$ is the solution of the functional integral equation:

$$\begin{aligned} 0 = \textstyle\sum_{j=1}^{J} f_j(z) E\Big[& \mathcal{L}_{j\theta\beta}\{\widetilde{Y}, \omega_1(\beta, \theta, \delta), \ldots, \omega_J(\beta, \theta, \delta), \zeta\} \\ & + \textstyle\sum_{k=1}^{J} \mathcal{L}_{jk\theta}\{\widetilde{Y}, \omega_1(\beta, \theta, \delta), \ldots, \omega_J(\beta, \theta, \delta), \zeta_0\}\theta_\beta(Z_k, \beta, \delta, \gamma) \Big| Z_j = z \Big]. \end{aligned} \quad (3.10)$$

Then, for any fixed value of γ, the profiled score function for β_0 evaluated at $\beta_0 = 0$, $\delta_0 = \widehat{\delta}$ and $\theta(z) = \widehat{\theta}(z, \widehat{\delta})$ is given by

$$\begin{aligned} \mathcal{T}_{n,\mathrm{pro}}(\gamma) &= n^{-1/2} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{J} [\{1 + \gamma\widehat{\theta}(Z_{ij}, \widehat{\delta})\}X_{ij} + \widehat{\theta}_\beta(Z_{ij}, 0, \widehat{\delta}, \gamma)] \\ &\quad \times \mathcal{L}_{j\theta}\{\widetilde{Y}, \widehat{\theta}(Z_{i1}, \widehat{\delta}) + S_{i1}^{\mathrm{T}}\widehat{\eta}, \ldots, \widehat{\theta}(Z_{iJ}, \widehat{\delta}) + S_{iJ}^{\mathrm{T}}\widehat{\eta}, \widehat{\zeta}\}. \end{aligned}$$

III.4.3.2. Asymptotic Theory

Denote $(\cdot) = \{\widetilde{Y}, \omega_1(\beta_0, \theta_0, \delta_0), ..., \omega_J(\beta_0, \theta_0, \delta_0), \zeta_0\}$ and denote $(\cdot_i)$ to be $(\cdot)$ evaluated at the i^{th} observation. Do all calculations at the null model $\beta_0 = 0$. Define $\theta_\delta(z, \delta_0)$ such that

$$0 = \textstyle\sum_{j=1}^{J} f_j(z) E\{\mathcal{L}_{j\theta\delta}(\cdot) + \sum_{k=1}^{J} \theta_\delta(Z_k, \delta_0)\mathcal{L}_{jk\theta}(\cdot) | Z_j = z\}.$$

Further define

$$\begin{aligned} \mathcal{M}_1 &= -\mathrm{cov}\Big[\mathcal{L}_\delta(\cdot) + \textstyle\sum_{j=1}^{J} \mathcal{L}_{j\theta}(\cdot)\theta_\delta(Z_j, \delta_0)\Big]; \\ \mathcal{M}_2 &= E\Big[\textstyle\sum_{j=1}^{J} \{1 + \gamma\theta_0(Z_j)\}X_j\{\mathcal{L}_{j\theta\delta}(\cdot) + \sum_{k=1}^{J} \theta_\delta(Z_k, \eta_0)\mathcal{L}_{jk\theta}(\cdot)\}^{\mathrm{T}}\Big]; \end{aligned}$$

$$\begin{aligned}\Psi_i(\gamma) &= \sum_{j=1}^{J}[X_{ij}\{1+\gamma\theta_0(Z_{ij})\}+\theta_\beta(Z_{ij},0,\delta_0,\gamma)]\mathcal{L}_{j\theta}(\cdot_i) \\ &\quad -\mathcal{M}_2\mathcal{M}_1^{-1}\{\mathcal{L}_\delta(\cdot_i)+\sum_{j=1}^{J}\mathcal{L}_{j\theta}(\cdot_i)\theta_\delta(Z_{ij},\delta_0)\}.\end{aligned}$$

Then we have the following result:

Result 4 Suppose that we are interested in testing $H_0 : \beta_0 = 0$. Assume that $h \propto n^{-\alpha}$ where $1/3 \leq \alpha \leq 1/5$. Then, for any fixed γ, the score function for β_0 can be written as

$$\mathcal{T}_{n,\text{pro}}(\gamma) = n^{-1/2}\sum_{i=1}^{n}\Psi_i(\gamma) + o_p(n^{-1/2}).$$

In addition, assume that, for any γ_1 and γ_2, $\mathcal{V}(\gamma_1,\gamma_2) = E\{\Psi(\gamma_1)\Psi^{\mathrm{T}}(\gamma_2)\}$ is finite. Then, under the hypothesis that $\beta_0 = 0$, $\mathcal{T}_{n,\text{pro}}(\gamma)$ as a function of $\gamma \in [L,R]$ converges weakly to a Gaussian process $\mathcal{W}(\gamma)$ with mean zero and covariance function $\mathcal{V}(\gamma_1,\gamma_2)$.

Using Result 4, we construct the test statistic and the critical values in the obvious analogy with Sections III.3.3-III.3.4. To implement this in practice though, we have to solve the integral equations for $\theta_\beta(\cdot)$ and $\theta_\delta(\cdot)$, which is very difficult to do. In the next section, we show how to estimate these quantities without directly solving the integral equations.

III.4.4. Computation of $\theta_\beta(\cdot)$ and $\theta_\delta(\cdot)$

The main difficulty in performing the score test is that for each γ, one has to compute $\widehat{\theta}_\beta(z,0,\delta_0,\gamma)$ and $\widehat{\theta}_\delta(z,0,\delta_0)$, the former of which is the solution of a integral equation (3.10), making implementation difficult. In this section we show that $\theta_\beta(z,0,\delta_0,\gamma)$ can be viewed as a regression function and hence can be computed via a nonparametric Gaussian repeated measures regression, which is easily computed and for which the exact solution is known, see Huggins (2006) and Lin, et al. (2004). The result can be

stated as follows: details are in the Appendix.

Result 5 Define $Q_{ij} = -X_{ij}\{1+\gamma\theta_0(Z_{ij})\}$. Let V_i be the $J \times J$ matrix with elements $v^{ijk} = -\mathcal{L}_{jk\theta}(\cdot_i)$. Then $\theta_\beta(z, 0, \delta_0, \gamma)$ is identified as the formal solution of the Gaussian repeated measures problem solved by Wang (2003) and Huggins (2006) with "responses" being the components of Q_{ij} and the inverse of the covariance matrix being V_i.

The algorithm for estimating $\theta_\beta(\cdot)$ now is quite simple. Define

$$\widehat{Q}_{ij} = -\{1 + \gamma\widehat{\theta}(Z_{ij}, \widehat{\delta})\}X_{ij}.$$

Then we construct each component of $\widehat{\theta}_\beta(z, 0, \widehat{\delta}, \gamma)$ by performing a nonparametric repeated measures regression under the null model with $\beta = 0$, with the response being the appropriate component of $\widehat{Q}_{ij}$ and the inverse of the covariance matrix being $\widehat{V}_i = (\widehat{v}^{ijk})$, where $\widehat{v}^{ijk} = -\mathcal{L}_{jk\theta}\{\widetilde{Y}_i, \widehat{\theta}(Z_{i1}, \widehat{\delta}) + S_{i1}^{\mathrm{T}}\widehat{\eta}, \ldots, \widehat{\theta}(Z_{iJ}, \widehat{\delta}) + S_{iJ}^{\mathrm{T}}\widehat{\eta}, \widehat{\zeta}\}$ and $\widehat{\theta}(z, \widehat{\delta})$ is computed under the null model with $\beta_0 = 0$.

One can estimate $\theta_\delta(\cdot)$ in a similar manner. We do this componentwise. Let $\mathcal{L}_{j\theta\delta,\ell}(\cdot)$ denote the ℓ^{th} component of $\mathcal{L}_{j\theta\delta}(\cdot)$, and similarly for $\theta_{\delta,\ell}(\cdot)$. Define $(R_{i1}^\ell, \ldots, R_{iJ}^\ell)^{\mathrm{T}} = -V_i^{-1}\{\mathcal{L}_{i1\theta\delta,\ell}(\cdot), \ldots, \mathcal{L}_{iJ\theta\delta,\ell}(\cdot)\}^{\mathrm{T}}$. Then $\theta_{\delta,\ell}(\cdot)$ can be thought as the Gaussian repeated measures regression of R_{ij}^ℓ on Z_{ij} pretending the inverse of the covariance matrix for the i^{th} cluster is V_i. In practice, one constructs $\widehat{\theta}_{\delta,\ell}(\cdot)$ using $\widehat{R}_{ij}^\ell$ and $(\widehat{V}_i)^{-1}$.

III.4.5. Special Case: Partially Linear Repeated Measurement Model

In this section we consider the partially linear Gaussian model as an example to demonstrate our methodology. Specifically, we consider the model

$$Y_{ij} \;=\; X_{ij}^{\mathrm{T}}\beta_0\{1+\gamma\theta_0(Z_{ij})\}+\theta_0(Z_{ij})+S_{ij}^{\mathrm{T}}\eta_0+\epsilon_{ij},$$

where $\widetilde{\epsilon}_i = (\epsilon_{i1},\ldots,\epsilon_{iJ})$ has a Normal$(0,\Sigma)$ distribution. We want to test for H_0 : $\beta_0 = 0$. The asymptotic theory is not affected by estimation of Σ, so here we assume it is known.

Let $\Sigma = (\sigma_{jk})_{j,k=1,\ldots,J}$ and $\Sigma^{-1} = V = (v^{jk})$. Then the loglikelihood function is given by

$$\mathcal{L} \;=\; -(1/2)\textstyle\sum_{q=1}^{J}\sum_{\ell=1}^{J} v^{q\ell}(Y_q-\mu_q)(Y_\ell-\mu_\ell),$$

where $\mu_j = X_j^{\mathrm{T}}\beta_0\{1+\gamma\theta_0(Z_j)\}+\theta_0(Z_j)+S_j^{\mathrm{T}}\eta_0$. Now we observe that when $\beta_0 = 0$,

$$\begin{aligned}
\mathcal{L}_{j\theta}(\cdot) &= \textstyle\sum_{\ell=1}^{J} v^{j\ell}(Y_\ell-\mu_\ell);\\
\mathcal{L}_{j\theta\beta}(\cdot) &= \gamma X_j\textstyle\sum_{\ell=1}^{J} v^{j\ell}(Y_\ell-\mu_\ell)-\sum_{\ell=1}^{J} v^{j\ell}X_\ell\{1+\gamma\theta_0(Z_\ell)\};\\
\mathcal{L}_{jk\theta}(\cdot) &= -v^{jk}.
\end{aligned}$$

For $\beta_0 = 0$, $\theta_\beta(z,0,\eta_0,\gamma)$ solves:

$$0 = \textstyle\sum_{j=1}^{J} f_j(z)E\Big(\sum_{k=1}^{J} v^{jk}[X_k\{1+\gamma\theta_0(Z_k)\}+\theta_\beta(Z_k,0,\eta_0,\gamma)]\Big|Z_j=z\Big). \quad (3.11)$$

Hence the profiled score function is given by

$$\begin{aligned}
\mathcal{T}_{n,\mathrm{pro}}(\gamma) \;=\; n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{J} v^{jk}[\{1+\gamma\widehat{\theta}(Z_{ij},\widehat{\eta})\}X_{ij}+\widehat{\theta}_\beta(Z_{ij},0,\widehat{\eta},\gamma)]\\
\times\{Y_{ik}-\widehat{\theta}(Z_{ik},\widehat{\eta})-S_{ik}^{\mathrm{T}}\widehat{\eta}\}.
\end{aligned}$$

Now one can construct the score test by using Result 4.

Remark 8 Referring to Section III.4.4, we observe that estimation of $\theta_\beta(\cdot)$ becomes much simpler in this case. Using the fact that $\mathcal{L}_{jk\theta}(\cdot) = -v^{jk}$, one can construct $\widehat{\theta}_\beta(\cdot)$ by performing a nonparametric *componentwise* Gaussian repeated measures regression of $\widehat{Q}_k = -\{1 + \gamma\widehat{\theta}(Z_k, \widehat{\eta})\}X_k$ on Z_k pretending the error covariance matrix to be Σ, where $\widehat{\theta}(z, \widehat{\eta})$ is computed under the null model with $\beta_0 = 0$. Similarly, one can estimate $\theta_\eta(\cdot)$ by performing a nonparametric Gaussian repeated measures regression of $-S_{ij}$ on Z_{ij} using Σ as the error covariance matrix.

III.4.6. Testing Under Working Independence

In practice, often working independence is used to simplify the computations in the presence of repeated measures. In this setup, one pretends that there is no correlation among the data. In our context, this leads to the assumption that $\sigma_{jk} = 0$ for $j \neq k$, and we work with the criterion function

$$\mathcal{L}^{\mathrm{WI}} \;=\; -(1/2)\textstyle\sum_{j=1}^{J}\sigma_{jj}^{-1}(Y_j - \mu_j)^2,$$

where $\mu_j = X_j^{\mathrm{T}}\beta_0\{1 + \gamma\theta_0(Z_j)\} + \theta_0(Z_j) + S_j^{\mathrm{T}}\eta_0$. Note that the use of this criterion function simplifies the calculations to a great extent. For any generic random variable W, define $\widetilde{W}_j = W_j - m_Z^W(Z_j)$ with

$$m_Z^W(z) \;=\; \textstyle\sum_{j=1}^{J}\sigma_{jj}^{-1}f_j(z)E(W_j|Z_j = z)/\sum_{j=1}^{J}\sigma_{jj}^{-1}f_j(z).$$

Under the hypothesis that $H_0 : \beta_0 = 0$, we then observe that now $\theta_\beta(\cdot)$ and $\theta_\eta(\cdot)$ have closed form expressions:

$$\theta_\beta(z, 0, \eta_0, \gamma) \;=\; -\{1 + \gamma\theta_0(z)\}m_Z^X(z);$$

$$\theta_\eta(z,\eta_0) \;=\; -m_Z^S(z).$$

The profiled score statistic is given by

$$\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma) \;=\; n^{-1/2}\textstyle\sum_{i=1}^n\sum_{j=1}^J\sigma_{jj}^{-1}\{1+\gamma\widehat{\theta}(Z_{ij},\widehat{\eta})\}\widetilde{X}_{ij,\mathrm{est}}\{Y_{ij}-\widehat{\theta}(Z_{ij},\widehat{\eta})-S_{ij}^{\mathrm{T}}\widehat{\eta}\},$$

where $\widetilde{X}_{ij,\mathrm{est}} = X_{ij} - \widehat{m}_Z^X(Z_{ij})$. One can compute $\widehat{m}_Z^X(z)$ by running a componentwise Gaussian repeated measures regression on X_{ij} and Z_{ij} using working independence setup.

Further define

$$\begin{aligned}\mathcal{M}_1 &= -\mathrm{cov}\Big[\textstyle\sum_{j=1}^J\sigma_{jj}^{-1}\widetilde{S}_j\{Y_j-\theta_0(Z_j)-S_j^{\mathrm{T}}\eta_0\}\Big];\\ \mathcal{M}_2 &= -E\Big[\textstyle\sum_{j=1}^J\sigma_{jj}^{-1}\{1+\gamma\theta_0(Z_j)\}X_j\widetilde{S}_j^{\mathrm{T}}\Big],\end{aligned}$$

Result 4 then translates to the following result:

Result 6 Assume that $h \propto n^{-\alpha}$ where $1/3 \le \alpha \le 1/5$. Then, under the assumption of working independence

$$\begin{aligned}\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma) \;=\; n^{-1/2}\textstyle\sum_{i=1}^n\sum_{j=1}^J\sigma_{jj}^{-1}\Big[\{1+\gamma\theta_0(Z_{ij})\}\widetilde{X}_{ij}+\mathcal{M}_2\mathcal{M}_1^{-1}\widetilde{S}_{ij}\Big]\\ \times\{Y_{ij}-\theta_0(Z_{ij})-S_{ij}^{\mathrm{T}}\eta_0\}+o_p(1).\end{aligned}$$

Define $\Psi_{ij}^*(\gamma) = \{1+\gamma\theta_0(Z_{ij})\}\widetilde{X}_{ij}+\mathcal{M}_2\mathcal{M}_1^{-1}\widetilde{S}_{ij}$ and let $\widehat{\Psi}_{ij}^*(\gamma)$ be the sample version. Under the null hypothesis, we estimate the covariance matrix of $\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}$ by

$$\mathcal{I}_{\beta_0,n}^{\mathrm{WI}} \;=\; n^{-1}\textstyle\sum_{i=1}^n\sum_{j=1}^J\sigma_{jj}^{-1}\widehat{\Psi}_{ij}^*(\gamma)\{\widehat{\Psi}_{ij}^*(\gamma)\}^{\mathrm{T}}.$$

The score statistic, maximized over γ, is then given by

$$\mathcal{T}_n^* \;=\; \max_{\gamma\in[L,R]} \mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma)^{\mathrm{T}}(\mathcal{I}_{\beta_0,n}^{\mathrm{WI}})^{-1}\mathcal{T}_{n,\mathrm{pro}}^{\mathrm{WI}}(\gamma).$$

Using Lemma 6, we can now implement the score test using the technique described in Section III.3.4. We start by generating

$$\mathcal{T}_0^{\mathrm{WI}}(\gamma) \;=\; n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{J}\sigma_{jj}^{-1}\widehat{\Psi}_{ij}^*(\gamma)\mathcal{Z}_{ij},$$

where $\mathcal{Z}_i = (Z_{i1},\ldots,Z_{iJ})^{\mathrm{T}}$, $i = 1,\ldots,n$ are independent random vectors generated from Normal$(0,\widehat{\Sigma})$. One can form $\widehat{\Sigma}$ as the sample covariance matrix of the residuals $\{Y_{ij} - \widehat{\theta}(Z_{ij},\widehat{\eta}) - S_{ij}^{\mathrm{T}}\widehat{\eta}\}$. The null distribution of $\mathcal{T}_n^*$ is then simulated by repeatedly generating

$$\mathcal{T}_0^* \;=\; \max_{\gamma\in[L,R]} \mathcal{T}_0^{\mathrm{WI}}(\gamma)^{\mathrm{T}}(\mathcal{I}_{\beta_0,n}^{\mathrm{WI}})^{-1}\mathcal{T}_0^{\mathrm{WI}}(\gamma).$$

Remark 9 We reiterate that one needs to estimate $\widehat{m}_Z^X(Z_{ij})$ and $\widehat{m}_Z^S(Z_{ij})$ to implement the score test. These quantities can be easily estimated by performing componentwise Gaussian repeated measures regressions of X_{ij} and S_{ij} on Z_{ij} using the working independence setup.

III.5. Simulations

III.5.1. Testing Without Repeated Measures

For the simulation for the test for $\beta_0 = 0$, we used the following conventions. We used 31 values of γ in the range $[-3,3]$. The variable $Z = \mathrm{Uniform}[-2,2]$, while the function $\theta_0(z) = \sin(2z)$ is distinctly nonlinear. In keeping with our data example, the sample size was $n = 1,400$.

We generated X in three ways.

- As a bivariate standard normal random variable.
- $X = (X_1, X_2)$ where $X_1 = \text{Bernoulli}(0.6)$ and $X_2 = \text{Normal}(0, 1)$.
- As two dummy variables. Thus, we first generated a standard normal random variable r, and $X_1 = I(r < -0.4)$ while $X_2 = I(r > 0.4)$.

We set $\beta_0 = c(1, 1)^{\mathrm{T}}$, where we set $c = 0.0, 0.01, ..., 0.15$ for power calculations. The true value of γ was varied: $\gamma_{\text{true}} = 0, 1, 2$. We ran simulations both with and without additional covariates S: in the former case, we set S to be generated from a univariate Normal$(0, 1)$ distribution and use $\eta_0 = 1$.

For each scenario, we ran $1,000$ simulated data sets. To estimate the significance level, we applied the method in Section III.3.4 with $1,500$ replications. The Epanechnikov kernel was used to carry out the computation. We used different bandwidth of the form $h = \kappa \times \text{std}(Z)n^{-1/5}$ with different values of κ ranging from 0.5 to 2. The results are very similar in each of those cases and hence we report the results for $\kappa = 1$ only. The results are displayed in Figures 6, 7 and 8. There three main conclusions are clear:

- The test level of our method is near-nominal, being 0.051 without S and 0.057 with S in the model.
- For the main effects model with $\gamma_{\text{true}} = 0$, our maximized score-type test loses only modest power compared to the efficient (in this case) main effects score test.
- When there are interactions, our methods greatly dominate the main effects score test as γ_{true} increases.

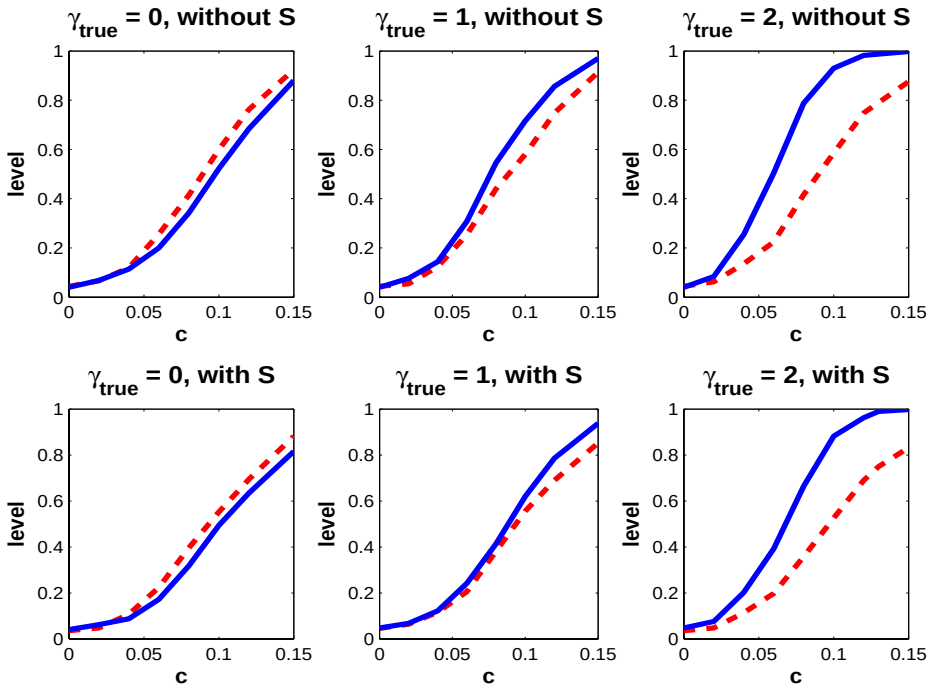

Fig. 6 Results of the simulation for testing whether $\beta = 0$ as described in Section III.5.1 using Kernel based calculations. Here X is a bivariate standard normal random variable. Solid line is our method, while the dashed line is the naive test which assumes $\gamma = 0$. The top rows gives power where there are no additional covariates S, while the bottom row includes a covariate S. The true value used was $\beta = c(1,1)^{\mathrm{T}}$: the horizontal axis plots the value of c and the vertical axis plots the corresponding power.

