Locally Efficient Estimators for Semiparametric Models With Measurement Error

Yanyuan MA and Raymond J. CARROLL

We derive constructive locally efficient estimators in semiparametric measurement error models. The setting is one in which the likelihood function depends on variables measured with and without error, where the variables measured without error can be modeled nonparametrically. The algorithm is based on backfitting. We show that if one adopts a parametric model for the latent variable measured with error and if this model is correct, then the estimator is semiparametric efficient; if the latent variable model is misspecified, then our methods lead to a consistent and asymptotically normal estimator. Our method further produces an estimator of the nonparametric function that achieves the standard bias and variance property. We extend the methodology to allow estimation of parameters in the measurement error model by additional data in the form of replicates or instrumental variables. The methods are illustrated through a simulation study and a data example, where the putative latent variable distribution is a shifted lognormal, but concerns about the effects of misspecification of this assumption and the linear assumption of another covariate demand a more model-robust approach. A special case of wide interest is the partial linear measurement error model. If one assumes that the model error and the measurement error are both normally distributed, then our estimator has a closed form. When a normal model for the unobservable variable is also posited, our estimator becomes consistent and asymptotically normally distributed for the general partially linear measurement error model, even without any of the normality assumptions under which the estimator is originally derived. We show that the method in fact reduces to a same estimator as that of Liang et al., thus demonstrating a previously unknown optimality property of their method.

KEY WORDS: Errors in variables; Estimating equations; Instrumental variables; Kernel regression; Latent variables; Local efficiency; Measurement error; Nonparametric regression; Partially linear model; Semiparametric methods.

1. INTRODUCTION

A common practice in facilitating increased model flexibility is through nonparametric modeling, resulting in widely used semiparametric models including partially linear models, generalized partially linear models, or semiparametric models containing a single index component. Measurement error problems in such a context are less well studied than their parametric counterparts, probably due to the difficulty of handling multiple infinite-dimensional parameters. In this article we consider a class of such semiparametric measurement error models. We construct estimators for the parametric part of the model that are root-n consistent and asymptotically normally distributed and also for the nonparametric part of the model that enjoys the usual bias and variance properties of the nonparametric estimation. We assume a parametric specification for the measurement error part of the model. The methods are based on a further parametric specification for the latent variable; if this model specification holds, then our methods are semiparametric efficient, whereas if the latent variable model is misspecified, then we still obtain root-n–consistent and asymptotically normal estimators. As far as we know, this is the first article on semiparametric measurement error models that proposes a general methodology for consistently estimating parametric and nonparametric parts without having to resort to a deconvolution method or to correctly specify a distributional model for the variable measured with error.

An example of such problem is as follows. In the Framingham Heart Study data (Kannel et al. 1986), consider a logistic regression of coronary heart disease, Y, on true systolic blood pressure, X, and age, Z, among the nonsmokers. The main interest is in the effect of systolic blood pressure on coronary heart disease. A model that allows for a flexible shape in age is

$$\Pr(Y = 1|X, Z) = H\{\mathcal{B}X + \theta(Z)\}, \tag{1}$$

where $H(\cdot)$ is the logistic distribution function. Of course, true systolic blood pressure is not observable, and instead we observe W, measured blood pressure. As described by Carroll, Ruppert, Crainiceanu, and Stefanski (2006, chap. 5), a reasonable model relating W and X is

$$\log(W - 50) = \log(X - 50) + U, \tag{2}$$

where U is normally distributed with mean 0 and variance σ_u^2: Carroll et al. estimated $\sigma_u^2 = .0126$ based on 1,615 df, so that for the purposes of illustration, we consider σ_u^2 as known. Although earlier analysis (Carroll et al. 2006) has assumed a linear model for the age effect, our analysis will show that the linearity assumption is somewhat violated, and hence the inclusion of an arbitrary function $\theta(Z)$ is needed.

There are various strategies for analyzing the model (1)–(2). An obvious and reasonable approach for this particular dataset is to make the further assumption that $\log(X - 50)$ is normally distributed independently of Z, although a moments analysis suggests that the distribution of this transformed variable is heavier-tailed than the normal, with a kurtosis of approximately 9.0. We then have a fully specified semiparametric model, and thus we would typically apply a standard semiparametric method, such as profile likelihood (Severini and Staniswalis 1994) or backfitting (Chen, Linton, and van Keilegom 2003).

The main point of our article is illustrated by the following considerations. Assuming that $\log(X - 50)$ is normally

Yanyuan Ma is Assistant Professor (E-mail: *ma@stat.tamu.edu*) and Raymond J. Carroll is Distinguished Professor (E-mail: *carroll@stat.tamu.edu*), Department of Statistics, Texas A&M University, College Station TX 77843. Ma's work was supported by a grant from the National Cancer Institute (CA74552). Carroll's work was supported by a grant from the National Cancer Institute (CA57030) and by the Texas A&M Center for Environmental and Rural Health through a grant from the National Institute of Environmental Health Sciences (P30-ES09106). The work was done during a visit by the authors to the Centre of Excellence for Mathematics and Statistics of Complex Systems at the Australian National University, whose support is gratefully acknowledged. The authors thank Naisyin Wang for many helpful comments. They also thank a referee for a very detailed reading of the manuscript and for pointing out the connection with instrumental variables.

Journal of the American Statistical Association
December 2006, Vol. 101, No. 476, Theory and Methods
DOI 10.1198/016214506000000519

distributed implies the assumption that X is a shifted lognormal random variable. However, there is inevitable concern that the analysis will be sensitive to this assumption; the fact that $\log(X - 50)$ has a kurtosis greater than 8.0 indicates a t-distribution with approximately 5 df. A good discussion of this issue has been given by Gustafson (2004, chap. 4.6). This concern is a major motivation for the class of functional measurement error methods, including the SIMEX estimator of Cook and Stefanski (1995), which is an approximately consistent estimator. In contrast, we seek methods that are fully consistent and semiparametric efficient when the shifted lognormal assumption is true and remain fully consistent when the assumption is false.

Our method is based on a computationally convenient backfitting method for estimating $\theta(\bullet)$, in conjunction with kernel-based local polynomial methods. Denote the response variable as Y, the predictor measured with error as X, and predictors measured without error as (S, Z). The likelihood function for Y given (X, S, Z) is

$$[Y|X, S, Z] = p\{y|x, s, z, \mathcal{B}, \theta(z)\} \tag{3}$$

for some unknown function $\theta(z)$ and parameter $\mathcal{B}$. But instead of observing X, we observe W, which is conditionally independent of Y given (X, S, Z). The likelihood function of W given (X, S, Z) is $p(w|x, s, z, \gamma_{\text{mem}})$ depending on a parameter γ_{mem}. Often γ_{mem} can be estimated using additional information. Here we separate the covariates measured without error into S and Z to allow both parametric and nonparametric entry of these covariates. Throughout the article, S can be ignored without affecting understanding of the methodology.

To complete a parametric likelihood specification, we need a model for the unobservable X given (S, Z), which we denote by $p_{\text{c}}(x|s, z, \xi_{\text{latent}})$ depending on a parameter ξ_{latent}, where the subscript means conjectured. We assume that ξ_{latent} can be estimated at root-n rate by an estimator, $\widehat{\xi}_{\text{latent}}$. We show how to construct estimators of $\mathcal{B}$ such that the following conditions hold:

- Whether or not $p_{\text{c}}(x|s, z, \xi_{\text{latent}})$ is correct, the estimator is consistent and asymptotically normally distributed, with limiting distribution independent of the method for estimating ξ_{latent}. If $p_{\text{c}}(x|s, z, \xi_{\text{latent}})$ is correct, then the estimator is semiparametric efficient.
- For any chosen distributional model of X given (S, Z), the estimator achieves the minimal estimation variance under such a model. That is, no further improvement for estimating $\mathcal{B}$ can be achieved through improved estimation of $\theta(Z)$.