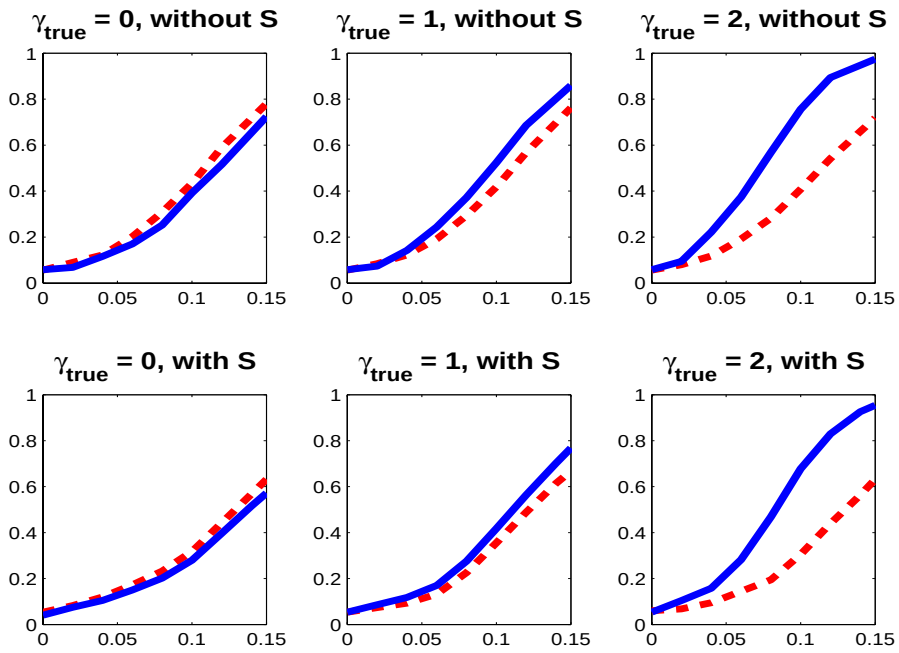

Fig. 7 Results of the simulation for testing whether $\beta = 0$ as described in Section III.5.1 using Kernel based calculations. Here $X = (X_1, X_2)$ where $X_1 = \text{Bernoulli}(0.6)$ and $X_2 = \text{Normal}(0, 1)$. Solid line is our method, while the dashed line is the naive test which assumes $\gamma = 0$. The top rows gives power where there are no additional covariates S, while the bottom row includes a covariate S. The true value used was $\beta = c(1, 1)^{\mathrm{T}}$: the horizontal axis plots the value of c and the vertical axis plots the corresponding power.

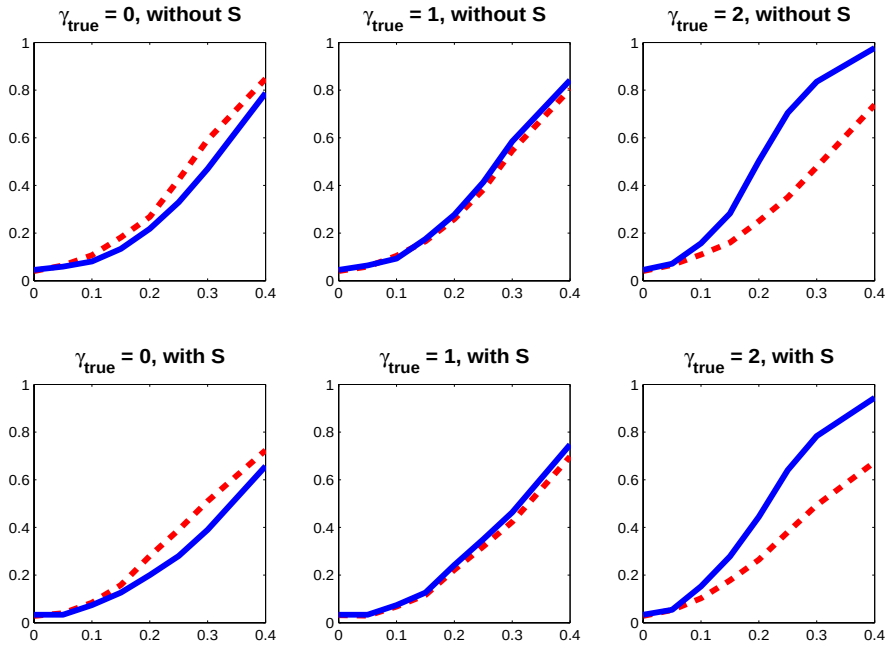

Fig. 8 Results of the simulation for testing whether $\beta = 0$ as described in Section III.5.1 using Kernel based calculations. Here $X = (X_1, X_2)$ is two dummy variables. Thus, we first generated a standard normal random variable r, and $X_1 = I(r < -0.4)$ while $X_2 = I(r > 0.4)$. Solid line is our method, while the dashed line is the naive test which assumes $\gamma = 0$. The top rows gives power where there are no additional covariates S, while the bottom row includes a covariate S. The true value used was $\beta = c(1, 1)^{\mathrm{T}}$: the horizontal axis plots the value of c and the vertical axis plots the corresponding power.

For comparison purposes, we repeated the simulation using penalized B-spline regression, using a second-order B-spline with 10 basis functions and with a second-order difference penalty. The smoothing parameter was chosen by GCV. The results were very similar to those obtained for kernel methods. The near equivalence of kernel and spline methods here is no surprise, since there is evidence in Gaussian cases that smoothing splines are equivalent to kernel methods (Silverman, 1984; Lin, et al., 2004). Recently, Li and Ruppert (2008) showed that penalized B-spline regression is also asymptotically equivalent to kernel regression methods in the Gaussian case.

III.5.2. Testing With Repeated Measures

We use the following setup for our simulations for testing $\beta_0 = 0$. We generate samples from the partially linear Gaussian repeated measures model: for $i = 1, \ldots, n$ and $j = 1, \ldots, J$,

$$Y_{ij} = X_{ij}^{\mathrm{T}}\beta_0 + \theta_0(Z_{ij})(1 + \gamma X_{ij}^{\mathrm{T}}\beta_0) + \epsilon_{ij},$$

with $n = 200$ and $J = 3$, where we take the true value of the parameter to be $\beta_0 = c(1, -1)^{\mathrm{T}}$ and set $c = 0, 0.01, \ldots, 0.06$ for power calculation. We set $\theta_0(z) = \sin(2z)$ to be the true function. We generated X from the standard bivariate normal distribution and Z from the Uniform$[-2, 2]$ distribution. The error vectors $(\epsilon_1, \ldots, \epsilon_J)^{\mathrm{T}}$ are generated from a multivariate normal distribution with covariance matrix $\Sigma = I + 0.6(\mathbf{1}\mathbf{1}^{\mathrm{T}} - I)$.

We use 11 values of γ in $[0, 2]$ to compute the test statistic. The true values of γ that are used to generate the data are taken to be $\gamma_{\mathrm{true}} = 0, 1, 2$. As in the previous simulation, we use the Epanechnikov kernel with bandwidth $h = \kappa \times \mathrm{std}(Z)n^{-1/5}$ where the value of κ ranged from 0.5 to 2. In this case also, we observe that the results are very similar for each of the bandwidth choices and hence we report the

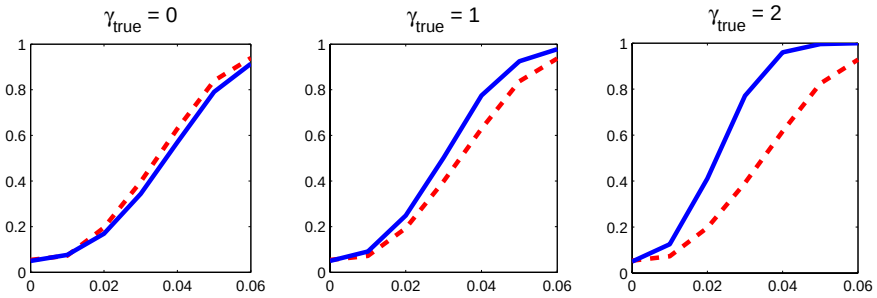

Fig. 9 Results of the simulation for testing whether $\beta_0 = 0$, as described in Section III.5.2. Solid line is our method, while the dashed line is the usual test. The true value used was $\beta = c(1, -1)^{\mathrm{T}}$: the horizontal axis plots the value of c and the vertical axis plots the corresponding power.

results for $\kappa = 1$. We generate 1,000 data sets for each case and for each data set we apply our method using 1,000 replications. The results are given in Figure 9. The level of our test is 0.051, which is very close to the nominal level of 0.05. It is evident that while our test loses very little power when $\gamma_{\text{true}} = 0$, it achieves great power gain in the presence of interaction as seen in cases where $\gamma_{\text{true}} = 1, 2$.

We redid the simulation using B-splines with 10 basis functions where the penalty parameter is estimated at the null model using GCV. The results are nearly identical to Figure 9, as one would expect in the Gaussian case.

III.6. Data Analysis

Chatterjee et al illustrated application of their methodology using a case-control study for investigation of association between colorectal adenoma, a precursor of colorectal cancer and *NAT2*, a candidate gene that is known to play important role in detoxification of certain aromatic carcinogen present in cigarette smoke. The study involved

about 700 cases and 700 controls who were genotyped for six known functional polymorphisms related to *NAT2* acetylation activity. The genotype data were used to construct diplotype information, i.e. the pair of haplotypes the subjects carried along their pair of homologous chromosomes. The frequency distribution of these diplotypes and associated acetylation phenotypes are shown in Table 4 of Chatterjee et al. In principle, the diplotypes are not observed directly and we can only assign diplotypes based on the unphased genotype data. However, in many instances such as this example, when we have very tightly linked SNPs, the phase ambiguity is often minimal, i.e., one can assign a very large proportion ($> 90\%$) of the subjects a specific diplotype with a very high probability (> 0.95). In such cases, it is easier to just remove those few people for whom the diplotypes are more uncertain and assume that for the rest of the people the diplotypes are known. In our data set, we removed a small number of people whose haplotypes were quite uncertain.

Chatterjee et al considered an omnibus test that can account for interaction of *NAT2* history with smoking history, defined as ever, former or never smokers. We consider a similar application involving *NAT2* diplotypes, but model the effect of CIG_STOP (years since stopping smoking) in a continuous fashion with nonparametric regression among smokers. Because of a few high-leverage values, we censored CIG_STOP at 45. In our analysis, the co-factor S included gender and 3 indicator dummy variables for age-level: between 60 and 65, between 65 and 70, and more than 70. For modeling the effect of *NAT2* diplotypes, we considered a series of 14 different analysis where in the k^{th} analysis we compare the risk associated with the k ($k = 1, \ldots, 14$) most common diplotypes in reference to the rest, with the associated design matrix X_k being defined by k corresponding dummy variables. To account for non-smokers in this analysis, we defined δ to be the indicator of smoking (ever vs never) and considered the following

Table 1. Significance levels (p-values) of the test for genetic effects in a regression model in which Z is years since stopped smoking. Age category and gender were modeled additively and parametrically. The analysis is done for the most common diplotype, the most common two diplotypes, and so on. The nonparametric regression was done using penalized order-2 B-splines with 10 segments, with penalization done via GCV.

	Our Method		$\gamma = 0$	
diplotypes	Test	p-value	Test	p-value
1	11.4	0.001	3.3	0.066
2	13.9	0.003	5.7	0.055
3	16.6	0.002	9.8	0.016
4	16.7	0.007	9.8	0.041
5	19.5	0.007	11.3	0.045
6	19.7	0.017	11.4	0.087
7	20.0	0.021	12.3	0.098
8	21.3	0.025	13.1	0.111
9	24.1	0.015	14.2	0.116
10	25.2	0.016	15.3	0.120
11	25.2	0.027	15.4	0.180
12	25.6	0.036	15.4	0.214
13	25.9	0.055	15.8	0.262
14	26.7	0.066	16.6	0.279

model:

$$\mathrm{pr}(D = 1|X, S, Z) = H\{(1 - \delta)\beta_0 + S^{\mathrm{T}}\beta_1 + X^{\mathrm{T}}\beta_2 + \delta\theta(Z) + \gamma\delta X^{\mathrm{T}}\beta_2\theta(Z)\}.$$

Modifying our methods to handle this slightly more complex model is straightforward: details are available from the authors.

Table 1 compares results of the proposed method for testing $\beta_2 = 0$ based on model (3.12) with those for a test for only the corresponding main effects of the diplotypes, ignoring *NAT2*-smoking interaction, i.e. assuming $\gamma = 0$. We observe that in each analysis, stronger evidence of association is seen in our new test. For example, when the 12 most common diplotypes were used, our method had a significance level of 0.036 versus a significance level of 0.214 for the main-effect based test. Interestingly, when all 14 common diplotypes are used, the significance level of

the proposed test was 0.066, quite close to that for the test used by Chatterjee et al, also using all the 14 diplotypes, but accounting for interaction with the categorical smoking history variable defined as never, former or current smoker.

CHAPTER IV

ESTIMATION VIA CORRECTED SCORES IN GENERAL SEMIPARAMETRIC REGRESSION MODELS WITH ERROR-PRONE COVARIATES

IV.1. Introduction

Ma and Carroll (2006), building upon work of Tsiatis and Ma (2004), develop a functional methodology for semiparametric measurement error models when a covariate measured precisely is modeled nonparametrically. Specifically, a response Y given covariates (X, S, Z) has the loglikelihood function $\mathcal{L}\{Y, X, S, \mathcal{B}_0, \theta_0(Z)\}$ for an unknown parameter $\mathcal{B}_0$ and an unknown function $\theta_0(\cdot)$. In the measurement error problem, X is unobserved, and instead they suppose that W is observed, where they assume that the distribution of W given X is specified parametrically. For example, the case considered here is the standard additive measurement error model

$$W_i = X_i + U_i, \quad U_i = \text{Normal}(0, \Sigma_u), \tag{4.1}$$

where U_i is independent of (Y_i, X_i, S_i, Z_i). Equation (4.1) may hold after a data transformation.

The method of Ma and Carroll works as follows. First, they specify a parametric distribution for X given (S, Z), with density function $p_{\text{c}}(x|s, z, \xi_{\text{latent}})$, where "c" stands for "conjectured". They assume that ξ_{latent} can be estimated at the rate $n^{1/2}$. Their method has two important properties:

- It is a *functional* measurement error method (Carroll, et al., 2006), in the sense that their estimates of $\mathcal{B}_0$ and $\theta_0(\cdot)$ are consistent and asymptotically normally

distributed no matter what the distribution for X is. In particular, it is consistent and asymptotically normally distributed even when the conjectured density function $p_{\mathrm{c}}(x|s, z, \xi_{\mathrm{latent}})$ for X is misspecified.

- If the density function $p_{\mathrm{c}}(x|s, z, \xi_{\mathrm{latent}})$ for X is specified correctly, their estimate of $\mathcal{B}_0$ is semiparametric efficient among all functional measurement error methods.

Despite these strengths and great generality, as described in detail Section IV.2.1, the Ma and Carroll method suffers from the fact that its implementation requires solving integral equations, which may be problematic for cases of large measurement error (Y. Ma, personal communication) and is not really practical for multivariate X, e.g., in longitudinal data settings. In addition, Ma and Carroll use a discrete approximation to solve the integral equations which leads their solution to be only approximately consistent.

In this chapter, we develop an alternative functional measurement error model for the standard additive measurement error model (4.1). Our method is based upon the idea of Monte Carlo corrected scores (Novick and Stefanski, 2002). While it uses complex-value arithmetic, our method is easily implemented, does not require the solution of integral equations, and its theory falls into the framework of standard profiling methods for criterion functions in semiparametric problems, thus for example yielding standard errors of parameter estimates as a by-product. One important aspect of our method is that despite being a corrected score based method it does not require the exact form of the corrected score. As long as one knows the log-likelihood (or criterion function) one can compute the "Monte Carlo corrected score" and our method is applicable even to those cases where the exact form of the corrected score

can not be derived.

Within the additive normal measurement error context, our method also applies to much more complex models than those considered by Ma and Carroll. For example, in longitudinal and repeated measures data, the underlying loglikelihood function might be of the form

$$\mathcal{L}\{Y_i, \widetilde{X}_i, \widetilde{S}_i, \mathcal{B}_0, \theta_0(Z_{i1}), ..., \theta_0(Z_{im})\}, \tag{4.2}$$

for a parameter $\mathcal{B}_0$, where the key is that unlike in the Ma and Carroll context, the nonparametric component $\theta_0(\cdot)$ is evaluated multiple times per individual, see Lin and Carroll (2006) for many examples and the theory and methods when X is observable. We use the notation $\widetilde{X}$ to indicate the possibility of a vector of covariates evaluated repeatedly, e.g.., $(X_{i1}, ..., X_{im})$. Once again, our method in this general context is a functional method, based on a criterion function, and with a theoretical development that follows easily from existing literature. It is worth pointing out that in (4.2), the Ma and Carroll method is not really practical. Suppose the covariate measures with error is time varying, so that $\widetilde{X}_i = (X_{i1}, \ldots, X_{im})$, where X_{ij} is of dimension p. The the integral equation to be solved is of dimension $m \times p$, clearly infeasible in many applications. In contrast, our method is easily applied.

An outline of this chapter is as follows. In Section IV.2, we review the basic method of Ma and Carroll, define our method as it applies to their problem, and derive its asymptotic theory. The last step is particularly easy because our formulation sits within standard semiparametric modeling for criterion functions. In Section IV.3, we consider the multivariate response and predictor partially linear measurement error model, an example of (4.2), showing that our method is as efficient numerically as the

semiparametric efficient functional method, the derivation of which is new. In Section IV.4, we redo the simulation of Ma and Carroll in the logistic partially linear model with a quadratic effect in X, showing that our method is as efficient numerically as theirs, even with larger measurement error, while being computationally much easier. In Section IV.5, we apply our method to Nevada Test Site (NTS) Thyroid Disease Study data and report the results. All technical details are collected in an appendix.

IV.2. Methodology

This section considers problems in which the likelihood function for the semiparametric model is of the form $\mathcal{L}\{Y, X, S, \mathcal{B}_0, \theta_0(Z)\}$. Section IV.3 discusses the more complex model (4.2).

IV.2.1. The Ma and Carroll Method

The method of Ma and Carroll (2006) works as follows. Let $\mathcal{Y} = (Y, W, S, Z)$ be the observed data. Let $\mathcal{B}_0$ be the true parameter in this model, and $\theta_0(z)$ the true function. The method requires a conjectured density for X given (S, Z), $p_{\rm c}(x|S, Z, \xi_{\rm latent})$ based upon a parameter $\xi_{\rm latent}$ that can be estimated with rate $n^{1/2}$. Let $\mathcal{S}_{\mathcal{B}}(\cdot)$ and $\mathcal{S}_{\theta}(\cdot)$ be the loglikelihood scores of the observed data for $\mathcal{B}$ and θ, respectively, computed under $p_{\rm c}(x|S, Z, \xi_{\rm latent})$, the conjectured model for X given (S, Z). Let expectations computed under the true model and the conjectured model for X given (S, Z) be denoted by "E" and "E_*", respectively. Then there exist functions $a_{\mathcal{B}}(X, S, Z)$ and $a_{\theta}(X, S, Z)$ such that

$$E\{\mathcal{S}_{\mathcal{B}}(\cdot)|X, S, Z\} = E[E_*\{a_{\mathcal{B}}(X, S, Z)|\mathcal{Y}\}|X, S, Z]; \tag{4.3}$$

$$E\{\mathcal{S}_\theta(\cdot)|X,S,Z\} = E[E_*\{a_\theta(X,S,Z)|\mathcal{Y}\}|X,S,Z]. \tag{4.4}$$

Ma and Carroll then form estimating functions $\mathcal{L}_\mathcal{B}(\cdot) = \mathcal{S}_\mathcal{B}(\cdot) - E_*\{a_\mathcal{B}(X,S,Z)|\mathcal{Y}\}$ and $\Psi_\theta(\cdot) = \mathcal{S}_\theta(\cdot) - E_*\{a_\theta(X,S,Z)|\mathcal{Y}\}$. These estimating functions are unbiased at $\mathcal{B}_0$ and $\theta_0(\cdot)$, i.e., have mean zero, even if the conjectured model for X given Z is incorrect. They then propose a backfitting algorithm similar to one described below in Section IV.2.3 for estimating $\mathcal{B}_0$ and $\theta_0(\cdot)$, but based upon the estimating functions $\mathcal{L}_\mathcal{B}(\cdot)$ and $\Psi_\theta(\cdot)$.

The main issue with implementation of the Ma and Carroll approach lies in solving the integral equations (4.3)-(4.4) while at the same time implementing backfitting. They propose to approximate the integrals by discretizing the support of X into a finite set $(x_1, ..., x_J)$. Let $p_{X,S,Z|\mathcal{Y}}(\cdot)$ be the conjectured conditional density of (X,S,Z) given $\mathcal{Y}$, and denote $p_i(\mathcal{Y}) = p_{X,S,Z|\mathcal{Y}}(x_i,S,Z|\mathcal{Y})$. In this case, (4.3) becomes

$$\sum_{i=1}^{J} a_\mathcal{B}(x_i,S,Z)E_*\{p_i(\mathcal{Y})|X,S,Z\} = E_*\{S_\mathcal{B}(\mathcal{Y})|X,S,Z\}. \tag{4.5}$$

Setting $X = x_1,\ldots,x_J$ in (4.5) will thus provide J linear equations, and then one subsequently solves the J-equation linear system to obtain $a_\theta(x_i,S,Z)$, $i = 1,\ldots,J$. A similar calculation is done for $a_\theta(x_i,S,Z)$.

The implementation difficulties in the Ma and Carroll approach are now clear. It is not obvious how one should choose the number of discretization points J, and presumably J will need to become fairly large if X is multivariate. In addition, the various conditional expectations in (4.5) may be more-or-less difficult to compute accurately. In contrast, the semiparametric-MCCS is simple to implement and, as we will see in our numerical examples, performs as well as Ma and Carroll method, even when the

measurement error in the model is relatively large.

IV.2.2. Semiparametric Monte-Carlo Corrected Scores

Let $K(\cdot)$ be a symmetric density function with finite support, let h be a bandwidth and define $K_h(v) = h^{-1}K(v/h)$.

If the true covariate, X, were observed then a profile likelihood estimation procedure for $\mathcal{B}_0$ and $\theta_0(\cdot)$ is discussed in Lin and Carroll (2006): for a fixed value of $\mathcal{B} = \mathcal{B}^*$, compute $\widehat{\theta}(z, \mathcal{B}^*)$ by solving the local-linear estimating equations

$$\sum_{i=1}^{n} K_h(Z_i - z)\{1, (Z_i - z)/h\}^{\mathrm{T}} \mathcal{L}_\theta \left\{Y_i, X_i, S_i, \mathcal{B}^*, \widehat{\alpha}_0 + \widehat{\alpha}_1(Z_i - z)/h\right\} = 0 \qquad (4.6)$$

for $\widehat{\alpha}_0$ and setting $\widehat{\theta}(z, \mathcal{B}^*) = \widehat{\alpha}_0$. Then, maximize

$$\sum_{i=1}^{n} \mathcal{L}\Big\{Y_i, X_i, S_i, \mathcal{B}, \widehat{\theta}(Z_i, \mathcal{B})\Big\} \qquad (4.7)$$

in $\mathcal{B}$ and set the maximizer $\widehat{\mathcal{B}}$ to be the estimate of $\mathcal{B}_0$.