One interesting example is the partially linear model with measurement error $Y = X\beta + \theta(Z) + \epsilon$ and $W = X + U$, where both ϵ and U are assumed to be normal. When $\theta(Z)$ is replaced by a constant θ in this model, Stefanski and Carroll (1987, eq. 3.5) derived an efficient estimator. We generalize this work to the partially linear model, deriving the semiparametric efficient estimator. When the latent variable X is also assumed to be normal, the resulting estimator is explicit and enjoys the robustness property of being consistent and asymptotically normal even if *all* of the normality assumptions are violated. We also show that this estimator is the same as one proposed by Liang, Härdle, and Carroll (1999), thus characterizing their estimator in terms of the optimality/suboptimality property under different conditions.

The article is organized as follows. In Section 2 we describe the estimating equations approach for the parametric measurement error models of Tsiatis and Ma (2004) and then define our methodology. Although our backfitting estimator can build on any consistent estimator of a parametric measurement error model, we choose to use their estimator due to the estimator's general applicability and its local efficiency. We give our main results in Section 3, with the limiting distribution of the estimator in Section 3.1, the local efficiency property in Section 3.2, the partially linear model in Section 3.3, and implementation of the methods in Section 3.4. In Section 3.5 we describe how the results change if the distribution of the measurement error needs to be estimated. We focus on additive measurement error, with either replicates of W (Sec. 3.6) or instruments (Sec. 3.7). In Section 4 we describe a simulation study, and in Section 5 we analyze the Framingham data. We give concluding remarks in Section 6, and collect all technical details in an Appendix.

2. PARAMETRIC ESTIMATING FUNCTIONS AND METHODS

Let $\mathcal{Y} = (Y, W, S, Z)$ be the observed data. Consider the special situation of (3) when $\theta(z) \equiv \alpha$. Let $\mathcal{B}_0$ and α_0 be the true parameters in this model, and let $\theta_0(z)$ be the true function. Thus $\theta_0(z) = \alpha_0$. In what follows, for simplicity we assume that the true value for the parameter γ_{mem} in the model for W given (X, S, Z), $\gamma_{0,\text{mem}}$, is known, and hence we suppress γ_{mem}; see Section 3.5 for a discussion on how to relax this condition. We also assume that there is an estimator $\widehat{\xi}_{\text{latent}}$ such that $n^{1/2}(\widehat{\xi}_{\text{latent}} - \xi_{*,\text{latent}}) = O_p(1)$ for some $\xi_{*,\text{latent}}$; if the model for X given (S, Z) is correctly specified, then $\xi_{*,\text{latent}} = \xi_{0,\text{latent}}$, the true value of ξ_{latent}.

Tsiatis and Ma (2004) constructed estimating functions $\mathcal{L}_{\mathcal{B}}(\bullet)$ for $\mathcal{B}$ and $\Psi_\theta(\bullet)$ for α that identify $(\mathcal{B}_0, \alpha_0)$. Here we use the notation Ψ_θ instead of Ψ_α to hint at the effect of Ψ_θ on estimating $\theta(Z)$ in the nonparametric case. Let $\mathcal{S}_{\mathcal{B}}(\bullet)$ and $\mathcal{S}_\theta(\bullet)$ be the log-likelihood scores for $\mathcal{B}$ and θ computed under $p_{\text{c}}(x|s, z)$, the conjectured model for X given (S, Z), that is, $\mathcal{S}_{\mathcal{B}}(\bullet) = \partial \log \int p_{\text{c}}(x|s, z)p(w|x, s, z)p(y|x, s, z)\, dx/\partial\mathcal{B}$ and $\mathcal{S}_\theta(\bullet) = \partial \log \int p_{\text{c}}(x|s, z)p(w|x, s, z)p(y|x, s, z)\, d\mu(X)/\partial\theta$. Let expectations computed under the true model and the assumed model for X given (S, Z) be denoted by "E" and "E_*." Then there exist functions $a_{\mathcal{B}}(X, S, Z)$ and $a_\theta(X, S, Z)$ such that

$$E\{\mathcal{S}_{\mathcal{B}}(\bullet)|X, S, Z\} = E\big[E_*\{a_{\mathcal{B}}(X, S, Z)|\mathcal{Y}\}|X, S, Z\big] \tag{4}$$

and

$$E\{\mathcal{S}_\theta(\bullet)|X, S, Z\} = E\big[E_*\{a_\theta(X, S, Z)|\mathcal{Y}\}|X, S, Z\big], \tag{5}$$

and form $\mathcal{L}_{\mathcal{B}}(\bullet) = \mathcal{S}_{\mathcal{B}}(\bullet) - E_*\{a_{\mathcal{B}}(X, S, Z)|\mathcal{Y}\}$ and $\Psi_\theta(\bullet) = \mathcal{S}_\theta(\bullet) - E_*\{a_\theta(X, S, Z)|\mathcal{Y}\}$. In particular, it follows that

$$0 = E\{\mathcal{L}_{\mathcal{B}}(\mathcal{Y}, \mathcal{B}_0, \alpha_0, \xi_{*,\text{latent}})|S, Z\} \tag{6}$$

and

$$0 = E\{\Psi_\theta(\mathcal{Y}, \mathcal{B}_0, \alpha_0, \xi_{*,\text{latent}})|S, Z\}. \tag{7}$$

Equations (6) and (7) form the backbone of our method that allows for a general unknown function $\theta(z)$. Let $K(z)$ be a

smooth symmetric density function, let h be a bandwidth, and define $K_h(z) = h^{-1}K(z/h)$. Then for every $(\mathcal{B}, \xi_{\text{latent}})$, define $\widehat{\theta}(z_0, \mathcal{B}, \xi_{\text{latent}})$ as the solution α to the local constant estimating equation

$$0 = \sum_{i=1}^{n} K_h(Z_i - z_0)\Psi_\theta\{\mathcal{Y}_i, \mathcal{B}, \alpha, \xi_{\text{latent}}\}. \tag{8}$$

The estimate $\widehat{\mathcal{B}}$ of $\mathcal{B}$ is defined as the solution to

$$0 = \sum_{i=1}^{n} \mathcal{L}_{\mathcal{B}}\{\mathcal{Y}_i, \mathcal{B}, \widehat{\theta}(Z_i, \mathcal{B}, \widehat{\xi}_{\text{latent}}), \widehat{\xi}_{\text{latent}}\}. \tag{9}$$

Equations (8) and (9) represent a type of backfitting algorithm, as opposed to the more commonly studied profile likelihood approaches (see Chen et al. 2003).

3. MAIN RESULTS

Our results split into a series of steps. We first describe the limiting distribution of the estimates of $\mathcal{B}_0$ when $\gamma_{0,\text{mem}}$ is known; recall that γ_{mem} is the parameter associated with the measurement error model and $\gamma_{0,\text{mem}}$ is its true value. We then describe the local semiparametric efficiency of our methods. Finally, we indicate the modifications necessary for the case where that γ_{mem} must be estimated. We state all of the results in the case where we use a local constant estimator for $\theta(Z)$.

3.1 Main Asymptotic Expansion

Set the definitions that $\mathcal{L}_{\mathcal{B}\mathcal{B}}$ is the partial derivative of $\mathcal{L}_{\mathcal{B}}$ with respect to $\mathcal{B}$, $\mathcal{L}_{\mathcal{B}\theta}$ is the partial derivative of $\mathcal{L}_{\mathcal{B}}$ with respect to θ, $\Psi_{\theta\theta}$ is the partial derivative of Ψ_θ with respect to θ, and $\Psi_{\theta\mathcal{B}}$ is the partial derivative of Ψ_θ with respect to $\mathcal{B}$. With the argument $(\bullet)$ being $\{\mathcal{Y}, \mathcal{B}_0, \theta_0(Z), \xi_{*,\text{latent}}\}$, also define $\Omega(Z) = E\{\Psi_{\theta\theta}(\bullet)|Z\}$, $\mathcal{U}(Z) = E\{\mathcal{L}_{\mathcal{B}\theta}(\bullet)|Z\}/\Omega(Z)$, and $\theta_{\mathcal{B}}(Z) = -E\{\Psi_{\theta\mathcal{B}}(\bullet)|Z\}/\Omega(Z)$. Further define that