However, in the presence of measurement errors, we observe W_i instead of X_i. Hence the estimation procedure given by (4.6) and (4.7), when applied based on W instead of X, produce biased estimates.

IV.2.3. Corrected Score Estimation

We follow the idea of Novick and Stefanski (2002) to solve this problem using methods based on corrected scores. Consider the complex variate

$$\widetilde{W}_{ib} \quad = \quad W_i + \iota V_{ib}, \quad b = 1, \ldots, B,$$

where $\iota = \sqrt{-1}$ and V_{ib} is a normal random vector generated by computer with mean 0 and covariance matrix Σ_u. Stefanski and Cook (1995) showed that if $f(\cdot)$ is an entire function then under integrability conditions

$$\mathrm{E}\{f(\widetilde{W}_{ib})|X_i\} = \mathrm{E}[\mathrm{Re}\{f(\widetilde{W}_{ib})\}|X_i] = f(X_i).$$

Assume that $\mathcal{L}(\cdot)$ is an entire function of its second argument. We define the corrected score as

$$\mathcal{R}_i(\cdot) \quad = \quad B^{-1} \sum_{b=1}^{B} \mathrm{Re}[\mathcal{L}\{Y_i, \widetilde{W}_{ib}, S_i, \mathcal{B}_0, \theta_0(Z_i)\}].$$

Note that, $\mathcal{R}(\cdot)$ is a real valued function of real arguments $\{Y_i, W_i, S_i, \widetilde{V}_i, \mathcal{B}_0, \theta_0(Z_i)\}$, where $\widetilde{V}_i = (V_{i1}, \ldots, V_{iB})$. Define $\mathcal{R}_\theta(\cdot)$ and $\mathcal{R}_\mathcal{B}(\cdot)$ as the derivatives of $\mathcal{R}(\cdot)$ with respect to θ and $\mathcal{B}$, respectively. Define $G_i(z,h) = \{1, (Z_i - z)/h\}$.For a fixed $\mathcal{B}^*$, we propose to estimate $\theta_0(z)$ by solving

$$0 = n^{-1} \sum_{i=1}^{n} K_h(Z_i - z) G_i(z,h)^{\mathrm{T}} \mathcal{R}_\theta \left\{ Y_i, W_i, S_i, \widetilde{V}_i, \mathcal{B}^*, \widehat{\alpha}_0 + \widehat{\alpha}_1 (Z_i - z)/h \right\} \qquad (4.8)$$

for $\widehat{\alpha}_0$ and setting $\widehat{\theta}(z, \mathcal{B}^*) = \widehat{\alpha}_0$.

There are two methods to estimate $\mathcal{B}$:

1. Profiling maximizes $n^{-1} \sum_{i=1}^{n} \mathcal{R} \left\{ Y_i, W_i, S_i, \widetilde{V}_i, \mathcal{B}, \widehat{\theta}(Z_i, \mathcal{B}) \right\}$ in $\mathcal{B}$. If we define

$\widehat{\theta_{\mathcal{B}}}(z,\mathcal{B})$ to be the derivative of $\widehat{\theta}(z,\mathcal{B})$ with respect to $\mathcal{B}$, then profiling solves

$$\begin{aligned} 0 &= n^{-1}\sum_{i=1}^{n}\Big[\mathcal{R}_{\mathcal{B}}\Big\{Y_i,W_i,S_i,\widetilde{V}_i,\mathcal{B},\widehat{\theta}(Z_i,\mathcal{B})\Big\} \\ &\qquad +\mathcal{R}_{\theta}\Big\{Y_i,W_i,S_i,\widetilde{V}_i,\mathcal{B},\widehat{\theta}(Z_i,\mathcal{B})\Big\}\widehat{\theta_{\mathcal{B}}}(Z_i,\mathcal{B})\Big]. \end{aligned} \tag{4.9}$$

Call the solution $\widehat{\mathcal{B}}_{\text{pf}}$.

2. Backfitting estimates $\mathcal{B}$ iteratively. Based on the current estimate, $\widehat{\mathcal{B}}_{\text{cur}}$, backfitting solves

$$n^{-1}\sum_{i=1}^{n}\mathcal{R}_{\mathcal{B}}\left\{Y_i,W_i,S_i,\widetilde{V}_i,\mathcal{B},\widehat{\theta}(Z_i,\widehat{\mathcal{B}}_{\text{cur}})\right\}=0. \tag{4.10}$$

Let $\widehat{\mathcal{B}}_{\text{bf}}$ be the backfitting estimator.

It is important to note that while $\mathcal{R}(\cdot)$ may not be a valid loglikelihood function, it is a *criterion function* in the sense of Lin and Carroll (2006). Also note that the results given in Lin and Carroll for the profiling and backfitting methods are true for any *criterion function* as long as various conditions are satisfied: these conditions translate to A1-A5, given in the Appendix.

IV.2.4. Asymptotic Properties

In this section, we derive the asymptotic properties of our method in the case that the measurement error covariance matrix Σ_u is known, see Section IV.2.6 for the case that it is estimated. We make use of the results of Lin and Carroll (2006). Define

$$\begin{aligned} \theta_{\mathcal{B}}(z,\mathcal{B}_0) &= -\frac{E[\mathcal{R}_{\theta\mathcal{B}}\{Y,W,S,\widetilde{V},\mathcal{B}_0,\theta(Z)\}|Z=z]}{E[\mathcal{R}_{\theta\theta}\{Y,W,S,\widetilde{V},\mathcal{B}_0,\theta(Z)\}|Z=z]}; \\ \Omega(z) &= f_Z(z)E\{\mathcal{R}_{\theta\theta}(\cdot)|Z=z\}; \end{aligned}$$

$$\mathcal{M} \;=\; E\{\mathcal{R}_{\mathcal{B}\mathcal{B}}(\cdot) + \mathcal{R}_{\mathcal{B}\theta}(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z, \mathcal{B}_0)\}.$$

Then the following result is a direct consequence of the main results of Lin and Carroll (2006).

Result 7 Assume that $(Y_i, Z_i, W_i, S_i), i = 1, \ldots, n$ are independent and identically distributed and $\widehat{\mathcal{B}}_{\mathrm{pf}}$ and $\widehat{\theta}(\cdot)$ are estimates obtained from (4.8) and (4.9). Also assume that $h \propto n^{-c}$ with $1/5 \leq c \leq 1/3$. Let $\theta^{(2)}(z)$ be the second derivative of $\theta(z)$ and $\phi_2 = \int z^2 K(z)dz$. Then,

$$\begin{aligned}
\widehat{\theta}(z, \widehat{\mathcal{B}}_{\mathrm{pf}}) - \theta_0(z) &= (h^2/2)\phi_2\theta^{(2)}(z) - n^{-1}\sum_{i=1}^{n} K_h(Z_i - z)\mathcal{R}_{i\theta}(\cdot)/\Omega(z) \\
&\quad -\theta_{\mathcal{B}}(z_0, \mathcal{B}_0)^{\mathrm{T}}\mathcal{M}^{-1}n^{-1}\sum_{i=1}^{n}\{\mathcal{R}_{i\mathcal{B}}(\cdot) + \mathcal{R}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i, \mathcal{B}_0)\} + o_p(n^{-1/2}); \\
n^{1/2}(\widehat{\mathcal{B}}_{\mathrm{pf}} - \mathcal{B}_0) &= -\mathcal{M}^{-1}n^{-1/2}\sum_{i=1}^{n}\{\mathcal{R}_{i\mathcal{B}}(\cdot) + \mathcal{R}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i, \mathcal{B}_0)\} + o_p(n^{-1/2}) \\
&\Rightarrow \quad \mathrm{Normal}(0, \mathcal{M}^{-1}\mathcal{F}\mathcal{M}^{-1}),
\end{aligned}$$

where $\mathcal{F} = \mathrm{cov}[\mathcal{R}_{\mathcal{B}}(\cdot) + \mathcal{R}_{\theta}(\cdot)\theta_{\mathcal{B}}(Z, \mathcal{B}_0)]$.

Result 8 Make the same assumptions as in Theorem 7 but assume $nh^4 \to 0$. Then the backfitting estimator $\mathcal{B}_{\mathrm{bf}}$ has the same limiting distribution as the profile estimator.

Remark 10 Estimation of the asymptotic variance of $\mathcal{B}_0$ is a straightforward exercise. To construct such estimates, all the expectations in the definitions of $\mathcal{M}$ and $\mathcal{F}$ are replaced by sums and all the regression functions are replaced by kernel estimates. Alternatively, one can use the bootstrap: Chen, et al. (2003) justify the use of the

bootstrap for estimating $\mathcal{B}_0$ in semiparametric models with general criterion functions.

IV.2.5. Special Case: Partially Linear Model

One common but important example is the partially linear measurement error model. Estimation in the partially linear model with error prone covariates are described in Liang, Hardle and Carroll (1999). In this section we derive the asymptotic distribution of our estimates explicitly and compare our estimates to that of Liang, et al.

Consider the model

$$Y_i = X_i^{\mathrm{T}}\gamma + \theta(Z_i) + \epsilon_i,$$

for $i = 1, \ldots, n$. Assume that $\epsilon = \text{Normal}(0, \sigma^2)$. Instead of observing X, we observe $W_i = X_i + U_i$, where U_i is independent of (X_i, Z_i, Y_i) and has a $\text{Normal}(0, \Sigma_{uu})$ distribution. Assume that Σ_{uu} is known. Define $\beta = (\gamma^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$. Then the loglikelihood is

$$\mathcal{L}\{Y, X, \theta(Z), \beta\} \;\; = \;\; -\log(\sigma^2)/2 - (2\sigma^2)^{-1}\{Y - X^{\mathrm{T}}\gamma - \theta(Z)\}^2.$$

Define $\widetilde{W}_{ib} = W_i + \iota V_{ib}$, where $V_{ib} = \text{Normal}(0, \Sigma_{uu})$ are independent random vectors generated by computer. Let $\widetilde{V}_i = (V_{i1}, \ldots, V_{ib})$. Then, the corrected score is

$$\begin{aligned}
&\mathcal{R}\{Y, W, \widetilde{V}, \theta(Z), \beta\} \\
&\quad = -\log(\sigma^2)/2 - B^{-1}\sum_{b=1}^{B} \text{Re}[(2\sigma^2)^{-1}\{Y - (W + \iota V_b)^{\mathrm{T}}\gamma - \theta(Z)\}^2] \\
&\quad = -\log(\sigma^2)/2 - (2\sigma^2)^{-1}[\{Y - W^{\mathrm{T}}\gamma - \theta(Z)\}^2 - \gamma^{\mathrm{T}}B^{-1}\sum_{b=1}^{B} V_b V_b^{\mathrm{T}}\gamma].
\end{aligned}$$

Also, define

$$\begin{aligned}\Gamma &= E[\{X - E(X|Z)\}(\epsilon - U^{\mathrm{T}}\gamma)^2\{X - E(X|Z)\}^{\mathrm{T}}] + E(UU^{\mathrm{T}}\epsilon^2) \\ &\quad + E\{(UU^{\mathrm{T}} - \Sigma_{uu})\gamma\gamma^{\mathrm{T}}(UU^{\mathrm{T}} - \Sigma_{uu})^{\mathrm{T}}\}; \\ \mathcal{S} &= \mathrm{cov}\{X - E(X|Z)\}; \\ \tau^2 &= E\{(\epsilon - U^{\mathrm{T}}\gamma)^2 - (\sigma^2 + \gamma^{\mathrm{T}}\Sigma_{uu}\gamma)\}^2.\end{aligned}$$

Then we have the following result:

Result 9 Let $\widehat{\gamma}$ and $\widehat{\sigma}^2$ denote the estimate based on our method. Then marginally,

$$\begin{aligned}n^{1/2}(\widehat{\gamma} - \gamma) &\rightarrow \mathrm{Normal}\{0, \mathcal{S}^{-1}\Gamma\mathcal{S}^{-1} + R_1(B)\}; \\ n^{1/2}(\widehat{\sigma}^2 - \sigma^2) &\rightarrow \mathrm{Normal}\{0, \tau^2 + R_2(B)\};\end{aligned}$$

where $R_1(B) = B^{-1}\mathcal{S}^{-1}E\{(VV^{\mathrm{T}} - \Sigma_{uu})\gamma\gamma^{\mathrm{T}}(VV^{\mathrm{T}} - \Sigma_{uu})^{\mathrm{T}}\}\mathcal{S}^{-1} \rightarrow 0$ and $R_2(B) = B^{-1}\mathrm{var}\{\gamma^{\mathrm{T}}(V_bV_b^{\mathrm{T}} - \Sigma_{uu})\gamma\}] \rightarrow 0$ as $B \rightarrow \infty$.

It is important to note that $R_1(B)$ and $R_2(B)$ vanish as $B \rightarrow \infty$, giving us the exact same result as in Liang, et al. (1999).

IV.2.6. Estimation of the Error Covariance Matrix

It is straightforward to modify our results to account for estimation of the measurement error covariance matrix Σ_u. The usual way to estimate Σ_u is via replication of the W-data, so as an illustration suppose that $W_i = (W_{i(1)} + W_{i(2)})/2$, where $W_{i(j)} = X_i + U_{i(j)}$ and $U_{i(j)} = \mathrm{Normal}(0, \Sigma_u)$. Then a root-n consistent estimate $\widehat{\Sigma}_u$ of Σ_u is the sample covariance matrix of the terms $D_i = (W_{i(1)} - W_{i(2)})/2$. Let $\gamma = \mathrm{vech}(\Sigma_u)$, where "vech" is the vector half, i.e., the vector of the unique elements of Σ_u. Then with $\widehat{\gamma} = \mathrm{vech}(\widehat{\Sigma}_u)$, we have that $\widehat{\gamma} - \gamma = n^{-1}\sum_i \mathrm{vech}(D_i - \Sigma_u) + o_p(n^{-1/2})$.

Since V_{ib} can be written as $\Sigma_u^{1/2} e_i$ with $e_i = \text{Normal}(0, I)$, we can redefine the criterion function as $\mathcal{R}\{Y, W, S, \Sigma_u^{1/2}\widetilde{e}_i, \mathcal{B}, \theta(Z, \mathcal{B}, \Sigma_u)\}$. Let $\mathcal{R}_\gamma(\cdot)$ be its derivative with respect to γ. Then following Section 4 of Lin and Carroll (2006), we have the following asymptotic expansion for the profile estimator, up to terms of order $o_p(1)$,

$$\begin{aligned} n^{1/2}(\widehat{\mathcal{B}}_{\text{pf}} - \mathcal{B}_0) &= -\mathcal{M}^{-1}[n^{-1/2}\sum_{i=1}^{n}\{\mathcal{R}_{i\mathcal{B}}(\cdot) + \mathcal{R}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i, \mathcal{B}_0)\} + \mathcal{M}_{\mathcal{B}\gamma} n^{1/2}(\widehat{\gamma} - \gamma)] \\ &= -\mathcal{M}^{-1} n^{-1/2}\sum_{i=1}^{n}[\mathcal{R}_{i\mathcal{B}}(\cdot) + \mathcal{R}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i, \mathcal{B}_0) + \mathcal{M}_{\mathcal{B}\gamma}\{\text{vech}(D_i - \Sigma_u)\}], \end{aligned}$$

where

$$\mathcal{M}_{\mathcal{B}\gamma} = \text{E}\{\mathcal{R}_{i\mathcal{B}\gamma}(\cdot) + \theta_{\mathcal{B}}(Z_i, \mathcal{B}_0)\mathcal{R}_{i\theta\gamma}^{\text{T}}(\cdot)\}.$$

The covariance of the asymptotic distribution of $n^{1/2}(\widehat{\mathcal{B}}_{\text{pf}} - \mathcal{B}_0)$ follows from the above expressions and a consistent estimator of asymptotic covariance matrix can be easily constructed, see Remark 10.

IV.3. Multivariate Measurement Error Models

In longitudinal and repeated measures data, the likelihood function when X is observed is given by (4.2). Use the notation $\theta(\widetilde{Z}_i) = \{\theta(Z_{i1}), ..., \theta(Z_{im})\}^{\text{T}}$. Instead of observing X_{ij}, we observe $T_{ij} = X_{ij} + U_{ij}$. Define $\widetilde{U}_i = (U_{i1}, \ldots, U_{im})^{\text{T}}$ and assume that $\text{vec}(\widetilde{U})$ has a Normal distribution with mean zero and covariance matrix Σ_u which is assumed known, see Remark 11 below for comments. Define $\widetilde{X}$, $\widetilde{Z}$, $\widetilde{S}$ and $\widetilde{T}$ similarly. Let $\widetilde{W}_{ib} = \widetilde{T}_i + \iota\widetilde{V}_{ib}$ for $b = 1, \ldots, B$, where $\text{vec}(\widetilde{V}_{ib}) = \text{Normal}(0, \Sigma_u)$. Then the MCCS criterion function is given by

$$\mathcal{L}_*(\cdot) = B^{-1}\sum_{b=1}^{B}\text{Re}\Big[\mathcal{L}\{\widetilde{Y}, \widetilde{W}_b, \widetilde{S}, \mathcal{B}_0, \theta_0(\widetilde{Z})\}\Big]. \tag{4.11}$$

Equation (4.11) is a criterion function in the sense of Lin and Carroll (2006), and their asymptotic results then apply.

IV.3.1. Special Case: The Partially Linear Model

We illustrate this approach in the multivariate partially linear measurement error model discussed in Lin and Carroll (2006). In particular, they considered the model

$$Y_{ij} = X_{ij}^{\mathrm{T}}\beta_0 + \theta_0(Z_{ij}) + \epsilon_{ij}, \tag{4.12}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where $\widetilde{\epsilon_i} = (\epsilon_{i1}, \ldots, \epsilon_{im})^{\mathrm{T}} = \mathrm{Normal}(0, \Sigma_\epsilon)$. Let $\mathcal{B} = (\beta, \Sigma_\epsilon)$ be the parameter of interest. Then the criterion function ignoring the measurement errors is given by

$$\begin{aligned}\mathcal{L}\{\widetilde{Y}, \widetilde{X}, \mathcal{B}, \theta(\widetilde{Z})\} &= (1/2)\log\{\det(\Sigma_\epsilon^{-1})\} \\ &\quad -(1/2)\{\widetilde{Y} - \widetilde{X}\beta - \theta(\widetilde{Z})\}^{\mathrm{T}}\Sigma_\epsilon^{-1}\{\widetilde{Y} - \widetilde{X}\beta - \theta(\widetilde{Z})\}.\end{aligned}$$

The Monte-Carlo Corrected Scores criterion function is given by

$$\begin{aligned}\mathcal{R}(\cdot) &= B^{-1}\sum_{b=1}^{B}\mathrm{Re}[\mathcal{L}\{\widetilde{Y}, \widetilde{W}_b, \mathcal{B}, \theta(\widetilde{Z})\}] \\ &= (1/2)\log\{\det(\Sigma_\epsilon^{-1})\} - (1/2)\{\widetilde{Y} - \widetilde{T}\beta - \theta(\widetilde{Z})\}^{\mathrm{T}}\Sigma_\epsilon^{-1}\{\widetilde{Y} - \widetilde{T}\beta - \theta(\widetilde{Z})\} \\ &\quad +(1/2)\beta^{\mathrm{T}}(B^{-1}\sum_{b=1}^{B}\widetilde{V}_b^{\mathrm{T}}\Sigma_\epsilon^{-1}\widetilde{V}_b)\beta.\end{aligned}$$

The backfitting algorithm is easy to apply in this case. Given the current estimates, $\widehat{\mathcal{B}}_{\mathrm{cur}} = (\widehat{\beta}_{\mathrm{cur}}, \widehat{\Sigma}_{\epsilon,\mathrm{cur}})$, the new estimates are given by

$$\begin{aligned}\widehat{\beta}_{\mathrm{new}} &= \Big[n^{-1}\sum_{i=1}^{n}\{\widetilde{T}_i^{\mathrm{T}}\widehat{\Sigma}_{\epsilon,\mathrm{cur}}^{-1}\widetilde{T}_i - B^{-1}\sum_{b=1}^{B}(\widetilde{V}_{ib}^{\mathrm{T}}\widehat{\Sigma}_{\epsilon,\mathrm{cur}}^{-1}\widetilde{V}_{ib})\}\Big]^{-1} \\ &\quad \times n^{-1}\sum_{i=1}^{n}\widetilde{T}_i^{\mathrm{T}}\widehat{\Sigma}_{\epsilon,\mathrm{cur}}^{-1}\{\widetilde{Y}_i - \widehat{\theta}(\widetilde{Z}_i, \widehat{\mathcal{B}}_{\mathrm{cur}})\};\end{aligned}$$

$$\widehat{\Sigma}_{\epsilon,\text{new}} = n^{-1}\sum_{i=1}^{n}\Big[\{\widetilde{Y}_i - \widetilde{T}_i\widehat{\beta}_{\text{cur}} - \widehat{\theta}(\widetilde{Z}_i, \widehat{\mathcal{B}}_{\text{cur}})\}\{\widetilde{Y}_i - \widetilde{T}_i\widehat{\beta}_{\text{cur}} - \widehat{\theta}(\widetilde{Z}_i, \widehat{\mathcal{B}}_{\text{cur}})\}^{\text{T}} - B^{-1}\sum_{b=1}^{B}(\widetilde{V}_{ib}\widehat{\beta}_{\text{cur}}\widehat{\beta}_{\text{cur}}^{\text{T}}\widetilde{V}_{ib}^{\text{T}})\Big].$$

Profile pseudolikelihood estimates are also easily constructed. Let $\mathcal{S}$ be a smoother matrix as in Lin et al. (2004) and define $\mathcal{Y} = (Y_{11}, \cdots, Y_{nm})^{\text{T}}$ and $\mathcal{T} = (\widetilde{T}_1^{\text{T}}, \cdots, \widetilde{T}_n^{\text{T}})^{\text{T}}$. Let $\mathcal{T}_* = (I - \mathcal{S})\mathcal{T}$, $\mathcal{Y}_* = (I - \mathcal{S})\mathcal{Y}$ and $\widetilde{\Sigma}_\epsilon = I_n \bigotimes \Sigma_\epsilon$. Then for given Σ_ϵ, the profile estimate of β is given by

$$\{\mathcal{T}_*^{\text{T}}\widetilde{\Sigma}_\epsilon^{-1}\mathcal{T}_* - \sum_i(B^{-1}\sum_b \widetilde{V}_{ib}^{\text{T}}\Sigma_\epsilon^{-1}\widetilde{V}_{ib})\}^{-1}\mathcal{T}_*^{\text{T}}\widetilde{\Sigma}_\epsilon^{-1}\mathcal{Y}_*.$$

A simple estimate of Σ_ϵ is to form the working independence estimate of β and to apply the above equation for $\widehat{\Sigma}_{\epsilon,new}$.