$$\mathcal{F} = E\big[\mathcal{L}_{\mathcal{B}\mathcal{B}}\{\mathcal{Y}, \mathcal{B}_0, \theta_0(Z), \xi_{*,\text{latent}}\} + \mathcal{L}_{\mathcal{B}\theta}\{\mathcal{Y}, \mathcal{B}_0, \theta_0(Z), \xi_{*,\text{latent}}\}\theta_{\mathcal{B}}^{\mathrm{T}}(Z, \mathcal{B}_0)\big]. \tag{10}$$

Theorem 1. Let $\mathcal{L}_{i,\mathcal{B}}(\bullet) = \mathcal{L}_{\mathcal{B}}(\mathcal{Y}_i, \mathcal{B}_0, \theta_0(Z_i), \xi_{*,\text{latent}}\}$, and similarly for other terms. Assume that the bandwidth h satisfies $nh^4 \to 0$ and $nh^2 \to \infty$. Then, whether the model for X given (S, Z) is specified correctly or not, the backfitting estimator $\widehat{\mathcal{B}}$ has the asymptotic expansion

$$-\mathcal{F}n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) = n^{-1/2}\sum_{i=1}^{n}\{\mathcal{L}_{i,\mathcal{B}}(\bullet) - \Psi_{i,\theta}(\bullet)\mathcal{U}(Z_i)\} + o_p(1). \tag{11}$$

Hence $n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0)$ is asymptotically normally distributed with mean 0 and covariance matrix $\mathcal{F}^{-1}\Sigma\mathcal{F}^{-\mathrm{T}}$, where $\Sigma = \text{cov}\{\mathcal{L}_{\mathcal{B}}(\bullet) - \Psi_\theta(\bullet)\mathcal{U}(Z)\}$.

In Theorem 1, the requirement that $nh^4 \to 0$ is the undersmoothing typically required for backfitting, a direct result of the bias of the local constant estimator.

Remark 1. Theorem 1 states that regardless of the correctness of the conjectured model for X given (S, Z), the estimator is guaranteed to yield a root-n consistent estimator for $\mathcal{B}$. Note that although the conditional expectations involved in $\mathcal{L}_{\mathcal{B}}$ and Ψ_θ can be under either the true or false model of $p(X|S, Z)$, the expectations involved in $\Omega(Z)$, $\theta_{\mathcal{B}}(Z)$, and $U(Z)$ must be calculated under the true $p(X|S, Z)$ for (11) to hold. Also note that $\theta_{\mathcal{B}}(Z) = -U(Z)$ if $\mathcal{L}_{\mathcal{B}}$ and Ψ_θ are calculated under the true model.

Remark 2. There are numerous possibilities for performing inference about $\mathcal{B}_0$, Chen et al. (2003) described conditions under which the bootstrap will be asymptotically valid for backfitting estimators. Alternatively, estimating equation ideas can be used. The term $\mathcal{F}$ is easily estimated by the sample average of the terms in (10). The difficulty in implementing an estimate of Σ via sandwich-type ideas is that it relies on an estimate of $\mathcal{U}(Z)$, and this depends on the *true* model for X given (S, Z). The device that we used was to estimate $\Omega(Z)$ and $\mathcal{U}(Z)$ by simple nonparametric regression devices, and then define $\widehat{\Sigma}$ to be the sample covariance matrix of the terms $\widehat{\mathcal{L}}_{i,\mathcal{B}}(\bullet) - \widehat{\Psi}_{i,\theta}(\bullet)\widehat{\mathcal{U}}(Z_i)$.

Remark 3. After obtaining the root-n consistent estimator $\widehat{\mathcal{B}}$, we can select a more standard bandwidth, $h = O(n^{-1/5})$, and perform one more iteration of (8) to obtain a nonparametric estimator of $\theta(Z)$ with standard bias and variance properties.