Remark 11 Estimation of the error covariance matrix Σ_u and its impact on limiting distribution theory for estimation of $\mathcal{B}_0$ is described in Section IV.2.6.

Remark 12 Note that as $B \to \infty$, our estimators converges to those given in Lin and Carroll (2006).

In fact, under the assumption that X is generated from a Gaussian distribution, Lin and Carroll's procedure (equivalently, our method with $B \to \infty$) performs very similar to the semiparametric efficient method, as we now show.

Suppose we assume a Gaussian distribution for X with mean μ_x and covariance matrix Σ_x; and for simplicity of notation we let β be a scalar. Then the criterion function becomes

$$\mathcal{L}_{\text{G}}(\cdot) = -(1/2)\log(|\mathcal{J}|) - (1/2)(\widetilde{Y} - \mathcal{V})^{\text{T}}\mathcal{J}^{-1}(\widetilde{Y} - \mathcal{V})$$

$$-(1/2)\log(|\Sigma_x+\Sigma_u|)-(1/2)(\widetilde{T}-\widetilde{\mu}_x)^{\mathrm{T}}(\Sigma_x+\Sigma_u)^{-1}(\widetilde{T}-\widetilde{\mu}_x),$$

where

$$\begin{aligned}\mathcal{V} &= \mathcal{V}\{\widetilde{T},\beta,\theta(\widetilde{Z}),\widetilde{\mu}_x,\Sigma_x\}=\beta\widetilde{\mu}_x+\theta(\widetilde{Z})+\beta\Sigma_x(\Sigma_x+\Sigma_u)^{-1}(\widetilde{T}-\mu_x);\\ \mathcal{J} &= \mathcal{J}(\beta,\Sigma_x,\Sigma_\epsilon)=\Sigma_\epsilon+\beta^2\Sigma_x(\Sigma_x+\Sigma_u)^{-1}\Sigma_u.\end{aligned}$$

By the results of Lin and Carroll (2006), the estimates based on $\mathcal{L}_{\mathrm{G}}(\cdot)$ are semiparametric efficient.

We compared the two methods via a simulation study. We set $m=3$ and $\beta=0.7$, $\theta(z)=0.5\cos(2z)-1$. We set Σ_ϵ to be identity matrix and $\Sigma_u=0.3I_3+0.2J_3$, where J_k denotes the $k\times k$ matrix with all the elements equal to one. We take $\Sigma_x=I_3$ and $\mu_x=(-1,-1,-1)^{\mathrm{T}}$. We generated Z from a Uniform$(0,\pi)$ distribution.

Under this setup, we generated 1000 data sets following the model given by (4.12) with $n=500$ samples each. Using each data set we estimated β using both the methods, with the bandwidth estimated as $\widehat{\sigma}_z n^{-1/3}$, where $\widehat{\sigma}_z$ is the sample standard deviation of Z. The estimates based on Lin and Carroll method and $\mathcal{L}_G(\cdot)$ have asymptotic root mean squared error (RMSE) of 0.06 and 0.057, respectively, evidence that the performance of both the methods is very close indeed.

IV.4. Simulation Study

We repeated the simulation study of Ma and Carroll (2006) to demonstrate our method. They considered the logistic regression model $\mathrm{logit}\{\mathrm{pr}(Y=1|X,Z)\}=\beta_1X+\beta_2X^2+\theta(Z)$, where $W=X+U$ and $U=\mathrm{Normal}(0,\sigma_u^2)$ with σ_u^2 known. They

Table 2. Mean, empirical standard errors (emp s.e.), root mean squared error (RMSE) and empirical coverage of 95% confidence intervals of β_1 and β_2 when $\sigma_u^2 = 0.16$. Results based on 1000 simulated data sets each with sample size $n = 500$. For each choice of $\theta(z)$, the top row presents the results using our method and bottom row shows the results from Ma and Carroll (2006).

	$\beta_1(=0.7)$				$\beta_2(=0.7)$			
	mean	emp s.e.	RMSE	95%	mean	emp s.e.	RMSE	95%
$\theta(z) = 0.5\cos(z) - 1$	0.638	0.261	0.268	0.942	0.653	0.149	0.156	0.940
	0.720	0.277	0.278	0.947	0.726	0.156	0.158	0.939
$\theta(z) = 0.5\cos(2z) - 1$	0.615	0.238	0.253	0.943	0.639	0.135	0.148	0.942
	0.727	0.276	0.277	0.951	0.728	0.155	0.158	0.940

set $\sigma_u^2 = 0.16$, $\mathcal{B} = (\beta_1, \beta_2) = (0.7, 0.7)$, and $n = 500$. We used $B = 150$ Monte Carlo iterations. They used two different forms $\theta(z)$,

1. $\theta(z) = 0.5\cos(z) - 1$

2. $\theta(z) = 0.5\cos(2z) - 1$

For both of the setups, X was generated from Normal$(-1, 1)$ and Z was generated from Uniform$(0, \pi)$.

Several bandwidth selection methods can be applied in this situation. One possibility is to use the "Direct plug-in" (DPI) method suggested by Ruppert, Sheather and Wand (1995). One can also opt for the globally fixed bandwidth $\widehat{\sigma}_z n^{-1/3}$, where $\widehat{\sigma}_z$ is the estimated standard deviation of Z. For comparison's sake, we use the global bandwidth $h_n = \widehat{\sigma}_z n^{-1/3}$, the same as in Ma and Carroll (2006). The Epanechnikov kernel was used to estimate the nonparametric function.

Technically, the logistic regression setup as described above does not fall into our framework as the logistic distribution function is not entire in the complex plane. However, Novick and Stefanski (2002) pointed out that for small measurement error

Table 3. Mean, empirical standard errors (emp s.e.), root mean squared error (RMSE) and empirical coverage of 95% confidence intervals of β_1 and β_2 using our method when $\sigma_u^2 = 0.5$. Here we generated 1000 simulated data sets each with sample size $n = 500$. Ma and Carroll (2006) did not consider this example.

	$\beta_1(= 0.7)$				$\beta_2(= 0.7)$			
	mean	emp s.e.	RMSE	95%	mean	emp s.e.	RMSE	95%
$\theta(z) = 0.5\cos(z) - 1$	0.621	0.272	0.283	0.948	0.633	0.161	0.175	0.943
$\theta(z) = 0.5\cos(2z) - 1$	0.600	0.250	0.269	0.942	0.618	0.155	0.175	0.945

variance one can still apply corrected score based methods, with only minor bias. The results are displayed in Table 2. It is evident that our method is comparable in both cases to that of Ma and Carroll in terms of mean squared error and coverage probability, albeit with the small bias expected from the fact that the logistic function is not entire on the complex plane.

The simulation was repeated for a much larger measurement error variance, $\sigma_u^2 = 0.5$ versus $\sigma_u^2 = 0.16$. The results are shown in Table 3. Again, our results indicate only a small bias and favorable coverage probability. Ma and Carroll did not report results for this situation so it is not possible to compare our method with theirs in this situation.

IV.5. Nevada Test Site Thyroiditis Data Example

In this section we apply our method to the Nevada test site (NTS) thyroid study data. The study was conducted in 1980's by the University of Utah. The original study is described in Stevens, et al. (1992), Kerber, et al. (1993) and Simon, et al. (1995). The main idea of the study was to relate the incidence of thyroid related disease to the exposure of radiation to the thyroid. In this study, $2,491$ individuals, who

were exposed to radiation as children, were tested for thyroid disease. The primary radiation exposure to the thyroid glands of these children came from the ingestion of milk and vegetables contaminated with radioactive isotopes of iodine. Recently, the dosimetry for the study was redone (Simon, et al., 2006), and the study results were reported in Lyon, et al. (2006).

Due to the fact that the actual radiation doses in foods or in the thyroid gland of the individuals are not available, the estimated radiation doses are well known to be contaminated with measurement errors. Many authors have studied and described measurement error properties and analysis in this context (Reeves, et al., 1998; Schafer, et al., 2001; Mallick, et al., 2002; Stram and Kopecky, 2003; Lubin, et al., 2004; Pierce and Kellerer, 2004; Schafer and Gilbert, 2006; Li, et al., 2007). A common approach is to build a large dosimetry model that attempts to convert the known data about above-ground nuclear testing to the radiation actually absorbed into the thyroid. Dosimetry calculations for individual subjects were based upon several variables, such as, age at exposure, gender, residence history, whether as a child the individual was breast-fed, and a diet questionnaire filled out by the parent focusing on milk consumption and vegetables. The data were then input into a complex model and for each individual, the point estimate of thyroid dose (the arithmetic mean of a lognormal distribution of dose estimates) and an associated error term (the geometric standard deviation) for the measurement error were reported.

It is typical to assume that radiation doses are estimated with a combination of Berkson measurement error and a classical type of measurement error (Reeves, et al., 1997). In the log-scale, true log-dose T is related to observed or calculated log-dose

W by a latent intermediate X via

$$\begin{aligned} T &= X + U_{\text{berk}}; \\ W &= X + U_{\text{class}}, \end{aligned}$$

where U_{berk} and U_{class} are the Berkson uncertainty and the classical uncertainty, respectively, with corresponding variances $\sigma^2_{u,\text{berk}}$ and $\sigma^2_{u,\text{class}}$ depending on the individual. It is typical to assume that the errors U_{berk} have Gaussian distributions. In the NTS study, the total uncertainty $\sigma^2_{u,\text{berk}} + \sigma^2_{u,\text{class}}$ is known but not the relative contributions. We will let 50% of the total uncertainty be classical in our illustration.

If the latent, X, could be observed then typically the total mean dose, $\exp(X + \sigma^2_{u,\text{berk}}/2)$ is taken to be the main predictor and we will take this as our target. We take the incidence of thyroiditis (inflammation of the thyroid gland), Y, as the response variable. In addition, we consider Z, the sex of the patient and A, age at exposure (standardized to have mean zero and variance 1), which are measured without measurement error. A typical parametric model relating total mean dose and gender to disease is the excess relative risk model

$$\text{pr}(Y = 1|X, Z) = H[\beta_0 + \beta_1 Z + \log\{1 + \gamma \exp(X + \sigma^2_{u,\text{berk}}/2)\}], \tag{4.13}$$

where $H(\cdot)$ is the logistic distribution function and γ is called the excess relative risk. We instead include A, age at exposure, nonparametrically in the model (4.13) as follows:

$$\text{pr}(Y = 1|X, Z) = H[\beta Z + \log\{1 + \gamma \exp(X + \sigma^2_{u,\text{berk}}/2)\} + \theta(A)], \tag{4.14}$$

where $\theta(\cdot)$ is an unknown function.

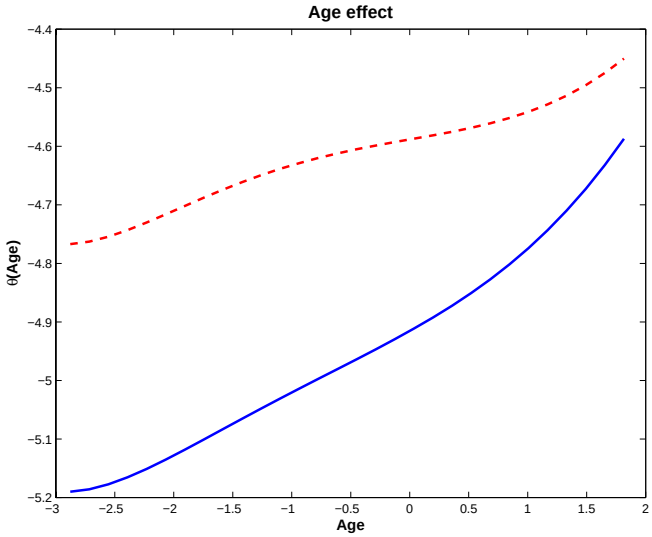

Fig. 10 Estimated age effect in the Nevada Test Site thyroiditis data. Solid line: the MCCS estimate. Dashed line: the naive estimate ignoring the presence of measurement error.

We employed our method discussed in Section IV.2 for the model given by (4.14). We compared our method to the naive method where one ignores the measurement error altogether. We used the Epanechnikov kernel and the bandwidth chosen was 1.5, but similar results were obtained for 1.0 and 2.0. For MCCS calculations, we used $B = 100$. The estimated effect of gender, $\widehat{\beta}_1 \approx 1.75$ for both the naive and MCCS method. This can be explained from the fact that gender and radiation dose for an individual are essentially independent and hence the effect of gender is not affected by measurement error in radiation dose.

The estimated value of the relative risk parameter was 8.54 for the naive method and 17.19 for the proposed MCCS method. The effect of age, A, is displayed in Figure 10 for both the naive and MCCS procedures. It is evident from the results that because

of the change in the estimate of the excess relative risk γ, there is a corresponding change in the estimated age effect when the presence of the measurement error is taken into account.

Remark 13 As noted in Section IV.4, the logistic regression setup does not fall into our framework as the logistic distribution function is not entire in the complex plane. To observe the performance of semiparametric-MCCS in this example, we compared our results to the well known SIMEX procedure (Cook and Stefanski, 1994; Stefanski and Cook, 1995). To apply SIMEX, we modeled the age effect parametrically by a quadratic polynomial. We used a quartic extrapolant for SIMEX and obtained the estimated value of the excess relative risk parameter to be 15.92. We can see that in this case the SIMEX estimate is not very different from what semiparametric-MCCS produces.

CHAPTER V

SUMMARY AND CONCLUSIONS

In Chapter II, we considered the problem of estimating population-level quantities κ_0 such as the mean, probabilities, etc. Previous literature on the topic applies only to the simple special case of estimating a population mean in the Gaussian partially linear model. The problem was motivated by an important issue in nutritional epidemiology, estimating the distribution of usual intake for episodically consumed food, where we considered a zero-inflated mixture measurement error model: such a problem is very different from the partially linear model, and the main interest is not in the population mean.

The key feature of the problem that distinguishes it from most work in semiparametric modeling is that the quantities of interest are based on both the parametric and the nonparametric parts of the model. Results were obtained for two general classes of semiparametric ones: (a) general semiparametric regression models depending on a function $\theta_0(Z)$; and (b) generalized linear single index models. Within these semiparametric frameworks, we suggested a straightforward estimation methodology, derived its limiting distribution, and showed semiparametric efficiency. An interesting part of the approach is that we also allow for partially missing responses.

In the case of standard semiparametric models, we have considered the case that the unknown function $\theta_0(Z)$ was a scalar function of a scalar argument. The results though readily extend to the case of a multivariate function of a scalar argument.

We have also assumed that $\kappa_0 = E[\mathcal{F}\{X, \theta_0(Z), \mathcal{B}_0\}]$ and $\mathcal{F}(\cdot)$ are scalar, which in principle excludes the estimation of the population variance and standard deviation. It is however readily seen that both $\mathcal{F}(\cdot)$ and κ_0 or κ_{SI} can be multivariate, and hence the obvious modification of our estimates is semiparametric efficient.

In Chapter III, we have developed methodology for efficient score test for genetic effect in general semiparametric models that can account for gene-environment interaction with nonparametrically specified environmental effects. The proposed procedure allows for repeated measurements.

We have noted that direct application of the usual likelihood based score test is generally invalid when standard bandwidth selection criteria are used, making the user rely on undersmoothing to achieve validity. This creates a difficulty in performing smoothing. To solve this problem, we proposed a profiled score statistic which can be performed using standard bandwidth selection procedures. We also found that these profiled score tests are efficient.

The main difficulty of performing the score test is that one has to estimate a function which itself is a solution of a complex integral equation. In case of repeatedly measured data, the solution generally does not have any closed form expression and hence some sort of numerical procedure is required for estimation. We overcome this problem by developing an easily implementable estimation procedure which does not involve solving integral equations and can be performed easily via standard software. The key idea lies in the fact that the target functions, based on their estimating equations, can be interpreted as Gaussian repeated measures regressions.

Simulations presented in the paper show that the proposed score-tests maintains the desired type-I error level, indicating that the asymptotic approximations work well for studies such as ours. Moreover, both simulation studies and the data example indicate that the proposed score test taking account of the interaction can achieve higher statistical power than naive tests which ignore interaction altogether. Future research areas of interest include extension of the score-test to account for the interaction of the genetic factors with several different, but biologically related, environmental factors, such as different biomarkers for a nutrient, simultaneously. In principle, the score-test can be extended using generalized additive models (GAM) to account for the effect of several different continuous exposures. Further theoretical development, however, is needed to establish the asymptotic theory for such procedures.

We address the problem of presence of measurement error in covariates in Chapter IV. We propose a Monte Carlo Corrected Score (MCCS) based method for estimation of parameters. To recap briefly, our method is a functional measurement error method, in that it makes no assumptions about the distribution of the error-prone covariate X. Its implementation is straightforward in any programming language that allows for complex-value arithmetic. Since the method is based upon profiling and backfitting for a criterion function, its theoretical development is straightforward, and standard errors are easily computed. In two examples, the multivariate partially linear model and the logistic model with quadratic effects of X, our method is numerically as efficient as the semiparametric efficient method. In fact, in the logistic case, our method performs well in presence of large measurement errors where the Ma and Carroll method faced computational problems (personal communication with Y. Ma).

We have focused on the case that the covariate Z modeled nonparametrically is univariate. However, the idea of building a semiparametric criterion function using Monte-Carlo corrected scores can be applied to more general problems, e.g., additive models.

REFERENCES

Bickel, P. J. (1982) On adaptive estimation. *Annals of Statistics*, **10**, 647–671.

Block, G., Hartman, A. M., Dresser, C. M., Carroll, M. D., Gannon, J. and Gardner, L. (1986) A data-based approach to diet questionnaire design and testing. *American Journal of Epidemiology*, **124**, 453–469.

Brown, R. L., Durbin, J. and Evans, J. M. (1975) Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, **37**, 149–192.

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

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006) *Measurement error in nonlinear models*, 2nd edn. Boca Raton: Chapman & Hall–CRC.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. and Wacholder, S. (2006) Powerful multi-locus tests for genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics*, 1002–1016.

Chen, X., Linton, O. and Van Keilegom, I. (2003) Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591–1608.

Claeskens, G. and Carroll, R. J. (2007) An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, **94**, 249–265.

Claeskens, G. and Van Keilegom, I. (2003) Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, **31**, 1852–1884.

Cook, J. R. and Stefanski, L. A. (1994) Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314–1328.

Davies, R. B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33–43.

Flegal, K. M., Carroll, M. D., Ogden, C. L. and Johnson, C. L. (2002) Prevalence and trends in obesity among US adults, 1999-2000. *Journal of the American Medical Association*, **288**, 1723–1727.

Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal smoothing in single-index models. *Annals of Statistics*, **21**, 157–178.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004) *Nonparametric and semiparametric models*. New York: Springer-Verlag.

Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, **408**, 986–995.

Hayes, R. B., Reding, D., Kopp, W., Subar, A. F., Bhat, N., Rothman, N., Caporaso, N., Ziegler, R. G., Johnson, C. C., Weissfeld, J. L., Hoover, R. N., Hartge, P., Palace, C. and Gohagan, J. K. (2000) Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (plco) cancer screening trial. *Controlled Clinical Trials*, **21(6 Suppl)**, 349S–355S.

Huggins, R. (2006) Understanding nonparametric estimation for clustered data. *Biometrika*, **93**, 486–489.

Kerber, R. L., Till, J. E., Simon, S. L., Lyon, J. L., Thomas, D. C., Preston-Martin, S., Rollison, M. L., Lloyd, R. D. and Stevens, W. (1993) A cohort study of thyroid disease in relation to fallout from nuclear weapons testing. *Journal of the American Medical Association*, **270**, 2076–2083.

Li, Y., Guolo, A., Owen Hoffman, F. and Carroll, R. J. (2007) Shared uncertainty in measurement error problems, with application to nevada test site fallout data. *Biometrics*, **63**, 1226–1236.

Li, Y. and Ruppert, D. (2008) On the asymptotics of penalized splines. *Biometrika*, **95**, 415–436.

Liang, H., Härdle, W. and Carroll, R. J. (1999) Estimation in a semiparametric partially linear errors-in-variables model. *Annals of Statistics*, **27**, 1519–1535.

Lin, D. Y. and Zou, F. (2004) Assessing genomewide statistical significance in linkage studies. *Genetic Epidemiology*, **27**, 202–214.

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

Lin, X. and Carroll, R. J. (2006) Semiparametric estimation in general repeated measures problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**, 69–88.

Lin, X., Wang, N., Welsh, A. H. and Carroll, R. J. (2004) Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika*, **91**, 177–193.

Lubin, J. H., Schafer, D. W., Ron, E., Stovall, M. and Carroll, R. J. (2004) A reanalysis of thyroid neoplasms in the israeli tinea capitis study accounting for dose uncertainties. *Radiation Research*, **161**, 359–368.

Lyon, J. L., Alder, S. C., Stone, M. B., Scholl, A., Reading, J. C., Holubkov, R., Sheng, X., White, G., Hegmann, K. T., Anspaugh, L., Hoffman, F., Simon, S. L., Thomas, B., Carroll, R. J. and Meikle, A. W. (2006) Thyroid disease associated with exposure to the nevada test site radiation: a reevaluation based on corrected dosimetry and examination data. *Epidemiology*, **17**, 604–614.

Ma, Y. and Carroll, R. J. (2006) Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, **101**, 1465–1474.

Ma, Y., Chiou, J. M. and Wang, N. (2006) Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*, **93**, 75–84.

Mallick, B., Hoffman, F. O. and Carroll, R. J. (2002) Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. *Biometrics*, **58**, 13–20.

Newey, W. K. (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics*, **5**, 99–135.

Newey, W. K., Hsieh, F. and Robins, J. M. (2004) Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, **72**, 947–962.

Novick, S. J. and Stefanski, L. A. (2002) Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, **97**, 472–481.

Pierce, D. A. and Kellerer, A. M. (2004) Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. *Biometrika*, **91**, 863–876.

Powell, J. L. and Stoker, T. M. (1996) Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, **75**, 291–316.

Reeves, G., Cox, D. R., Darby, S. C. and Whitley, E. (1998) Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine*, **17**, 2157–2177.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric regression*, vol. 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

Schafer, D. W. and Gilbert, E. S. (2006) Some statistical implications of does uncertainty in radiation dose-response analyses. *Radiation Research*, **166**, 303–312.

Schafer, D. W., Lubin, J. H., Ron, E., Stovall, M. and Carroll, R. J. (2001) Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry. *Biometrics*, **57**, 689–697.

Sepanski, J. H., Knickerbocker, R. and Carroll, R. J. (1994) A semiparametric correction for attenuation. *Journal of the American Statistical Association*, **89**, 1366–1373.

Severini, T. A. and Staniswalis, J. G. (1994) Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89**, 501–511.

Severini, T. A. and Wong, W. H. (1992) Profile likelihood and conditionally parametric models. *Annals of Statistics*, **20**, 1768–1802.

Silverman, B. W. (1984) Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, **12**, 898–916.

Simon, S., Till, J. E., Lloyd, R. D., Kerber, R., Thomas, D. C., Preston-Martin, S., Lyon, J. L. and Stevens, W. (1995) The utah leukemia case–control study: dosimetry methodology and results. *Health Physics*, **68**, 460–471.

Simon, S. L., Anspaugh, L. R., Hoffman, F. O., Scholl, A. E., Stone, M. B., Thomas, B. A. and Lyon, J. L. (2006) 2004 update of dosimetry for the utah thyroid cohort study. *Radiation Research*, **165**, 208–222.

Stefanski, L. A. and Cook, J. R. (1995) Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, **90**, 1247–1256.

Stevens, W., Till, J., Thomas, D., Lyon, J., Kerber, R., Preston-Martin, S., Simon, S., Rallison, M. and Lloyd, R. (1992) Assessment of leukemia and thyroid disease in relation to fallout in utah: report of a cohort study of thyroid disease and radioactive fallout from the nevada test site. Tech. rep., University of Utah, Salt Lake City.

Stram, D. O. and Kopecky, K. J. (2003) Power and uncertainty analysis of epidemiological studies of radiation-related disease risk in which dose estimates are based on a complex dosimetry system: some observations. *Radiation Research*, **160**, 408–417.

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A. and Rosenfeld, S. (2001) Comparative validation of the block, willett and national cancer institute food frequency questionnaires: the eating at america's table study. *American Journal of Epidemiology*, **154**, 1089–1099.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical analysis of finite mixture distributions*. Chichester: John Wiley & Sons Ltd.

Tooze, J. A., Grunwald, G. K. and Jones, R. H. (2002) Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*, **11**, 341–355.

Tsiatis, A. A. and Ma, Y. (2004) Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, **91**, 835–848.

Tukey, J. W. (1949) One degree of freedom for non-additivity. *Biometrics*, **5**, 232–242.

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

Wang, Q., Linton, O. and Härdle, W. (2004) Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, **99**, 334–345.

Woteki, C. E. (2003) Integrated nhanes: uses in national policy. *Journal of Nutrition*, **133**, 582S–584S.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER II

In what follows, the arguments for $\mathcal{L}$ and its derivatives are in the form $\mathcal{L}(\cdot) = \mathcal{L}\{Y, X, \mathcal{B}_0, \theta_0(Z)\}$. The arguments for $\mathcal{F}$ and its derivatives are $(\cdot) = \{X, \theta_0(Z), \mathcal{B}_0\}$. Also, please note that in our arguments about semiparametric efficiency, we use the symbol d exactly as it was used by Newey (1990). It does not stand for differential.

A.1. Assumptions and Remarks

A.1.1. General Considerations

The main results needed for the asymptotic distribution of our estimator are (2.5) and (2.6). The single-index model assumptions are given already in Carroll et al. (1997).

Results (2.5) and (2.6) hold under smoothness and moment conditions for the likelihood function, and under smoothness and boundedness conditions for $\theta(\cdot)$. The strength of these conditions depends on the generality of the problem. For the partially linear Gaussian model of Wang et al. (2004), because the profile likelihood estimator of β is an explicit function of regressions of Y and X on Z, the conditions are simply conditions about uniform expansions for kernel regression estimators, as in for example Claeskens and Van Keilegom (2003). For generalized partially linear models, Severini and Staniswalis (1994) give a series of moment and smoothness conditions towards this end. For general likelihood problems, Claeskens and Carroll (2007) state that the conditions needed are as follows.

(C1) The bandwidth sequence $h_n \to 0$ as $n \to \infty$, in such a way that $nh_n/\log(n) \to \infty$ and $h_n \geq \{\log(n)/n\}^{1-2/\lambda}$ for λ as in condition (C4).

(C2) The kernel function K is a symmetric, continuously differentiable pdf on $[-1, 1]$ taking on the value zero at the boundaries. The design density $f(\cdot)$ is differentiable on an interval $B = [b_1, b_2]$, the derivative is continuous, and $\inf_{z\in B} f(z) > 0$. The function $\theta(\cdot, \mathcal{B})$ has 2 continuous derivatives on B and is also twice differentiable with respect to $\mathcal{B}$.

(C3) The Kullback-Leibler distance between $\mathcal{L}\{\cdot, \cdot, \mathcal{B}, \theta(\cdot, \mathcal{B})\}$, and $\mathcal{L}\{\cdot, \cdot, \mathcal{B}', \theta(\cdot, \mathcal{B}')\}$ is strictly positive for $\mathcal{B} \neq \mathcal{B}'$. For every (y, x), third partial derivatives of $\mathcal{L}\{y, x, \mathcal{B}, \theta(z)\}$ with respect to $\mathcal{B}$ exist and are continuous in $\mathcal{B}$. The 4th partial derivative exists for almost all (y, x). Further, mixed partial derivatives $\frac{\partial^{r+s}}{\partial \mathcal{B}^r \partial v^s}\mathcal{L}\{y, x, \mathcal{B}, v\}|_{v=\theta(z)}$, with $0 \leq r, s \leq 4, r + s \leq 4$ exist for almost all (y, x) and $E\{\sup_{\mathcal{B}} \sup_v \left|\frac{\partial^{r+s}}{\partial \mathcal{B}^r \partial v^s}\mathcal{L}\{y, x, \mathcal{B}, v\}\right|^2\} < \infty$. The Fisher information, $G(z)$, possesses a continuous derivative and $\inf_{z\in B} G(z) > 0$.

(C4) There exists a neighborhood $\mathcal{N}\{\mathcal{B}_0, \theta_0(z)\}$ such that

$$\max_{k=1,2} \sup_{z\in B} \left\| \sup_{(\mathcal{B},\theta)\in\mathcal{N}\{\mathcal{B}_0,\theta_0(z)\}} \left| \frac{\partial^k}{\partial\theta^k} \log\{\mathcal{L}(Y, X, \mathcal{B}, \theta)\} \right| \right\|_{\lambda,z} < \infty$$

for some $\lambda \in (2, \infty]$, where $\|\cdot\|_{\lambda,z}$ is the L^λ-norm, conditional on $Z = z$. Further,

$$\sup_{z\in B} E_z \left[\sup_{(\mathcal{B},\theta)\in\mathcal{N}\{\mathcal{B}_0,\theta_0(z)\}} |\frac{\partial^3}{\partial\theta^3} \log\{\mathcal{L}(Y, X, \mathcal{B}, \theta)\}| \right] < \infty.$$

The above regularity conditions are the same as those used in a local likelihood setting where one wishes to obtain strong uniform consistency of the local likelihood estimators. Condition (C3) requires the 4th partial derivative of the log profile likelihood to have a bounded second moment, it further requires the Fisher information matrix

to be invertible and to be differentiable with respect to z. Condition (C4) requires a bound on the first and second derivatives of the log profile likelihood and of the first moment of the third partial derivative, in a neighborhood of the true parameter values.

A.1.2. Compactly Supported Z

Multiple reviewers of earlier drafts of this paper commented that the assumption that Z be compactly supported with density positive on this support is too strong.

However, this assumption is completely standard in the kernel-based semiparametric literature for estimation of $\mathcal{B}_0$, because it is needed for uniform expansions for estimation of $\theta_0(\cdot)$. The assumption is made in the founding papers on semiparametric likelihood estimation (Severini and Wong 1992, p. 1875, part e); the first paper on generalized linear models (Severini and Staniswalis 1994, p. 511, assumption D), the first paper on efficient estimation of partially linear single index models (Carroll et al. 1997, p. 485, condition 2a); and the precursor paper to ours that is focused on estimation of the population mean in a partially linear model (Wang et al. 2004, p. 341, condition C.T). The uniform expansions for local likelihood given in Claeskens and van Keilegom (2003) also make this assumption, see their page 1869, condition R0. Thus, our assumption on the design density of Z is a standard one.

The reason this assumption is made has to do with kernel technology, where proofs generally require a uniform expansion for the kernel regression, or at least uniform in all observed values of Z which is the same thing. The Nadaraya-Watson estimator, for example, has a denominator that is a density estimate, and the condition on Z stops this denominator from getting too close to zero. Ma et al. (2006), who make

the same assumption (their condition 6 on page 83), state that it is necessary to avoid "pathological cases".

A.2. Proof of Result 1

A.2.1. Asymptotic Expansion

We first show (2.9). First note that $\mathcal{L}$ is a loglikelihood function conditioned on (X, Z), so that we have

$$\begin{aligned} E\{\delta\mathcal{L}_{\theta\theta}(\cdot)|X,Z\} &= -E\{\delta\mathcal{L}_{\theta}(\cdot)\mathcal{L}_{\theta}(\cdot)|X,Z\}; \\ E\{\delta\mathcal{L}_{\theta\mathcal{B}}(\cdot)|X,Z\} &= -E\{\delta\mathcal{L}_{\theta}(\cdot)\mathcal{L}_{\mathcal{B}}(\cdot)|X,Z\}. \end{aligned} \tag{A.1}$$

By a Taylor expansion,

$$\begin{aligned} n^{1/2}(\widehat{\kappa}-\kappa_0) &= n^{-1/2}\sum_{i=1}^{n}\Big[\mathcal{F}_i(\cdot)-\kappa_0+\{\mathcal{F}_{i\mathcal{B}}(\cdot)+\mathcal{F}_{i\theta}(\cdot)\theta_{\mathcal{B}}(Z_i,\mathcal{B}_0)\}^{\mathrm{T}}(\widehat{\mathcal{B}}-\mathcal{B}_0) \\ &\qquad +\mathcal{F}_{i\theta}(\cdot)\{\widehat{\theta}(Z_i,\mathcal{B}_0)-\theta_0(Z_i)\}\Big]+o_p(1) \\ &= \mathcal{M}_2^{\mathrm{T}}n^{1/2}(\widehat{\mathcal{B}}-\mathcal{B}_0) \\ &\qquad +n^{-1/2}\sum_{i=1}^{n}\Big[\mathcal{F}_i(\cdot)-\kappa_0+\mathcal{F}_{i\theta}(\cdot)\{\widehat{\theta}(Z_i,\mathcal{B}_0)-\theta_0(Z_i)\}\Big]+o_p(1). \end{aligned}$$

Because $nh^4 \to 0$, using (2.5), we see that

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}\mathcal{F}_{i\theta}(\cdot)\{\widehat{\theta}(Z_i,\mathcal{B}_0)-\theta_0(Z_i)\} \\ &= -n^{-1/2}\sum_{i=1}^{n}\mathcal{F}_{i\theta}(\cdot)n^{-1}\sum_{j=1}^{n}\delta_j K_h(Z_j-Z_i)\mathcal{L}_{j\theta}(\cdot)/\Omega(Z_i)+o_p(1) \\ &= -n^{-1/2}\sum_{i=1}^{n}\delta_i\mathcal{L}_{i\theta}(\cdot)n^{-1}\sum_{j=1}^{n}K_h(Z_j-Z_i)\mathcal{F}_{j\theta}(\cdot)/\Omega(Z_j)+o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n}\delta_i D_i(\cdot)+o_p(1), \end{aligned}$$

the last step following because the interior sum is a kernel regression converging to D_i, see Carroll et al. (1997) for details. Result (2.9) now follows from (2.6). The limiting variance (2.10) is an easy calculation, noting that (A.1) implies that

$$\begin{aligned} E\{\delta\epsilon\mathcal{L}_\theta(\cdot)|Z\} &= E\{\delta\mathcal{L}_\theta(\cdot)\mathcal{L}_\mathcal{B}(\cdot) + \delta\mathcal{L}_\theta(\cdot)\mathcal{L}_\theta(\cdot)\theta_\mathcal{B}(Z,\mathcal{B}_0)|Z\} \\ &= -E\{\delta\mathcal{L}_{\mathcal{B}\theta}(\cdot) + \delta\mathcal{L}_{\theta\theta}(\cdot)\theta_\mathcal{B}(Z,\mathcal{B}_0)|Z\} = 0 \end{aligned} \tag{A.2}$$

by the definition of $\theta_\mathcal{B}(\cdot)$ given at (2.7), and hence the last two terms in (2.9) are uncorrelated. We will use (A.2) repeatedly in what follows.

A.2.2. Pathwise Differentiability

We now turn to the semiparametric efficiency, using results of Newey (1990). The relevant text of his paper is in his Section 3, especially through his equation (9). A parameter $\kappa = \kappa(\Theta)$ is pathwise differentiable under two conditions. The first is that $\kappa(\Theta)$ is differentiable for all smooth parametric submodels: in our case, the parametric submodels include $\mathcal{B}$, parametric submodels for $\theta(\cdot)$, and parametric submodels for the distribution of (X, Z) and the probability function $\text{pr}(\delta = 1|X, Z)$. This condition is standard in the literature and fairly well required. Our motivating example clearly satisfies this condition.

The second condition is that there exists a random vector d such that $E(d^\text{T}d) < \infty$, and $\partial\kappa(\Theta)/\partial\Theta = E(dS_\Theta^\text{T})$, where S_Θ is the loglikelihood score for the parametric submodel. Newey notes that pathwise differentiability also holds if the first condition holds, and if there is a regular estimator in the semiparametric problem. Generally, as Newey notes, finding a suitable random variable d can be difficult.

Assuming pathwise differentiability, which as stated above we later show, the efficient influence function is calculated by projecting d onto the nuisance tangent space. One innovation here is that we can calculate the efficient influence function without having an explicit representation for d.

Our development below will consist of two steps. In the first, we will assume pathwise differentiability, and derive the efficient score function under that assumption. Using this derivation, we will then exhibit a random variable d that has the requisite property.

A.2.3. Efficiency

Recall that $\text{pr}(\delta = 1|X, Z) = \pi(X, Z)$. Let $f_{X,Z}(x, z)$ be the density function of (X, Z). Let the model under consideration be denoted by M_0. Now consider a smooth parametric submodel M_λ, with $f_{X,Z}(x, z, \alpha_1)$, $\theta(z, \alpha_2)$ and $\pi(X, Z, \alpha_3)$ in place of $f_{X,Z}(x, z)$, $\theta_0(z)$ and $\pi(X, Z)$ respectively. Then under M_λ the loglikelihood is given by

$$\begin{aligned} L(\cdot) &= \delta\mathcal{L}(\cdot) + \delta\log\{\pi(X, Z, \alpha_3)\} + (1 - \delta)\log\{1 - \pi(X, Z, \alpha_3)\} \\ &\quad +\log\{f_{X,Z}(X, Z, \alpha_1)\}, \end{aligned}$$

where $(\cdot)$ represents the argument $\{Y, X, \theta(Z, \alpha_2), \mathcal{B}_0\}$. Then the score functions in this parametric submodel are given by

$$\begin{aligned} \partial L(\cdot)/\partial\mathcal{B} &= \delta\mathcal{L}_\mathcal{B}(\cdot); \\ \partial L(\cdot)/\partial\alpha_1 &= \partial\log\{f_{X,Z}(X, Z, \alpha_1)\}/\partial\alpha_1; \\ \partial L(\cdot)/\partial\alpha_2 &= \delta\mathcal{L}_\theta(\cdot)\partial\theta(Z, \alpha_2)/\partial\alpha_2; \end{aligned}$$

$$\partial L(\cdot)/\partial\alpha_3 \;=\; \{\partial\pi(X,Z,\alpha_3)/\partial\alpha_3\}\{\delta-\pi(X,Z,\alpha_3)\}/[\pi(X,Z,\alpha_3)\{1-\pi(X,Z,\alpha_3)\}].$$

Thus, the tangent space is spanned by the functions $\delta\mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}}$, $s_f(x,z)$, $\delta\mathcal{L}_\theta(\cdot)g(Z)$, $a(X,Z)\{\delta-\pi(X,Z)\}$, where $s_f(x,z)$ is any function with mean 0, while $g(z)$ and $a(X,Z)$ are any functions. For computational convenience, we rewrite the tangent space as the linear span of four subspaces $\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3,\mathcal{T}_4$ that are orthogonal to each other (see below) and defined as follows:

$$\begin{aligned}
\mathcal{T}_1 &= \delta\mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}} + \delta\mathcal{L}_\theta(\cdot)\theta^{\mathrm{T}}_{\mathcal{B}}(Z,\mathcal{B}_0)\\
\mathcal{T}_2 &= s_f(x,z)\\
\mathcal{T}_3 &= \delta\mathcal{L}_\theta(\cdot)g(Z)\\
\mathcal{T}_4 &= a(X,Z)\{\delta-\pi(X,Z)\}.
\end{aligned}$$

To show that these spaces are orthogonal, we first note that by assumption, the data are missing at random, and hence $\mathrm{pr}(\delta=1|Y,X,Z)=\pi(X,Z)$. This means that $\mathcal{T}_4$ is orthogonal to the other three spaces. Note also that, by assumption, $E\{\mathcal{L}_{\mathcal{B}}(\cdot)|X,Z\}=E\{\mathcal{L}_\theta(\cdot)|X,Z\}=0$. This shows that $\mathcal{T}_2$ is orthogonal to $\mathcal{T}_1$ and $\mathcal{T}_3$. It remains to show that $\mathcal{T}_1$ and $\mathcal{T}_3$ are orthogonal, which we showed in (A.2). Thus, the spaces $\mathcal{T}_1$-$\mathcal{T}_4$ are orthogonal.

Note that, under model M_λ,

$$\kappa_0 \;=\; \int \mathcal{F}\{X,\theta(Z,\alpha_2),\mathcal{B}_0\}f_{X,Z}(x,z,\alpha_1)\,dxdz.$$

Hence we have that

$$\begin{aligned}
\partial\kappa_0/\partial\mathcal{B} &= E\{\mathcal{F}_{\mathcal{B}}(\cdot)\};\\
\partial\kappa_0/\partial\alpha_1 &= E[\mathcal{F}(\cdot)\partial\log\{f_{X,Z}(X,Z,\alpha_1)\}/\partial\alpha_1];
\end{aligned}$$

$$\begin{aligned}\partial\kappa_0/\partial\alpha_2 &= E\{\mathcal{F}_\theta(\cdot)\partial\theta(Z,\alpha_2)/\partial\alpha_2\};\\ \partial\kappa_0/\partial\alpha_3 &= 0.\end{aligned}$$

Now, by pathwise differentiability and equation (7) of Newey (1990), there exists a random variable d, which we need not compute, such that

$$E\{\mathcal{F}_{\mathcal{B}}(\cdot)\} = E[d\{\delta\mathcal{L}_{\mathcal{B}}(\cdot)\}]; \tag{A.3}$$

$$E\{\mathcal{F}(\cdot)s_f(X,Z)\} = E\{ds_f(X,Z)\}; \tag{A.4}$$

$$E\{\mathcal{F}_\theta(\cdot)g(Z)\} = E\{d\delta\mathcal{L}_\theta(\cdot)g(Z)\}; \tag{A.5}$$

$$0 = E[da(X,Z)\{\delta-\pi(X,Z)\}]. \tag{A.6}$$

Next we compute the projections of d into $\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3$ and $\mathcal{T}_4$. First note that, by (A.4), for any function $s_f(X,Z)$ with expectation zero, we have $E[\{d-\mathcal{F}(\cdot)+\kappa_0\}s_f(X,Z)] = 0$, which implies that the projection of d into $\mathcal{T}_2$ is given by

$$\Pi(d|\mathcal{T}_2) = \mathcal{F}(\cdot)-\kappa_0. \tag{A.7}$$

Also, by (A.1) and (A.5), for any function $g(Z)$, we have

$$\begin{aligned}&E[\{d-\delta D(\cdot)\}\delta g(Z)\mathcal{L}_\theta(\cdot)]\\ &\quad= E\{\mathcal{F}_\theta(\cdot)g(Z)\}+E[\delta g(Z)\mathcal{L}_\theta^2(\cdot)E\{\mathcal{F}_\theta(\cdot)|Z\}/E\{\delta\mathcal{L}_{\theta\theta}(\cdot)|Z\}]\\ &\quad= 0,\end{aligned}$$

and hence the projection of d onto $\mathcal{T}_3$ is given by

$$\Pi(d|\mathcal{T}_3) = \delta D(\cdot). \tag{A.8}$$

In addition, by (A.3) and (A.5),

$$E[\{d-\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\delta\epsilon\}\delta\epsilon^{\mathrm{T}}] = E\{\mathcal{F}_{\mathcal{B}}^{\mathrm{T}}(\cdot)\}-E\{\mathcal{F}_\theta(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z,\mathcal{B}_0)\}-E(\mathcal{M}_2^{\mathrm{T}}\mathcal{M}_1^{-1}\delta\epsilon\epsilon^{\mathrm{T}})$$

$$= \quad 0.$$

Hence the projection of d into $\mathcal{T}_1$ is given by

$$\Pi(d|\mathcal{T}_1) = \delta \mathcal{M}_2^{\mathrm{T}} \mathcal{M}_1^{-1} \epsilon. \tag{A.9}$$

Also by (A.6), we have $\Pi(d|\mathcal{T}_4) = 0$. Using (A.7), (A.8) and (A.9) we get the efficient influence function for κ_0 is

$$\psi_{eff} = \Pi(d|\mathcal{T}_1) + \Pi(d|\mathcal{T}_2) + \Pi(d|\mathcal{T}_3) + \Pi(d|\mathcal{T}_4) = \mathcal{F}(\cdot) - \kappa_0 + \delta \mathcal{M}_2^{\mathrm{T}} \mathcal{M}_1^{-1} \epsilon + \delta D(\cdot),$$

which is same as (2.9), hence completing the proof under the assumption of pathwise differentiability. In the calculations that follow, we will write $\mathcal{F}_{\mathcal{B}}$ rather than $\mathcal{F}_{\mathcal{B}}(\cdot)$, a rather than $a(X, Z)$, etc.

We now show pathwise differentiability, and hence semiparametric efficiency, i.e., we show that (A.3)-(A.6) hold for $d = \mathcal{F} - \kappa_0 + \delta D + \delta \mathcal{M}_2^T \mathcal{M}_1^{-1} \epsilon$.