3.2 Local Semiparametric Efficiency

In our model, there are two infinite-dimensional nuisance parameters, $\theta(\bullet)$ and the model for X given (S, Z). In Theorem 2 we show that the optimal efficiency with respect to $\theta(\cdot)$ is already achieved in our estimator; the optimal efficiency with respect to the model of $p(X|S, Z)$ can be achieved if we posit the model correctly. One implication of Theorem 2 is that there is no further improvement possible by, for example, better estimating $\theta(Z)$.

Theorem 2. The backfitting estimator $\widehat{\mathcal{B}}$ is locally efficient with respect to the assumed density function $p(x|s, z, \xi_{\text{latent}})$ for X given (S, Z). That is, it is semiparametric efficient if the conjectured model for X given (S, Z) is correct, and is consistent if the conjectured model for X given (S, Z) is incorrect.

The proof is given in the Appendix. The calculations leading to the proof of Theorem 2 are also useful in showing that when the conjectured model for X is true, the limiting covariance matrix of the estimate of $\mathcal{B}_0$ has a simple and familiar form. We state the result in Corollary 1 and give the proof in the Appendix.

Corollary 1. If the conjectured model is correct, then $\mathcal{F} = -\Sigma$, and the asymptotic covariance matrix of $n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0)$ is $-\mathcal{F}^{-1} = \Sigma^{-1}$.

Remark 4. The referee asked us to compare our work here with that of Liang, Wang, Robins, and Carroll (2004). Those authors worked in a missing-data context instead of a measurement error context, with the same two infinite-dimensional components. They computed a double projection and found it infeasible to implement. Thus they resorted to an ad hoc

but computationally convenient approach that lacked full efficiency. We too gave up the double-projection approach and replaced one of these projections with backfitting; however, our method is locally semiparametric efficient. The methods of proof in the two approaches are completely different.

3.3 Special Case: The Partially Linear Model

The partially linear model is $Y = X^{\mathrm{T}}\beta + \theta(Z) + \epsilon$, where $\epsilon = \text{normal}(0, \sigma_\epsilon^2)$. The parameter $\mathcal{B} = (\beta^{\mathrm{T}}, \sigma_\epsilon^2)^{\mathrm{T}}$. We assume that $W = X + U$, where $U = \text{normal}(0, \xi_{\text{latent}})$. As we shall see, the normality assumption for ϵ and U is used only in the construction of the estimators and does not affect their validity.

Define $\delta = W + Y\xi_{\text{latent}}\beta/\sigma_\epsilon^2$. Following Stefanski and Carroll (1987), the forms of $\mathcal{L}_\mathcal{B}$ and Ψ_θ are calculated as

$$\mathcal{L}_\mathcal{B}(\bullet) = (A_1^{\mathrm{T}}, A_2)^{\mathrm{T}}, \tag{12}$$

$$A_1 = \left\{Y - \frac{\delta^{\mathrm{T}}\beta + \theta}{1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2}\right\}E_*(X|\delta),$$

$$A_2 = Y^2(1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2) - 2Y(\theta + \delta^{\mathrm{T}}\beta) - \sigma_\epsilon^2 + \frac{(\delta^{\mathrm{T}}\beta + \theta)^2}{1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2},$$

where A_1 is the component associated with β and A_2 is associated with σ_ϵ^2. E_* is computed under the assumed model. In addition,

$$\Psi_\theta(\bullet) = Y - \frac{\delta^{\mathrm{T}}\beta + \theta}{1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2}. \tag{13}$$

This is a locally efficient estimator when both $p(W|X)$ and $p(Y|X, Z)$ are assumed to be normal. Expanding δ in terms of W and Y, (12) and (13) become

$$A_1 = \frac{Y - W^{\mathrm{T}}\beta - \theta}{1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2}E_*(X|\delta),$$

$$A_2 = \frac{(Y - W^{\mathrm{T}}\beta - \theta)^2 - \sigma_\epsilon^2 - \beta^{\mathrm{T}}\xi_{\text{latent}}\beta}{1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2},$$

$$\Psi_\theta = \frac{Y - W^{\mathrm{T}}\beta - \theta}{1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2}.$$