To verify (A.3), we see that

$$\begin{aligned}
E(d\delta\mathcal{L}_{\mathcal{B}}) &= E[(\mathcal{F} - \kappa_0 + \delta D + \delta \mathcal{M}_2^T \mathcal{M}_1^{-1} \epsilon)\delta\mathcal{L}_{\mathcal{B}}] \\
&= E[\delta D \mathcal{L}_{\mathcal{B}} + \delta \mathcal{M}_2^T \mathcal{M}_1^{-1} \epsilon \mathcal{L}_{\mathcal{B}}] \\
&= -E\{\mathcal{L}_\theta \frac{E(\mathcal{F}_\theta|Z)}{E(\delta\mathcal{L}_{\theta\theta}|Z)} \mathcal{L}_{\mathcal{B}} \delta\} + E\{\delta\mathcal{L}_{\mathcal{B}}(\mathcal{L}_{\mathcal{B}} + \mathcal{L}_\theta \theta_{\mathcal{B}})^{\mathrm{T}}\} \mathcal{M}_1^{-1} \mathcal{M}_2 \\
&= E\{\delta\mathcal{L}_{\theta\mathcal{B}} \frac{E(\mathcal{F}_\theta|Z)}{E(\delta\mathcal{L}_{\theta\theta}|Z)}\} - E\{\delta(\mathcal{L}_{\mathcal{B}\mathcal{B}} + \mathcal{L}_{\mathcal{B}\theta}\theta_{\mathcal{B}}^{\mathrm{T}})\} \mathcal{M}_1^{-1} \mathcal{M}_2 \\
&= -E(\mathcal{F}_\theta \theta_{\mathcal{B}}) + \mathcal{M}_2 \\
&= E(\mathcal{F}_{\mathcal{B}}).
\end{aligned}$$

To verify (A.4), we see that

$$\begin{aligned} E(ds_f) &= E\{(\mathcal{F} - \kappa_0 + \delta D + \delta \mathcal{M}_2^T \mathcal{M}_1^{-1} \epsilon) s_f\} \\ &= E(\mathcal{F} s_f) - \kappa_0 E(s_f) + E\{E(\delta D + \delta \mathcal{M}_2^T \mathcal{M}_1^{-1} \epsilon | X, Z) s_f\} \\ &= E(\mathcal{F} s_f). \end{aligned}$$

To verify (A.5), we see that

$$\begin{aligned} E(d\delta \mathcal{L}_\theta g) &= E\{(\mathcal{F} - \kappa_0 + \delta D + \delta \mathcal{M}_2^T \mathcal{M}_1^{-1} \epsilon) \delta \mathcal{L}_\theta g\} \\ &= E(D \mathcal{L}_\theta \delta g) + \mathcal{M}_2^T \mathcal{M}_1^{-1} E(\epsilon \mathcal{L}_\theta \delta g) \\ &= -E\{\mathcal{L}_\theta \frac{E(\mathcal{F}_\theta | Z)}{E(\delta \mathcal{L}_{\theta\theta} | Z)} \mathcal{L}_\theta \delta g\} + \mathcal{M}_2^T \mathcal{M}_1^{-1} E\{(\mathcal{L}_\mathcal{B} + \mathcal{L}_\theta \theta_\mathcal{B}) \mathcal{L}_\theta \delta g\} \\ &= E(\mathcal{F}_\theta g) - \mathcal{M}_2^T \mathcal{M}_1^{-1} E\{(\mathcal{L}_{\mathcal{B}\theta} + \mathcal{L}_{\theta\theta} \theta_\mathcal{B}) \delta g\} \\ &= E(\mathcal{F}_\theta g) - \mathcal{M}_2^T \mathcal{M}_1^{-1} E\{E(\delta \mathcal{L}_{\mathcal{B}\theta} + \delta \mathcal{L}_{\theta\theta} \theta_\mathcal{B} | Z) g\} \\ &= E(\mathcal{F}_\theta g), \end{aligned}$$

where again we have used (A.2). Finally, because the responses are missing at random, (A.6) is immediate. This completes the proof.

A.3. Sketch of Lemma 1

We have that

$$\widehat{\kappa}_{\text{marg}} = n^{-1} \sum_{i=1}^{n} \Big[\frac{\delta_i}{\widehat{\pi}_{\text{marg}}(Z_i)} \mathcal{G}(Y_i) + \Big\{ 1 - \frac{\delta_i}{\widehat{\pi}_{\text{marg}}(Z_i)} \Big\} \mathcal{F}\{X_i, \widehat{\theta}(Z_i, \widehat{\mathcal{B}}), \widehat{\mathcal{B}}\} \Big] = A_1 + A_2.$$

By calculations that are similar to those above, and using (2.11), it is readily shown that

$$A_1 = n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\pi_{\text{marg}}(Z_i)} \mathcal{G}(Y_i)$$

$$-n^{-1}\sum_{i=1}^{n}\{\delta_i-\pi_{\rm marg}(Z_i)\}E\Big[\frac{\delta_i\mathcal{G}(Y_i)}{\{\pi_{\rm marg}(Z_i)\}^2}|Z_i\Big]+o_p(n^{-1/2}).$$

We can write

$$\begin{aligned}
A_2 &= B_1+B_2+o_p(n^{-1/2});\\
B_1 &= n^{-1}\sum_{i=1}^{n}\Big\{1-\frac{\delta_i}{\pi_{\rm marg}(Z_i)}\Big\}\mathcal{F}\{X_i,\widehat{\theta}(Z_i,\widehat{\mathcal{B}}),\widehat{\mathcal{B}}\}\\
B_2 &= n^{-1}\sum_{i=1}^{n}\frac{\delta_i\mathcal{F}\{X_i,\widehat{\theta}(Z_i,\widehat{\mathcal{B}}),\widehat{\mathcal{B}}\}}{\{\pi_{\rm marg}(Z_i)\}^2}\{\widehat{\pi}_{\rm marg}(Z_i)-\pi_{\rm marg}(Z_i)\}.
\end{aligned}$$

Using (2.5) and (2.6), it is easy to show that

$$\begin{aligned}
B_1 &= n^{-1}\sum_{i=1}^{n}\Big\{1-\frac{\delta_i}{\pi_{\rm marg}(Z_i)}\Big\}\mathcal{F}_i(\cdot)+\mathcal{M}_{2,\rm marg}\mathcal{M}_1^{-1}n^{-1}\sum_{i=1}^{n}\delta_i\epsilon_i\\
&\quad +n^{-1}\sum_{i=1}^{n}\delta_i D_{i,\rm marg}(\cdot)+o_p(n^{-1/2}).
\end{aligned}$$

Using (2.11) once again, we see that

$$B_2=n^{-1}\sum_{i=1}^{n}\{\delta_i-\pi_{\rm marg}(Z_i)\}E\Big[\frac{\delta_i\mathcal{F}_i(\cdot)}{\{\pi_{\rm marg}(Z_i)\}^2}|Z_i\Big]+o_p(n^{-1/2}).$$

Collecting terms, and noting that

$$0=E\Big[\frac{\delta_i\{\mathcal{G}(Y_i)-\mathcal{F}_i(\cdot)\}}{\{\pi_{\rm marg}(Z_i)\}^2}|Z_i\Big],$$

this proves (2.12).

A.4. Sketch of Lemma 2

We have that

$$\widehat{\kappa}=n^{-1}\sum_{i=1}^{n}\Big[\frac{\delta_i}{\pi(X_i,Z_i,\widehat{\zeta})}\mathcal{G}(Y_i)+\Big\{1-\frac{\delta_i}{\pi(X_i,Z_i,\widehat{\zeta})}\Big\}\mathcal{F}\{X_i,\widehat{\theta}(Z_i,\widehat{\mathcal{B}}),\widehat{\mathcal{B}}\}\Big]=A_1+A_2,$$

say. By a simple Taylor series expansion,

$$\begin{aligned} A_1 &= n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{\pi(X_i,Z_i,\zeta)}\mathcal{G}(Y_i) \\ &\quad -E\Big\{\frac{1}{\pi(X,Z,\zeta)}\mathcal{G}(Y)\pi_\zeta(X,Z,\zeta)\Big\}^{\mathrm{T}} n^{-1}\sum_{i=1}^{n}\psi_{i\zeta}+o_p(n^{-1/2}). \end{aligned}$$

In addition,

$$\begin{aligned} A_2 &= B_1+B_2+o_p(n^{-1/2}); \\ B_1 &= n^{-1}\sum_{i=1}^{n}\Big\{1-\frac{\delta_i}{\pi(X_i,Z_i,\zeta)}\Big\}\mathcal{F}\{X_i,\widehat{\theta}(Z_i,\widehat{\mathcal{B}}),\widehat{\mathcal{B}}\}; \\ B_2 &= n^{-1}\sum_{i=1}^{n}\frac{\delta_i\mathcal{F}\{X_i,\widehat{\theta}(Z_i,\widehat{\mathcal{B}}),\widehat{\mathcal{B}}\}}{\{\pi(X_i,Z_i,\zeta)\}^2}\pi_\zeta(X_i,Z_i,\zeta)^{\mathrm{T}}(\widehat{\zeta}-\zeta)+o_p(n^{-1/2}). \end{aligned}$$

Using the fact that

$$0=E\Big\{1-\frac{\delta_i}{\pi(X_i,Z_i,\zeta)}|X,Z\Big\},$$

it follows easily that

$$B_1=n^{-1}\sum_{i=1}^{n}\Big\{1-\frac{\delta_i}{\pi(X_i,Z_i,\zeta)}\Big\}\mathcal{F}_i(\cdot)+o_p(n^{-1/2}).$$

It also follows that

$$B_2=E\Big\{\frac{1}{\pi(X,Z,\zeta)}\mathcal{F}(\cdot)\pi_\zeta(X,Z,\zeta)\Big\}^{\mathrm{T}} n^{-1}\sum_{i=1}^{n}\psi_{i\zeta}(\cdot)+o_p(n^{-1/2}).$$

Collecting terms and using the fact that $E\{\mathcal{G}(Y)|X,Z\}=\mathcal{F}(\cdot)$, the result follows.

A.5. Proof of Result 2

A.5.1. Asymptotic Expansion

We first show the expansion (2.15). Recall that $\mathcal{B} = (\gamma, \beta)$. The only things that differ with the calculations of Carroll et al. (1997) is that we add in terms involving δ_i and we need not worry about any constraint on γ, and thus we avoid items like their $P\alpha$ on their page 487.

In their equation (A.12), they show that

$$n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) = n^{-1/2}\mathcal{Q}^{-1}\sum_{i=1}^{n}\delta_i\mathcal{N}_i\epsilon_i + o_p(1). \tag{A.10}$$

Define $H(u) = [E\{\rho_2(\cdot)|U = u\}]^{-1}$. In their equations (A.13), Carroll et al. (1997) show that

$$\begin{aligned}\widehat{\theta}(R + S^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}}) - \theta_0(R + S^{\mathrm{T}}\gamma_0) &= \theta_0^{(1)}(R + S^{\mathrm{T}}\gamma_0)S^{\mathrm{T}}(\widehat{\gamma} - \gamma_0) \\ &\quad + \widehat{\theta}(R + S^{\mathrm{T}}\gamma_0, \widehat{\mathcal{B}}) - \theta_0(R + S^{\mathrm{T}}\gamma_0) + o_p(n^{-1/2}).\end{aligned} \tag{A.11}$$

Also, in their equation (A.11), they show that

$$\begin{aligned}\widehat{\theta}(u, \widehat{\mathcal{B}}) - \theta_0(u) &= n^{-1}\sum_{i=1}^{n}\delta_i K_h(U_i - u)\epsilon_i H(u)/f(u) \\ &\quad - H(u)[E\{\delta\Lambda\rho_2(\cdot)|U = u\}]^{\mathrm{T}}(\widehat{\mathcal{B}} - \mathcal{B}_0) + o_p(n^{-1/2}).\end{aligned} \tag{A.12}$$

Carroll et al. (1997) did not consider an estimate of ϕ. Make the definition

$$\mathcal{G}\{\phi, Y, X, \mathcal{B}, \theta(U)\} = \mathcal{D}_\phi(Y, \phi) - [Yc\{X^{\mathrm{T}}\beta + \theta(U)\} - \mathcal{C}\{c(\cdot)\}]/\phi^2.$$

Of course, $\mathcal{G}(\cdot)$ is the likelihood score for ϕ. If there are no arguments, we denote

$\mathcal{G} = \mathcal{G}\{\phi_0, Y, X, \mathcal{B}_0, \theta_0(R + S^{\mathrm{T}}\gamma_0)\}$. The estimating function for ϕ solves

$$0 = n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{G}\{\widehat{\phi}, Y_i, X_i, \widehat{\mathcal{B}}, \widehat{\theta}(R_i + S_i^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}})\}.$$

Since $\mathcal{G}$ is a likelihood score, it follows that

$$E[\mathcal{G}_\phi\{\phi_0, Y, X, \mathcal{B}_0, \theta_0(R + S^{\mathrm{T}}\gamma_0)\}|X, R, S] = -E\{\mathcal{G}^2|X, R, S\}.$$

By a Taylor series,

$$\begin{aligned} E(\delta\mathcal{G}^2)n^{1/2}(\widehat{\phi} - \phi_0) &= n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{G}\{\phi_0, Y_i, X_i, \widehat{\mathcal{B}}, \widehat{\theta}(R_i + S_i^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}})\} + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{G}_i + E(\delta\mathcal{G}_{\mathcal{B}}^{\mathrm{T}})n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) \\ &\quad + n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{G}_{i\theta}\{\widehat{\theta}(R_i + S_i^{\mathrm{T}}\widehat{\gamma}, \widehat{\mathcal{B}}) - \theta_0(R_i + S_i^{\mathrm{T}}\gamma_0)\} + o_p(1). \end{aligned}$$

However, it is readily verified that $E(\delta\mathcal{G}_{\mathcal{B}}|X, R, S) = 0$ and that $E(\delta\mathcal{G}_\theta|X, R, S) = 0$. It thus follows via a simple calculation using (A.11) that

$$\begin{aligned} E(\delta\mathcal{G}^2)n^{1/2}(\widehat{\phi} - \phi_0) &= n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{G}_i + n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{G}_{i\theta}\{\widehat{\theta}(U_i, \mathcal{B}_0) - \theta_0(U_i)\} + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^{n} \delta_i \mathcal{G}_i + o_p(1), \end{aligned}$$

the last step following from an application of (A.12).

With some considerable algebra, (2.15) now follows from calculations similar to those in the proof of Result 1. The variance calculation follows because it is readily shown that for any function $h(U)$,

$$0 = E[(\mathcal{N}\epsilon)\{\delta h(U)\epsilon)\}]. \tag{A.13}$$

A.5.2. Efficiency

We now turn to semiparametric efficiency. Recall that the GPLSIM follows the form (2.20) with $X^{\mathrm{T}}\beta_0 + \theta_0(R + S^{\mathrm{T}}\gamma_0)$, and that $U = R + S^{\mathrm{T}}\gamma_0$. It is immediate that $V\{\mu(t)\} = \mu^{(1)}(t)/c^{(1)}(t)$, that $c^{(1)}(t) = \rho_1(t)$ and that $\rho_2(t) = \rho_1^2(t)V\{\mu(t)\} = c^{(1)}(t)\mu^{(1)}(t)$. We also have that

$$\begin{aligned} E(\epsilon|X,Z) &= 0; && \text{(A.14)} \\ E(\epsilon^2|X,Z) &= E\Big([Y-\mu\{X^{\mathrm{T}}\beta_0+\theta_0(U)\}]^2|X,Z\Big)[\rho_1\{X^{\mathrm{T}}\beta_0+\theta_0(U)\}]^2 \\ &= \mathrm{var}(Y|X,Z)[\rho_1\{X^{\mathrm{T}}\beta_0+\theta_0(U)\}]^2 \\ &= \phi\rho_2(\cdot). && \text{(A.15)} \end{aligned}$$

Let the semiparametric model be denoted as M_0. Consider a parametric submodel M_λ with $f_{X,Z}(X,Z;\nu_1)$, $\theta_0(R + S^{\mathrm{T}}\gamma_0, \nu_2)$ and $\pi(X,Z,\nu_3)$. The joint loglikelihood of Y, X and Z under M_λ is given by

$$\begin{aligned} L(\cdot) &= (\delta/\phi)\Big(Yc\{X^{\mathrm{T}}\beta_0+\theta_0(R+S^{\mathrm{T}}\gamma_0,\nu 2)\} - \mathcal{C}[c\{X^{\mathrm{T}}\beta_0+\theta_0(R+S^{\mathrm{T}}\gamma_0,\nu 2)\}]\Big) \\ &\quad +\delta\mathcal{D}(Y,\phi) + \log\{f_{X,Z}(X,Z,\nu_1)\} \\ &\quad +\delta\log\{\pi(X,Z,\nu_3)\} + (1-\delta)\log\{1-\pi(X,Z,\nu_3)\}. \end{aligned}$$

As before, recall that $\epsilon = \rho_1(\cdot)\{Y-\mu(\cdot)\} = c^{(1)}(\cdot)\{Y-\mu(\cdot)\}$. Then the score functions evaluated at M_0 are

$$\begin{aligned} \partial L/\partial\beta &= \delta X c^{(1)}(\cdot)\{Y-\mu(\cdot)\}/\phi = \delta X\epsilon/\phi \\ \partial L/\partial\gamma &= \delta\theta^{(1)}(U)Sc^{(1)}(\cdot)\{Y-\mu(\cdot)\}\phi = \delta\theta^{(1)}(U)S\epsilon/\phi \\ \partial L/\partial\nu_1 &= s_f(X,Z) \\ \partial L/\partial\nu_2 &= \delta h(U)c^{(1)}(\cdot)\{Y-\mu(\cdot)\}/\phi = \delta h(U)\epsilon/\phi \\ \partial L/\partial\nu_3 &= a(X,Z)\{\delta-\pi(X,Z)\}; \end{aligned}$$

$$\partial L/\partial\phi \;=\; \delta\mathcal{D}_\phi(Y,\phi) - \delta[Yc(\cdot) - \mathcal{C}\{c(\cdot)\}]/\phi^2 = \delta\mathcal{G},$$

where $\mathcal{D}_\phi(Y,\phi)$ is the derivative of $\mathcal{D}(Y,\phi)$ with respect to ϕ, $s_f(X,Z)$ is a mean zero function and $h(U)$ and $a(X,Z)$ are any functions. This means that the tangent space is spanned by

$$\Big(\mathcal{T}_1 = \delta\{S^{\mathrm{T}}\theta_0^{(1)}(U), X^{\mathrm{T}}\}\epsilon/\phi, \mathcal{T}_2 = s_f(X,Z), \mathcal{T}_3 = \delta h(U)\epsilon/\phi,$$
$$\mathcal{T}_4 = a(X,Z)\{\delta - \pi(X,Z)\}, \mathcal{T}_5 = \delta\mathcal{G}\Big).$$

An orthogonal basis of the tangent space is given by $[\mathcal{T}_1 = \delta\mathcal{N}^{\mathrm{T}}\epsilon,\ \mathcal{T}_2 = s_f(X,Z),\ \mathcal{T}_3 = \delta h(U)\epsilon,\ \mathcal{T}_4 = a(X,Z)\{\delta - \pi(X,Z)\}]$ and $\mathcal{T}_5 = \delta\mathcal{G}$; the orthogonality is a straightforward calculation. Now notice that

$$\kappa_0 = \int \mathcal{F}\{x, \theta_0(z;\nu_2), \mathcal{B}_0, \phi_0\} f_{X,Z}(x,z;\gamma)\, dxdz$$

and hence

$$\begin{aligned}
\partial\kappa_0/\partial\beta &= E\{\mathcal{F}_\beta(\cdot)\};\\
\partial\kappa_0/\partial\gamma &= E\{\mathcal{F}_\theta(\cdot)\theta^{(1)}(U)S\};\\
\partial\kappa_0/\partial\nu_1 &= E\{\mathcal{F}(\cdot)s_f(X,Z)\};\\
\partial\kappa_0/\partial\nu_2 &= E[\mathcal{F}_\theta(\cdot)h(Z)];\\
\partial\kappa_0/\partial\nu_3 &= 0;\\
\partial\kappa_0/\partial\phi &= E\{\mathcal{F}_\phi(\cdot)\}.
\end{aligned}$$

As before, we first assume pathwise differentiability to construct the efficient score. We verify this later.

By equation (7) of Newey (1990) there is a random quantity d such that

$$\begin{aligned} E(d\delta X\epsilon/\phi) &= E\{\mathcal{F}_\beta(\cdot)\}; && \text{(A.16)} \\ E\{d\delta\theta^{(1)}(U)S\epsilon/\phi\} &= E\{\mathcal{F}_\theta(\cdot)\theta^{(1)}(U)S\}; && \text{(A.17)} \\ E\{ds_f(X,Z)\} &= E\{\mathcal{F}(\cdot)s_f(X,Z)\}; && \text{(A.18)} \\ E\{d\delta h(U)\epsilon/\phi\} &= E\{\mathcal{F}_\theta(\cdot)h(U)\}; && \text{(A.19)} \\ E[da(X,Z)\{\delta-\pi(X,Z)\}] &= 0; && \text{(A.20)} \\ E(d\delta\mathcal{G}) &= E\{\mathcal{F}_\phi(\cdot)\}. && \text{(A.21)} \end{aligned}$$

Now we compute the projection of d into the tangent space. It is immediate that $\Pi(d|\mathcal{T}_2) = \mathcal{F}(\cdot) - \kappa_0$ and that $\Pi(d|\mathcal{T}_4) = 0$. Since

$$E[\{\delta J(U)\epsilon\}\{\delta h(U)\epsilon/\phi\}] = E\{h(U)\mathcal{F}_\theta(\cdot)\},$$

it is readily shown that $\Pi(d|\mathcal{T}_3) = \delta J(U)\epsilon$. It is a similarly direct calculation to show that $\Pi(d|\mathcal{T}_1) = D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\epsilon$. Finally, $\Pi(d|\mathcal{T}_5) = \delta\mathcal{G}E\{\mathcal{F}_\phi(\cdot)\}/E(\delta\mathcal{G}^2)$.

These calculations thus show that, assuming pathwise differentiability, the efficient influence function for κ_0 is

$$\Psi = D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\epsilon + \mathcal{F}(\cdot) - \kappa_0 + \delta J(U)\epsilon + \delta\mathcal{G}E\{\mathcal{F}_\phi(\cdot)\}/E(\delta\mathcal{G}^2).$$

Hence from (2.15) we see that $\widehat{\kappa}_{\mathrm{SI}}$ has the semiparametric optimal influence function and hence is asymptotically efficient.

A.5.3. Pathwise Differentiability

For $d = D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\epsilon + \mathcal{F}(\cdot) - \kappa_0 + \delta J(U)\epsilon + \delta\mathcal{G}E\{\mathcal{F}_\phi(\cdot)\}/E(\delta\mathcal{G}^2)$, we have to show that (A.16) - (A.21) hold. Let

$$\begin{aligned} d_1 &= D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\epsilon; \\ d_2 &= \mathcal{F}(\cdot) - \kappa_0; \\ d_3 &= \delta J(U)\epsilon; \\ d_4 &= 0; \\ d_5 &= \delta\mathcal{G}E\{\mathcal{F}_\phi(\cdot)\}/E(\delta\mathcal{G}^2). \end{aligned}$$

Then, $d = d_1 + \ldots + d_5$. Since $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, $\mathcal{T}_4$ and $\mathcal{T}_5$ are orthogonal and $d_i \in \mathcal{T}_i$ for $i = 1, \ldots, 5$, we have

$$E(d_1\mathcal{T}_1) = E(D^{\mathrm{T}}\mathcal{Q}^{-1}\delta\mathcal{N}\mathcal{N}^{\mathrm{T}}\epsilon^2) = \phi E(D^{\mathrm{T}}); \tag{A.22}$$

$$E(d_2\mathcal{T}_2) = E[\{\mathcal{F}(\cdot) - \kappa_0\}s_f(X,Z)] = E\{\mathcal{F}(\cdot)s_f(X,Z)\}; \tag{A.23}$$

$$E(d_3\mathcal{T}_3) = E\{\delta J(U)h(U)\epsilon^2\} = E\{\pi(X,Z)J(U)h(U)\phi\rho_2(\cdot)\}; \tag{A.24}$$

$$E(d_4\mathcal{T}_4) = 0; \tag{A.25}$$

$$E(d_5\mathcal{T}_5) = E[\delta\mathcal{G}^2E\{\mathcal{F}_\phi(\cdot)\}/E(\delta\mathcal{G}^2)] = E\{\mathcal{F}_\phi(\cdot)\}; \tag{A.26}$$

$$E(d_i\mathcal{T}_j) = 0, i \neq j. \tag{A.27}$$

To verify (A.16) and (A.17), we have to prove

$$E\begin{bmatrix} d\delta\theta^{(1)}(U)S\epsilon/\phi \\ d\delta X\epsilon/\phi \end{bmatrix}^{\mathrm{T}} = E\begin{bmatrix} \mathcal{F}_\theta(\cdot)\theta^{(1)}(U)S \\ \mathcal{F}_\beta(\cdot) \end{bmatrix}^{\mathrm{T}}$$

Recall that, $\Lambda = \{\theta^{(1)}(U)S^{\mathrm{T}}, X^{\mathrm{T}}\}^{\mathrm{T}}$. So,

$$E\begin{bmatrix} d\delta\theta^{(1)}(U)S\epsilon/\phi \\ d\delta X\epsilon/\phi \end{bmatrix}^{\mathrm{T}} = E(d\delta\Lambda^{\mathrm{T}}\epsilon/\phi)$$

$$
\begin{aligned}
&= E\Big\{d\delta\Big(\mathcal{N}^{\mathrm{T}} + [E\{\delta\rho_2(\cdot)|U_i\}]^{-1}E\{\delta_i\Lambda_i^{\mathrm{T}}\rho_2(\cdot)|U_i\}\Big)\epsilon/\phi\Big\} \\
&= E\{d\delta\mathcal{N}^{\mathrm{T}}\epsilon/\phi\} \\
&\quad + E\Big\{d\delta\Big([E\{\delta\rho_2(\cdot)|U_i\}]^{-1}E\{\delta_i\Lambda_i^{\mathrm{T}}\rho_2(\cdot)|U_i\}\Big)\epsilon/\phi\Big\} \\
&= E(d\mathcal{T}_1/\phi) + E(d\delta h(U)\epsilon/\phi) \\
&= B_1 + B_2,
\end{aligned}
$$

where $h(U) = [E\{\delta\rho_2(\cdot)|U\}]^{-1}E\{\delta\Lambda^{\mathrm{T}}\rho_2(\cdot)|U\}$. Hence, using (A.22), (A.24) and (A.27), we see $B_1 = E(d_1\mathcal{T}_1/\phi) = E(D^{\mathrm{T}})$ and

$$
\begin{aligned}
B_2 &= E(d_3\delta h(U)\epsilon/\phi) \\
&= E\{\pi(X,Z)J(U)h(U)\rho_2(\cdot)\} \\
&= E\{\delta J(U)h(U)\rho_2(\cdot)\} \\
&= E[J(U)h(U)E\{\delta\rho_2(\cdot)|U\}] \\
&= E\Big\{\mathcal{F}_\theta(\cdot)[E\{\delta\rho_2(\cdot)|U\}]^{-1}E\{\delta\Lambda^{\mathrm{T}}\rho_2(\cdot)|U\}\Big\},
\end{aligned}
$$

and hence

$$
\begin{aligned}
B_1 + B_2 &= E(D^{\mathrm{T}}) + E\Big\{\mathcal{F}_\theta(\cdot)[E\{\delta\rho_2(\cdot)|U\}]^{-1}E\{\delta\Lambda^{\mathrm{T}}\rho_2(\cdot)|U\}\Big\} \\
&= E\begin{bmatrix}\mathcal{F}_\theta(\cdot)\theta^{(1)}(U)S \\ \mathcal{F}_\beta(\cdot)\end{bmatrix}^{\mathrm{T}}.
\end{aligned}
$$

To verify (A.19), we use (A.24) and (A.27) and get

$$
\begin{aligned}
E\{d\delta h(U)\epsilon/\phi\} &= E(d_3\mathcal{T}_3/\phi) \\
&= E\{\pi(X,Z)J(U)h(U)\rho_2(\cdot)\} \\
&= E\{\delta J(U)h(U)\rho_2(\cdot)\} \\
&= E[J(U)h(U)E\{\delta\rho_2(\cdot)|U\}] \\
&= E[\mathcal{F}_\theta(\cdot)h(U)].
\end{aligned}
$$

Finally, (A.18) follows directly from (A.23) and (A.27), (A.20) follows directly from (A.25) and (A.27) and (A.21) follows directly from (A.26) and (A.27).

A.6. Proof of Lemma 3

Denote the model under consideration by M_0. Now consider any regular parametric submodel M_λ, with $f_{X,Z}(x,z,\alpha_1)$ and $\theta(z,\alpha_2)$ in place of $f_{X,Z}(x,z)$ and $\theta_0(z)$ respectively. For the model M_λ we have the joint loglikelihood of Y, X and Z,

$$L(y,z,x) = \mathcal{L}(\cdot) + \log\{f_{X,Z}(x,z,\alpha_1)\},$$

where $(\cdot)$ represents the argument $\{Y, X, \theta(Z,\alpha_2), \mathcal{B}_0\}$. The score functions are given by,

$$\begin{aligned}
\partial L/\partial \mathcal{B} &= \mathcal{L}_{\mathcal{B}}(\cdot);\\
\partial L/\partial \alpha_1 &= \partial \log\{f_{X,Z}(x,z,\alpha_1)\}/\partial\alpha_1;\\
\partial L/\partial \alpha_2 &= \mathcal{L}_\theta(\cdot)\partial\theta(z,\alpha_2)/\partial\alpha_2;
\end{aligned}$$

The tangent space is spanned by $S_\lambda = \{\mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}}, s_f(x,z)^{\mathrm{T}}, \mathcal{L}_\theta(\cdot)g(z)^{\mathrm{T}}\}$ or equivalently by

$$\mathcal{T} = \{\mathcal{T}_1 = \mathcal{L}_{\mathcal{B}}(\cdot)^{\mathrm{T}} + \mathcal{L}_\theta(\cdot)\theta_{\mathcal{B}}^{\mathrm{T}}(Z,\mathcal{B}_0) = \epsilon^{\mathrm{T}}, \mathcal{T}_2 = s_f(X,Z)^{\mathrm{T}}, \mathcal{T}_3 = g(Z)^{\mathrm{T}}\mathcal{L}_\theta(\cdot)\}.$$

where $s_f(x,z)$ is any function with expectation 0, and $g(z)$ is any function of z. Note that, under model M_λ, $\kappa_0 = \int Y \exp\{\mathcal{L}(\cdot)\} f_{X,Z}(x,z,\alpha_1) dydxdz$. Hence we have

$$\begin{aligned}
\partial\kappa_0/\partial\mathcal{B} &= E\{Y\mathcal{L}_{\mathcal{B}}(\cdot)\} = E\{Y(\partial L/\partial\mathcal{B})\};\\
\partial\kappa_0/\partial\alpha_1 &= E\{Y s_f(X,Z)\} = E\{Y(\partial L/\partial\alpha_1)\};\\
\partial\kappa_0/\partial\alpha_2 &= E\{Y\mathcal{L}_\theta(\cdot)g(Z)\} = E\{Y(\partial L/\partial\alpha_2)\}.
\end{aligned}$$

Hence we see that κ_0 is pathwise differentiable and $d = Y$. The projection of d into $\mathcal{T}$ is then given by

$$\begin{aligned}
\Pi(d|\mathcal{T}_1) &= E(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\epsilon; \\
\Pi(d|\mathcal{T}_2) &= E(Y|X,Z) - \kappa_0; \\
\Pi(d|\mathcal{T}_3) &= \mathcal{L}_\theta(\cdot)E\{Y\mathcal{L}_\theta(\cdot)|Z\}/E[\{\mathcal{L}_\theta(\cdot)\}^2|Z],
\end{aligned}$$

and hence the efficient influence function is

$$\Pi(d|\mathcal{T}) = E(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\epsilon + \{E(Y|X,Z) - \kappa_0\} + \mathcal{L}_\theta(\cdot)E\{Y\mathcal{L}_\theta(\cdot)|Z\}/E[\{\mathcal{L}_\theta(\cdot)\}^2|Z].$$

But we see that the influence function of the sample mean is $Y - \kappa_0$. Hence the sample mean is semiparametric efficient if and only if (2.19) holds.

A.7. Proof of Lemma 4

It suffices to consider only the case that $\phi = 1$ is known, since the estimates of β_0 and $\theta_0(z)$ do not depend on the value of ϕ.

It is convenient to write $c\{\eta(x,z)\}$ as $d(x,z)$, and to denote the derivative of $d(x,z)$ with respect to $\theta_0(z)$ as $d_\theta(x,z)$. Note that the derivative with respect to β is $d_\beta(x,z) = Xd_\theta(x,z)$. Direct calculations show that

$$\begin{aligned}
\mathcal{L}_\theta(\cdot) &= d_\theta(X,Z)\{Y - \mu(X,Z)\}; \\
\mathcal{L}_\beta(\cdot) &= Xd_\theta(X,Z)\{Y - \mu(X,Z)\}; \\
\theta_\beta(Z) &= -\frac{E[Xd_\theta^2(X,Z)V\{\mu(X,Z)\}|Z]}{E[d_\theta^2(X,Z)V\{\mu(X,Z)\}|Z]}; \\
\epsilon &= \{X + \theta_\beta(Z)\}d_\theta(X,Z)\{Y - \mu(X,Z)\}; \\
E(Y\epsilon) &= E[\{X + \theta_\beta(Z)\}d_\theta(X,Z)V\{\mu(X,Z)\}];
\end{aligned}$$

$$\mathcal{L}_\theta(\cdot)\frac{E\{Y\mathcal{L}_\theta(\cdot)|Z\}}{E\{\mathcal{L}^2_\theta(\cdot)|Z\}} \;=\; \{Y-\mu(X,Z)\}d_\theta(X,Z)\frac{E[d_\theta(X,Z)V\{\mu(X,Z)\}|Z]}{E[d^2_\theta(X,Z)V\{\mu(X,Z)\}|Z]}.$$

If $d_\theta(x,z)$ depends only on z, then $\theta_\beta(Z) = -E[XV\{\mu(X,Z)\}|Z]/E[V\{\mu(X,Z)\}|Z]$, $E(Y\epsilon)=0$ and also

$$1 \equiv d_\theta(X,Z)\frac{E[d_\theta(X,Z)V\{\mu(X,Z)\}|Z]}{E[d^2_\theta(X,Z)V\{\mu(X,Z)\}|Z]}, \tag{A.28}$$

so that by Lemma 3 the sample mean is semiparametric efficient.

The cases that the sample mean is not semiparametric efficient are the following. Consider problems not of canonical exponential forms. First of all, it cannot be semiparametric efficient if $E(Y\epsilon)=0$ and $d_\theta(x,z)$ depends on x, for then (A.28) fails. This means then that $d_\theta(x,z)$ cannot be a function of x, i.e., the data must follow a canonical exponential family.

If $E(Y\epsilon)\neq 0$, we must have that

$$1 \equiv d_\theta(X,Z)\Big(E(Y\epsilon^{\mathrm{T}})\mathcal{M}_1^{-1}\{X+\theta_\beta(Z)\} + \frac{E[d_\theta(X,Z)V\{\mu(X,Z)\}|Z]}{E[d^2_\theta(X,Z)V\{\mu(X,Z)\}|Z]}\Big). \tag{A.29}$$

Examples that (A.29) fails to hold are easily constructed. Because the term inside the parenthesis in (A.29) is linear in X and a function of Z, (A.29) can only hold in principle if $d(x,z) = c\{x^{\mathrm{T}}\beta+\theta(z)\} = a + b\log\{x^{\mathrm{T}}\beta+\theta(z)\}$ for known constants (a,b).

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER III

For simplicity of notation, we first consider only the case that there are no nuisance parameters ζ_0. The more general case is a simple extension and is presented later.

B.1. Proof of Result 3

To prove the results, we rely on several technical conditions that we do not state here explicitly for the sake of saving space. These conditions are well known and standard in smoothing theory. Refer to Claeskens and Van Keilegom (2003), Claeskens and Carroll (2007) and Lin and Carroll (2006) among many others for the details of these assumptions. As stated just before Result 3, we require that the linear expansions for the parametric and nonparametric parts given in Lin and Carroll (2006) hold to order $o_p(n^{-1/2})$, the latter uniformly.

B.1.1. Expansion of $\mathcal{T}_n(\gamma)$

Let $\theta^{(j)}(\cdot)$ be the j^{th} derivative of $\theta(\cdot)$ with respect to z_0. Let $f_Z(z_0)$ be the density function of Z. Make the definitions

$$\begin{aligned}\Omega(z_0) &= E[\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z)\}|Z = z_0];\\ \theta_\eta(z_0, \eta_0) &= -E[S\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z)\}|Z = z_0]/\Omega(z_0).\end{aligned}$$

Note that $S_i + \theta_\eta(Z_i, \eta_0) = \widetilde{S}_i$, and recall

$$\mathcal{M} = -\mathrm{cov}[\{S + \theta_\eta(Z, \eta_0)\}\mathcal{L}_\theta\{Y, S^{\mathrm{T}}\eta_0 + \theta(Z)\}].$$

Then using Lin and Carroll (2006) we have that uniformly in z_0,

$$\widehat{\theta}(z_0,\eta_0)-\theta_0(z_0,\eta_0)=-n^{-1}\textstyle\sum_{i=1}^n K_h(Z_i-z_0)\mathcal{L}_\theta\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}/\{f_Z(z_0)\Omega(z_0)\}$$
$$+(\phi_2 h^2/2)\theta_0^{(2)}(z_0)+O_p\left\{h^4+\log(n)/(nh)\right\}; \qquad \text{(B.1)}$$

$$\widehat{\eta}-\eta_0=-\mathcal{M}^{-1}n^{-1}\textstyle\sum_{i=1}^n \widetilde{S}_i\mathcal{L}_\theta\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta(Z_i)\}+o_p(n^{-1/2}). \qquad \text{(B.2)}$$

The score statistic for β is, via Taylor series,

$$\begin{aligned}\mathcal{T}_{n,\mathrm{adj}}(\gamma) &= n^{-1/2}\textstyle\sum_{i=1}^n\{1+\gamma\theta(Z_i)\}\widetilde{X}_i\mathcal{L}_\theta\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}\\ &\quad +n^{-1/2}\textstyle\sum_{i=1}^n\mathcal{S}_{1i}(\gamma)\{\widehat{\theta}(Z_i)-\theta_0(Z_i)\}\\ &\quad +n^{-1/2}\textstyle\sum_{i=1}^n\mathcal{S}_{2i}(\gamma)(\widehat{\eta}-\eta_0)+o_p(1)\\ &= A_{1n}+A_{2n}+A_{3n}+o_p(1),\end{aligned}$$

where

$$\begin{aligned}\mathcal{S}_{1i}(\gamma) &= \widetilde{X}_i\left[\gamma\mathcal{L}_\theta\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}+\{1+\gamma\theta_0(Z_i)\}\mathcal{L}_{\theta\theta}\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}\right];\\ \mathcal{S}_{2i}(\gamma) &= \gamma\theta_\eta(Z_i)\widetilde{X}_i\mathcal{L}_\theta\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}\\ &\quad +\{1+\gamma\theta_0(Z_i)\}\widetilde{X}_i\widetilde{S}_i^{\mathrm{T}}\mathcal{L}_{\theta\theta}\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}.\end{aligned}$$

By definition of $\widetilde{X}$, it is easy to see that to order $o_p(1)$,

$$A_{2n} = -n^{-1/2}\textstyle\sum_{i=1}^n\mathcal{L}_\theta\{Y_i,S_i^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}E\{\mathcal{S}_{1i}(\gamma)|Z_i\}/\Omega(Z_i)=0,$$

where we have used (3.6) and (B.1). Also, using (B.2) and definition of $\mathcal{N}$ we obtain

$$A_{3n} = -\mathcal{N}\mathcal{M}^{-1}n^{-1/2}\textstyle\sum_{i=1}^n\widetilde{S}_i\mathcal{L}_\theta\{Y_i,S^{\mathrm{T}}\eta_0+\theta_0(Z_i)\}+o_p(1).$$

The result now follows by collecting all the terms. It is readily seen that the expansion is uniform in $\gamma\in[L,R]$.

B.1.2. Weak Convergence

Weak convergence is trivial. Examining the form of the test statistic $\mathcal{T}_{n,\text{adj}}(\gamma)$ in (3.8), we see that it is linear in γ and can be written as $U_n + \gamma V_n$, where (U_n, V_n) are jointly asymptotically normally distributed.

B.2. Proof of Result 4

Define $\Omega(z) = \sum_{j=1}^{J} f_j(z) E\{\mathcal{L}_{jj\theta}(\bullet)|Z_j = z\}$ and

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

where $f_j(z)$ is the density of Z_j and $f_{jk}(z_1, z_2)$ is the bivariate density of (Z_j, Z_k), assumed to have bounded support and are positive on the support. Let $\mathcal{G}(z_1, z_2)$ be the solution to

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

Using the results of Lin and Carroll (2006) we obtain that uniformly in z,

$$\begin{aligned}
\widehat{\theta}(z, \eta_0) - \theta_0(z) &= (\phi h^2/2) b(z) - n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J} K_h(Z_{ij} - z)\mathcal{L}_{ij\theta}(\cdot)/\Omega(z) \\
&\quad + n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{ij\theta}(\cdot)\mathcal{G}(z, Z_{ij})/\Omega(z) \\
&\quad + O_p\{h^4 + \log(n)/(nh)\}; \qquad \text{(B.3)} \\
\widehat{\eta} - \eta_0 &= -\mathcal{M}_1^{-1} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\{S_{ij} + \theta_\eta(Z_{ij}, \eta_0)\}\mathcal{L}_{ij\theta}(\cdot) \\
&\quad + o_p(n^{-1/2}). \qquad \text{(B.4)}
\end{aligned}$$

Define

$$\mathcal{T}_{k,n}(\gamma) = \sum_{j=1}^{J}[X_j\{1+\gamma\theta_0(Z_j)\} + \theta_\beta(Z_j,0,\eta_0,\gamma)]\mathcal{L}_{jk\theta}(\bullet);$$
$$\mathcal{T}_{\eta,n}(\gamma) = \sum_{j=1}^{J}\sum_{k=1}^{J}[X_j\{1+\gamma\theta_0(Z_j)\} + \theta_\beta(Z_j,0,\eta_0,\gamma)]\{S_k+\theta_\eta(Z_k,\eta_0)\}^{\mathrm{T}}\mathcal{L}_{jk\theta}(\bullet).$$

It is easily shown that

$$\begin{aligned}\mathcal{T}_{n,\mathrm{adj}}(\gamma) &= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}[X_{ij}\{1+\gamma\theta_0(Z_{ij})\} + \theta_\beta(Z_{ij},0,\eta_0,\gamma)]\mathcal{L}_{ij\theta}(\bullet)\\ &\quad +n^{-1/2}\sum_{i=1}^{n}\mathcal{T}_{i\eta,n}(\gamma)(\widehat{\eta}-\eta_0)\\ &\quad +n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\{\widehat{\theta}(Z_{ik},\eta_0)-\theta_0(Z_{ik})\}\\ &\quad +n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{ij\theta}(\bullet)\{\widehat{\theta}_\beta(Z_{ij},0,\widehat{\eta},\gamma)-\theta_\beta(Z_{ij},0,\eta_0,\gamma)\}+o_p(1).\end{aligned}$$

Using (B.4) and the fact that $E\{\mathcal{T}_{\eta,n}(\gamma)\} = \mathcal{M}_2$, it is easy to see that

$$\begin{aligned}&n^{-1/2}\sum_{i=1}^{n}\mathcal{T}_{i\eta,n}^{\mathrm{T}}(\gamma)(\widehat{\eta}-\eta_0)\\ &\quad = -\mathcal{M}_2\mathcal{M}_1^{-1}n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\{S_{ij}+\theta_\eta(Z_{ij},\eta_0)\}\mathcal{L}_{ij\theta}(\cdot)+o_p(1).\end{aligned}$$

Next, using (B.3), we now derive that up to terms of $o_p(1)$,

$$\begin{aligned}&n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\{\widehat{\theta}(Z_{ik},\eta_0)-\theta_0(Z_{ik})\}\\ &\quad = -n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\Big[n^{-1}\sum_{r=1}^{n}\sum_{j=1}^{J}K_h(Z_{rj}-Z_{ik})\mathcal{L}_{rj\theta}(\bullet)/\Omega(Z_{ik})\Big]\\ &\qquad +n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\Big[n^{-1}\sum_{r=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{rj\theta}(\bullet)\mathcal{G}(Z_{ik},Z_{rj})/\Omega(Z_{ik})\Big]\\ &\quad = n^{-1/2}\sum_{r=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{rj\theta}(\bullet)\{C_1(Z_{rj})+C_2(Z_{rj})\},\end{aligned}$$

where we define

$$\begin{aligned}C_1(z,\gamma) &= -\sum_{k=1}^{J}f_k(z)E\{\mathcal{T}_{ik,n}(\gamma)|Z_k=z\}/\Omega(z);\\ C_2(z,\gamma) &= E\Big[\sum_{k=1}^{J}E\{\mathcal{T}_{ik,n}(\gamma)|Z_k\}\mathcal{G}(Z_k,z)/\Omega(Z_k)\Big].\end{aligned}$$

We now note that $\sum_{k=1}^{J} f_k(z) E\{\mathcal{T}_{ik,n}(\gamma)|Z_k = z\} = 0$ by definition of $\theta_\beta(\cdot)$ with $\beta_0 = 0$ and hence $C_1(z,\gamma) = C_2(z,\gamma) = 0$.

Finally, we recognize that $\widehat{\theta}_\beta(\cdot)$ is the repeated measures regression of Q_{ij} on Z_{ij} and hence yields an asymptotic expansion similar to (B.3). Together with the fact that $E\{\mathcal{L}_{j\theta}(\cdot)|X,S,Z\} = 0$, it is now straightforward to show that the fourth term in the expansion of $\mathcal{T}_{n,\text{adj}}(\gamma) = o_p(1)$, completing the proof.