If we posit a suitable model on X (e.g., X is normal) so that $E_*(X|\delta)$ is a linear function of δ but is otherwise functionally independent of $\theta(\cdot)$, say $a + b\delta$, then we find that

$$A_1 = \frac{a(Y - W^{\mathrm{T}}\beta - \theta)}{1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2} + \{a_\theta(YW - WW^{\mathrm{T}}\beta - \theta W + Y^2\sigma_\epsilon^{-2}\xi_{\text{latent}}\beta - Y\sigma_\epsilon^{-2}\beta^{\mathrm{T}}W\xi_{\text{latent}}\beta - Y\sigma_\epsilon^{-2}\theta\xi_{\text{latent}}\beta)\} /(1 + \beta^{\mathrm{T}}\xi_{\text{latent}}\beta/\sigma_\epsilon^2).$$

Specifically, when the model for X is posited to be normal(μ, Ω), the expressions for a and b are as given in the Appendix. The fact that b does not depend functionally on θ will be useful later on. Under such conditions, it can be verified that A_1, A_2, and Ψ_θ all have mean 0 even if we do not have the normality for Y or W conditional on X. This implies that as long as we posit a normal model for X, our estimator is always consistent. In addition, if the true model $p(X)$, $p(W|X)$, and $p(Y|X, Z)$ are all normal, then our estimator is also efficient.

Although the estimator proposed here appears to be very different from that of the one in Liang et al. (1999), it is exactly the same estimator. See the Appendix for a proof of the equivalence.

3.4 Implementation

Implementation our method involves iterations of estimating $\theta(z_0)$ at $z_0 = Z_1, \ldots, Z_n$ from solving (8) and estimating $\mathcal{B}$ from solving (9). Equations (8) and (9) are usually solved through a Newton–Raphson algorithm, which would only require evaluating $\mathcal{L}_\mathcal{B}$ and Ψ_θ at fixed values of θ and $\mathcal{B}$. However, except for some special situations, such as in generalized linear models, $\mathcal{L}_\mathcal{B}$ and Ψ_θ do not have a closed form. In fact, each evaluation of $\mathcal{L}_\mathcal{B}$ and Ψ_θ requires solving for $a_\mathcal{B}$ and a_θ from the integral equations (4) and (5).

We propose solving these integral equations using a discretization technique. To remain focused, we describe details of solving for $a_\mathcal{B}(X, S, Z)$; the same technique applies to $a_\theta(X, S, Z)$ as well. We consider a set of grid points $x_1, \ldots, x_m$ and try to obtain $a_\mathcal{B}(x_i, S, Z)$ for each observed value of (S, Z). Under this scheme, denote $p_i(\mathcal{Y}) = p_{X,S,Z|\mathcal{Y}}(x_i, S, Z|\mathcal{Y})$, in which case (4) becomes

$$\sum_{i=1}^{m} a_\mathcal{B}(x_i, S, Z)E\{p_i(\mathcal{Y})|X, S, Z\} = E\{S_\mathcal{B}(\mathcal{Y})|X, S, Z\}. \tag{14}$$

Setting $X = x_1, \ldots, x_m$ in (14) thus will provide m linear equations, and we subsequently solve the m-equation linear system to obtain $a_\theta(x_i, S, Z)$, $i = 1, \ldots, m$.

The implementation of the discretization technique corresponds to positing a discrete model for $p(X|S, Z)$ and carrying out the computation under this posited model. Note that to ensure sufficient accuracy, we need to make sure that the discretization points are sufficiently dense on the support of the true $p(X|S, Z)$. Alternative methods for solving the integral equation also exist (see Kress 1999, chap. 15). Any suitable numerical method for solving integral equations can be used to solve (4) and (5), but determining the specific method to use must be based on the specific problem.