B.3. Proof of Result 5

Under the null hypothesis, $\theta_\beta(z,0,\delta_0,\gamma)$ solves

$$0 = \sum_{j=1}^{J} f_j(z) E\Big(\sum_{k=1}^{J}[X_k\{1+\gamma\theta_0(Z_k)\} + \theta_\beta(Z_k,0,\delta_0,\gamma)]\mathcal{L}_{jk\theta}(\cdot)\Big|Z_j = z\Big). \quad \text{(B.5)}$$

Recall that $K_h(z) = h^{-1}K(z/h)$ and $G_h(z) = (1, z/h)^{\mathrm{T}}$. Consider the problem of solving for $\{m(z), m^{(1)}(z)\}$,

$$\begin{aligned} 0 &= n^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{J} K_h(Z_{ij}-z)G(Z_{ij}-z) \\ &\quad \times\Big[\textstyle\sum_{k\neq j=1}^{J} v^{ijk}\{Q_{ik} - m(Z_{ik})\} + v^{ijj}Q_{ij} - v^{ijj}G(Z_{ij}-z)^{\mathrm{T}}\{m(z), m^{(1)}(z)\}^{\mathrm{T}}\Big] \end{aligned}$$

where $v^{ijk} = -\mathcal{L}_{ijk\theta}(\cdot)$. Define $F_n(z) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J} v^{ijj}K_h(Z_{ij}-z)G(Z_{ij}-z)G(Z_{ij}-z)^{\mathrm{T}}$. The solution then satisfies

$$\begin{aligned} F_n(z)\{m(z), m^{(1)}(z)\}^{\mathrm{T}} &= n^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{J} K_h(Z_{ij}-z)G(Z_{ij}-z) \\ &\quad \times\Big[\textstyle\sum_{k\neq j=1}^{J} v^{ijk}\{Q_{ik} - m(Z_{ik})\} + v^{ijj}Q_{ij}\Big]. \end{aligned}$$

Notice that

$$F_n(z) = \sum_{j=1}^{J} E(v^{jj}|Z_j = z)\begin{bmatrix} f_j(z) & 0 \\ 0 & \phi_2 \end{bmatrix} + o_p(1),$$

where $\phi_2 = \int z^2 k(z)\, dz$. Hence, taking limit of both the sides we obtain that $m(z)$ satisfies

$$\begin{aligned}\sum_{j=1}^{J} E(v^{jj}|Z_j = z) f_j(z) m(z) &= \sum_{j=1}^{J} f_j(z) \sum_{k\neq j=1}^{J} E[v^{jk}\{Q_k - m(Z_k)\}|Z_j = z] \\ &\quad + \sum_{j=1}^{J} f_j(z) E(v^{jj} Q_j | Z_j = z),\end{aligned}$$

which is identical to (B.5) with $m(z) = \theta_\beta(z, 0, \delta_0, \gamma)$. This completes the argument.

B.4. Proof of Result 3 With Nuisance Parameters

Let $\theta^{(j)}(\cdot)$ be the j^{th} derivative of $\theta(\cdot)$ with respect to z_0. Let $f_Z(z_0)$ be the density function of Z. Define $\Omega(z_0) = E[\mathcal{L}_{\theta\theta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z = z_0]$ and recall that $\delta_0 = (\eta_0^{\mathrm{T}}, \zeta_0^{\mathrm{T}})^{\mathrm{T}}$ and

$$\begin{aligned}\theta_\delta(z_0, \delta_0) &= -E[\mathcal{L}_{\theta\delta}\{Y, S^{\mathrm{T}}\eta_0 + \theta_0(Z), \zeta_0\}|Z = z_0]/\Omega(z_0); \\ \epsilon &= \mathcal{L}_\delta\{Y, S^{\mathrm{T}}\eta_0 + \theta(Z), \zeta_0\} + \theta_\delta(Z, \delta_0)\mathcal{L}_\theta\{Y, S^{\mathrm{T}}\eta_0 + \theta(Z), \zeta_0\}; \\ \mathcal{M} &= -E(\epsilon\epsilon^{\mathrm{T}}).\end{aligned}$$

Then using Lin and Carroll (2006) we have that uniformly in z_0,

$$\begin{aligned}\widehat{\theta}(z_0, \delta_0) - \theta_0(z_0, \delta_0) = &-n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\}/\{f_Z(z_0)\Omega(z_0)\} \\ &+(\phi_2 h^2/2)\theta_0^{(2)}(z_0) + O_p\left\{h^4 + \log(n)/(nh)\right\}; \qquad \text{(B.6)} \\ \widehat{\delta} - \delta_0 = &-\mathcal{M}^{-1} n^{-1}\sum_{i=1}^{n}\epsilon_i + o_p(n^{-1/2}). \qquad \text{(B.7)}\end{aligned}$$

The score statistic for β is, with a first-order Taylor series,

$$\begin{aligned}\mathcal{T}_{n,\text{gen}}(\gamma) &= n^{-1/2}\sum_{i=1}^{n}\{1 + \gamma\theta_0(Z_i)\}\widetilde{X}_i\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\} \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\mathcal{S}_{1i}(\gamma)\{\widehat{\theta}(Z_i) - \theta_0(Z_i)\} + n^{-1/2}\sum_{i=1}^{n}\mathcal{S}_{2i}(\gamma)(\widehat{\delta} - \delta_0) + o_p(1)\end{aligned}$$

$$= A_{1n} + A_{2n} + A_{3n} + o_p(1),$$

where

$$\begin{aligned}
\mathcal{S}_{1i}(\gamma) &= \widetilde{X}_i[\gamma\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\} + \{1+\gamma\theta_0(Z_i)\}\mathcal{L}_{\theta\theta}\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\}]; \\
\mathcal{S}_{2i}(\gamma) &= \widetilde{X}_i\Big(\{1+\gamma\theta_0(Z_i)\}[\mathcal{L}_{\theta\delta}\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\} \\
&\qquad +\mathcal{L}_{\theta\theta}\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\}\theta_\delta(Z_i, \delta_0)]^{\mathrm{T}} \\
&\qquad +\gamma\theta_\delta^{\mathrm{T}}(Z_i, \delta_0)\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\}\Big).
\end{aligned}$$

Using the fact that $h \propto n^{-\alpha}$ where $1/3 \leq \alpha \leq 1/5$, we obtain, to order $o_p(1)$,

$$A_{2n} = -n^{-1/2}\textstyle\sum_{i=1}^n \mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\}E\{\mathcal{S}_{1i}(\gamma)|Z_i\}/\Omega(Z_i) = 0,$$

where we have used (B.6) and definition of $\widetilde{X}$. Also, using (B.7) we obtain

$$A_{3n} = -\mathcal{N}\mathcal{M}^{-1}n^{-1/2}\textstyle\sum_{i=1}^n \epsilon_i + o_p(1).$$

Collecting all the terms we now see that the score statistic is, up to terms of $o_p(1)$,

$$\mathcal{T}_{\mathrm{adj},n}(\gamma) = n^{-1/2}\textstyle\sum_{i=1}^n \Big[\{1+\gamma\theta_0(Z_i)\}\widetilde{X}_i\mathcal{L}_\theta\{Y_i, S_i^{\mathrm{T}}\eta_0 + \theta_0(Z_i), \zeta_0\} - \mathcal{N}\mathcal{M}^{-1}\epsilon_i\Big].$$

The proof of weak convergence and tightness of $\mathcal{T}_{\mathrm{adj},n}$ follows along the same line as in the main text.

B.5. Proof of Result 4 with Nuisance Parameter

Make the definitions of $\Omega(z)$ and $\mathcal{G}(z_1, z_2)$ as in Section III.4.3 using the general likelihood. We use results from Lin and Carroll (2006) to see that

$$\widehat{\theta}(z, \delta_0) - \theta_0(z) = (\phi h^2/2)b(z) - n^{-1}\textstyle\sum_{i=1}^n\sum_{j=1}^J K_h(Z_{ij} - z)\mathcal{L}_{ij\theta}(\cdot)/\Omega(z)$$

$$+n^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{ij\theta}(\cdot)\mathcal{G}(z,Z_{ij})/\Omega(z)+O_P\{h^4+\log(n)/(nh)\};$$

$$\widehat{\delta}-\delta_0 = -\mathcal{M}_1^{-1}n^{-1}\textstyle\sum_{i=1}^{n}\{\mathcal{L}_\delta(\cdot)+\sum_{j=1}^{J}\mathcal{L}_{ij\theta}(\cdot)\theta_\delta(Z_{ij},\delta_0)\}+o_p(n^{-1/2}).$$

Define

$$\begin{aligned}
\mathcal{T}_{k,n}(\gamma) &= \gamma X_k\mathcal{L}_{k\theta}(\bullet)+\textstyle\sum_{j=1}^{J}[\{1+\gamma\theta_0(Z_j)\}X_j+\theta_\beta(Z_j,0,\delta_0,\gamma)]\mathcal{L}_{jk\theta}(\bullet);\\
\mathcal{T}_{\delta,n}(\gamma) &= \textstyle\sum_{j=1}^{J}\gamma X_j\theta_\delta^{\mathrm{T}}(Z_j,\delta_0)\mathcal{L}_{j\theta}(\bullet)\\
&\quad +\textstyle\sum_{j=1}^{J}[X_j\{1+\gamma\theta_0(Z_j)\}+\theta_\beta(Z_j,0,\delta_0,\gamma)]\\
&\quad \times\{\mathcal{L}_{j\theta\delta}(\bullet)+\textstyle\sum_{k=1}^{J}\theta_\delta(Z_k,\delta_0)\mathcal{L}_{jk\theta}(\bullet)\}^{\mathrm{T}}.
\end{aligned}$$

Using a Taylor's series expansion, $\mathcal{T}_n(\gamma)$ can be written as

$$\begin{aligned}
\mathcal{T}_{n,\mathrm{adj}}(\gamma) &= n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{J}[X_{ij}\{1+\gamma\theta_0(Z_{ij})\}+\theta_\beta(Z_{ij},0,\delta_0,\gamma)]\mathcal{L}_{ij\theta}(\bullet)\\
&\quad +n^{-1/2}\textstyle\sum_{i=1}^{n}\mathcal{T}_{i\delta,n}(\gamma)(\widehat{\delta}-\delta_0)\\
&\quad +n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\{\widehat{\theta}(Z_{ik},\delta_0)-\theta_0(Z_{ik})\}+o_p(1).
\end{aligned}$$

The second term in the right hand side can be written as

$$\begin{aligned}
&n^{-1/2}\textstyle\sum_{i=1}^{n}\mathcal{T}_{i\delta,n}^{\mathrm{T}}(\gamma)(\widehat{\delta}-\delta_0)\\
&\qquad = -\mathcal{M}_2\mathcal{M}_1^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\Big\{\mathcal{L}_\delta(\cdot)+\sum_{j=1}^{J}\mathcal{L}_{ij\theta}(\cdot)\theta_\delta(Z_{ij},\delta_0)\Big\}+o_p(1).
\end{aligned}$$

Using the expansion of $\widehat{\theta}(z,\delta_0)$, we can write the third term, up to terms of order $o_p(1)$, as

$$\begin{aligned}
&n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\{\widehat{\theta}(Z_{ik},\delta_0)-\theta_0(Z_{ik})\}\\
&\quad = -n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\Big[n^{-1}\sum_{r=1}^{n}\sum_{j=1}^{J}K_h(Z_{rj}-Z_{ik})\mathcal{L}_{rj\theta}(\bullet)/\Omega(Z_{ik})\Big]\\
&\qquad +n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{J}\mathcal{T}_{ik,n}(\gamma)\Big[n^{-1}\sum_{r=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{rj\theta}(\bullet)\mathcal{G}(Z_{ik},Z_{rj})/\Omega(Z_{ik})\Big]\\
&\quad = n^{-1/2}\textstyle\sum_{r=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{rj\theta}(\bullet)C_1(Z_{rj},\gamma)+n^{-1/2}\sum_{r=1}^{n}\sum_{j=1}^{J}\mathcal{L}_{rj\theta}(\bullet)C_2(Z_{rj},\gamma),
\end{aligned}$$

where we define

$$\begin{aligned} C_1(z,\gamma) &= -\textstyle\sum_{k=1}^{J} f_k(z) E\{\mathcal{T}_{ik,n}(\gamma)|Z_k = z\}/\Omega(z); \\ C_2(z,\gamma) &= E\Big[\textstyle\sum_{k=1}^{J} E\{\mathcal{T}_{ik,n}(\gamma)|Z_k\}\mathcal{G}(Z_k,z)/\Omega(Z_k)\Big]. \end{aligned}$$

Now we note that by definition of $\theta_\beta(z,\beta_0,\delta_0,\gamma)$ with $\beta_0 = 0$ we have

$$0 = \textstyle\sum_{j=1}^{J} f_j(z) E\Big(\sum_{k=1}^{J}[X\{1+\gamma\theta_0(Z_k)\} + \theta_\beta(Z_k,0,\delta_0,\gamma)]\mathcal{L}_{jk\theta}(\cdot)\Big|Z_j = z\Big).$$

Hence we obtain that $C_1(z,\gamma) = C_2(z,\gamma) = 0$. Now the result follows by collecting all the terms.

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER IV

We first state the required conditions below and then provide a sketch of Result 9.

C.1. Regularity conditions.

We require the following conditions.

A1. Z is absolutely continuous and has compact support $\mathcal{Z}$, its density $f_Z(\cdot)$ is differentiable on $\mathcal{Z}$, the derivative is continuous and $\inf_{z\in\mathcal{Z}} f_Z(z) > 0$. Moreover $\sup_{z\in\mathcal{Z}} |\theta_0(z)| \leq M < \infty$.

A2. Assume that $\mathcal{B} \in \mathbf{B}$, where $\mathbf{B}$ is a compact subset of $\mathcal{R}^k$. For $\mathcal{B} \neq \mathcal{B}' \in \mathbf{B}$, the Kullback-Leibler distance between $\mathcal{L}(Y, X, S, \mathcal{B}, \theta)$ and $\mathcal{L}(Y, X, S, \mathcal{B}', \theta')$ is strictly positive.

A3. $\mathcal{L}(\cdot, x, \cdot, \mathcal{B}, \theta)$ is an entire function with respect to x. Denote the k^{th} derivative with respect to x as $\mathcal{L}^{(k)}(\cdot, x, \cdot, \mathcal{B}, \theta)$, $k = 0, 1, \cdots$. For every (y, x, s) third partial derivatives of $\mathcal{L}^{(k)}(y, x, s, \mathcal{B}, \theta)$ with respect to $\mathcal{B}$ exist and are continuous. Furthermore mixed partial derivatives $\frac{\partial^{r+t}}{\partial\mathcal{B}^r\partial\theta^t}\mathcal{L}^{(k)}(\cdot, \cdot, \cdot, \mathcal{B}, \theta)$ with $0 \leq r, t, \leq 4, r + t \leq 4$, exist for almost all (y, x, s) and $\mathrm{E}\{\sup_{\mathcal{B}\in\mathbf{B}} \sup_{|\theta|\leq M} |\frac{\partial^{r+t}}{\partial\mathcal{B}^r\partial\theta^t}\mathcal{L}(Y, X, S, \mathcal{B}, \theta)|^2\} < \infty$.

A4. The Fisher information matrix

$$G(Z) = \frac{\partial}{\partial(\mathcal{B}^T, \theta^T)^T \partial(\mathcal{B}^T, \theta^T)} \mathrm{E}\{\mathcal{L}(Y, X, S, \mathcal{B}_0, \theta)|_{\theta=\theta_0(Z)}|Z\}$$

possesses a continuous derivative and $\inf_{z\in\mathcal{Z}} G(z) > 0$.

A5. There exists a neighborhood $\mathcal{N}\{\mathcal{B}_0, \theta_0(z)\}$ such that

$$\max_{k=1,2} \sup_{z\in\mathcal{Z}} \left|\left| \sup_{(\mathcal{B},\theta)\in\mathcal{N}\{\mathcal{B}_0,\theta_0(z)\}} \left| \frac{\partial^k}{\partial\theta^k} \mathcal{L}(Y, Z, S, \mathcal{B}, \theta) \right| \right|\right|_{\lambda,z} < \infty$$

for some $\lambda \in (2, \infty]$, where $||\cdot||_{\lambda,z}$ is the L^λ norm, conditioned on $Z = z$.

C.2. Sketch of Result 9

Define $\epsilon^* = Y - W^{\mathrm{T}}\gamma - \theta(Z)$. Direct calculations yield

$$\begin{aligned}
\mathcal{R}(\bullet) &= -\log(\sigma^2)/2 - (\epsilon^{*2} - \gamma^{\mathrm{T}} B^{-1} \sum_{b=1}^{B} V_b V_b^{\mathrm{T}} \gamma)/(2\sigma^2); \\
\mathcal{R}_\beta(\bullet) &= \begin{pmatrix} (W\epsilon^* + B^{-1}\sum_{b=1}^{B} V_b V_b^{\mathrm{T}}\gamma)/\sigma^2 \\ -1/(2\sigma^2) + [\epsilon^{*2} - \gamma^{\mathrm{T}} B^{-1} \sum_{b=1}^{B} V_b V_b^{\mathrm{T}}\gamma]/(2\sigma^4) \end{pmatrix}; \\
\mathcal{R}_\theta(\bullet) &= \epsilon^*/\sigma^2; \\
\mathcal{R}_{\beta\beta}(\bullet) &= \begin{bmatrix} \mathcal{R}_{\beta\beta,11}(\bullet) & \mathcal{R}_{\beta\beta,12}(\bullet) \\ \mathcal{R}_{\beta\beta,21}(\bullet) & \mathcal{R}_{\beta\beta,22}(\bullet) \end{bmatrix}; \\
\mathcal{R}_{\beta\beta,11}(\bullet) &= (-WW^{\mathrm{T}} + B^{-1} \sum_{b=1}^{B} V_b V_b^{\mathrm{T}})/\sigma^2; \\
\mathcal{R}_{\beta\beta,12}(\bullet) &= -[W\epsilon^* + B^{-1} \sum_{b=1}^{B} V_b V_b^{\mathrm{T}}\gamma]/\sigma^4; \\
\mathcal{R}_{\beta\beta,21}(\bullet) &= -[W\epsilon^* + B^{-1} \sum_{b=1}^{B} V_b V_b^{\mathrm{T}}\gamma]/\sigma^4; \\
\mathcal{R}_{\beta\beta,22}(\bullet) &= 1/(2\sigma^4) - [\epsilon^{*2} - \gamma^{\mathrm{T}} B^{-1} \sum_{b=1}^{B} V_b V_b^{\mathrm{T}}\gamma]/\sigma^6; \\
\mathcal{R}_{\beta\theta}(\bullet) &= \begin{pmatrix} -W/\sigma^2 \\ -\epsilon^*/\sigma^4 \end{pmatrix}; \\
\mathcal{R}_{\theta\theta}(\bullet) &= -1/\sigma^2.
\end{aligned}$$

Using these, we see that

$$\theta_\beta = -E\{\mathcal{R}_{\beta\theta}(\bullet)|Z\}/E\{\mathcal{R}_{\theta\theta}(\bullet)|Z\} = -\begin{Bmatrix} E(W|Z) \\ 0 \end{Bmatrix};$$

$$E\{\mathcal{R}_{\beta\beta}(\bullet)\} = \begin{bmatrix} -E(XX^{\mathrm{T}})/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix};$$

$$E\{\mathcal{R}_{\beta\theta}(\bullet)\theta_\beta(Z,\beta)^{\mathrm{T}}\} = \begin{bmatrix} (\sigma^2)^{-1}E\{WE(W|Z)^{\mathrm{T}}\} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} E\{XE(X|Z)^{\mathrm{T}}\}/\sigma^2 & 0 \\ 0 & 0 \end{bmatrix};$$

$$\mathcal{M} = \begin{bmatrix} -\mathcal{S}/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

Also, let

$$\begin{aligned} \mathcal{K} &= \mathcal{R}_\beta + \mathcal{R}_\theta\theta_\beta = (1/\sigma^2)\begin{pmatrix} \{W - E(W|Z)\}\epsilon^* + B^{-1}\sum_{b=1}^B V_b V_b^{\mathrm{T}}\gamma \\ -1/2 + [\epsilon^{*2} - \gamma^{\mathrm{T}}B^{-1}\sum_{b=1}^B V_b V_b^{\mathrm{T}}\gamma]/(2\sigma^2) \end{pmatrix} \\ &= (1/\sigma^2)\begin{pmatrix} \mathcal{K}_1 \\ \mathcal{K}_2 \end{pmatrix}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathrm{cov}(\mathcal{K}_1) &= \mathrm{cov}[\{X - E(X|Z) + U\}(\epsilon - U^{\mathrm{T}}\gamma) + B^{-1}\sum_{b=1}^B V_b V_b^{\mathrm{T}}\gamma] \\ &= \mathrm{cov}[\{X - E(X|Z)\}(\epsilon - U^{\mathrm{T}}\gamma) + U\epsilon - UU^{\mathrm{T}}\gamma + B^{-1}\sum_{b=1}^B V_b V_b^{\mathrm{T}}\gamma] \\ &= \mathrm{cov}[\{X - E(X|Z)\}(\epsilon - U^{\mathrm{T}}\gamma)] + \mathrm{cov}(U\epsilon) + \mathrm{cov}\{(UU^{\mathrm{T}} - \Sigma_{uu})\gamma\} \\ &\quad + \mathrm{cov}\{B^{-1}\sum_{b=1}^B (V_b V_b^{\mathrm{T}} - \Sigma_{uu})\gamma\} \\ &= \mathrm{cov}[\{X - E(X|Z)\}(\epsilon - U^{\mathrm{T}}\gamma)] + \mathrm{cov}(U\epsilon) + \mathrm{cov}\{(UU^{\mathrm{T}} - \Sigma_{uu})\gamma\} \\ &\quad + B^{-1}\mathrm{cov}\{(V_b V_b^{\mathrm{T}} - \Sigma_{uu})\gamma\} \\ &= \Gamma + B^{-1}\mathrm{cov}\{(V_b V_b^{\mathrm{T}} - \Sigma_{uu})\gamma\}. \end{aligned}$$

Also,

$$\begin{aligned} \mathrm{cov}(\mathcal{K}_2) &= (4\sigma^4)^{-1}\{\mathrm{var}(\epsilon^{*2}) + \mathrm{var}(\gamma^{\mathrm{T}}B^{-1}\sum_{b=1}^B V_b V_b^{\mathrm{T}}\gamma)\} \\ &= (4\sigma^4)^{-1}[E\{(\epsilon - U^{\mathrm{T}}\gamma)^4\} - (\sigma^2 + \gamma^{\mathrm{T}}\Sigma_{uu}\gamma)^2 \\ &\qquad + B^{-2}\sum_{b=1}^B \mathrm{var}\{\gamma^{\mathrm{T}}(V_b V_b^{\mathrm{T}} - \Sigma_{uu})\gamma\}] \end{aligned}$$

$$= \quad E\{(\epsilon - U^{\mathrm{T}}\gamma)^2 - (\sigma^2 + \gamma^{\mathrm{T}}\Sigma_{uu}\gamma)\}^2 + B^{-1}\mathrm{var}\{\gamma^{\mathrm{T}}(V_b V_b^{\mathrm{T}} - \Sigma_{uu})\gamma\}],$$

and the result follows from Result 7.

VITA

ARNAB MAITY

Department of Statistics

Texas A&M University

3143 TAMU

College Station, TX 77843-3143

c/o Raymond J. Carroll, Ph.D.

EDUCATION

2008 Ph.D., Statistics, Texas A&M University

2005 Master of Science, Statistics, Texas A&M University

2003 Bachelor of Statistics, Statistics, Indian Statistical Institute

RESEARCH INTERESTS

Kernel methods, Measurement error, Nonparametric regression, Repeated measures, Semiparametric regression.