3.5 Estimation of the Measurement Error Distribution

In practice, the parameter $\gamma_{0,\text{mem}}$ governing the measurement error distribution of W given (X, S, Z) may be unknown and must be estimated. We discuss two methods for doing this in the situation of additive normal error. The first method applies to cases with partial or complete replication. For this problem, our methodology is readily extended, and asymptotics follow from our previous results with only a slight change in notation. The second method involves using instrumental variables. It is not generally well known that instruments allow for estimation of the measurement error variance, but this is possible.

3.6 The Use of Replicates

One basic idea is to form an estimating function for $\gamma_{0,\text{mem}}$ and then append it to the estimating function $\mathcal{L}_\mathcal{B}(\bullet)$. Consider, for example, the standard additive measurement error model $W_{ij} = X_i + U_{ij}$, where $j = 1, \ldots, m$, and assume that

$U_{ij} \sim \text{normal}(0, \gamma_{0,\text{mem}})$ are independent and identically distributed. In this case, define $\psi_{\gamma_{\text{mem}}}(W_{i1}, \ldots, W_{im}, \gamma_{\text{mem}}) = (m-1)^{-1}\sum_{j=1}^{m}(W_{ij} - \overline{W}_i)^2 - \gamma_{\text{mem}}$. Then, in place of $\mathcal{L}_{\mathcal{B}}(\mathcal{Y}, \mathcal{B}, \gamma_{\text{mem}}, \theta, \xi_{\text{latent}})$, we use

$$\begin{aligned}\mathcal{L}_{\mathcal{B},\text{mem}}&(\mathcal{Y}, \mathcal{B}, \gamma_{\text{mem}}, \theta, \xi_{\text{latent}})\\ &= \left\{\mathcal{L}_{\mathcal{B}}^{\mathrm{T}}(\mathcal{Y}, \mathcal{B}, \gamma_{\text{mem}}, \theta, \xi_{\text{latent}}),\right.\\ &\qquad \left.\psi_{\gamma_{\text{mem}}}(W_{i1}, \ldots, W_{im}, \gamma_{\text{mem}})\right\}^{\mathrm{T}}.\end{aligned}$$

Our results in Theorem 1 apply with $\mathcal{L}_{\mathcal{B}}(\bullet)$ simply replaced by $\mathcal{L}_{\mathcal{B},\text{mem}}(\bullet)$.

This plug-in method enjoys the robustness properties of our general methodology, but it need not be semiparametric efficient, wherein we would need to include the unknown $\gamma_{0,\text{mem}}$ in the parameter of interest $\mathcal{B}$ and change the full-data likelihood to incorporate the replicates. This requires a rederivation of $\mathcal{L}_{\mathcal{B}}$ and Ψ_θ. Let $W_i = (W_{i1}, \ldots, W_{im})^{\mathrm{T}}$ and let $\mathcal{Y}_i = (Y_i, W_i, S_i, Z_i)$ be the observed data. Replace the parameter $\mathcal{B}$ with β and define the new $\mathcal{B} = (\beta^{\mathrm{T}}, \gamma_{\text{mem}}^{\mathrm{T}})^{\mathrm{T}}$. Using this new notation, the form of the observed data likelihood does not change. Thus Ψ_θ has exactly the same expression as in the case where γ_{mem} is known; $\mathcal{L}_{\mathcal{B}} = (\mathcal{L}_\beta^{\mathrm{T}}, \mathcal{L}_\gamma^{\mathrm{T}})^{\mathrm{T}}$, where $\mathcal{L}_\beta$ has exactly the same expression as $\mathcal{L}_{\mathcal{B}}$ in the known γ_{mem} case and $\mathcal{L}_\gamma$ is obtained by replacing all of the

$$E\left[\frac{\partial \log\{p(Y_i|X_i, S_i, Z_i, \beta, \theta)\}}{\partial\beta}\Big| W_i, Y_i, S_i, Z_i\right] \quad \text{with}$$

$$E\left[\frac{\partial \log\{p(W_i|X_i, S_i, Z_i, \gamma_{\text{mem}})\}}{\partial\gamma_{\text{mem}}}\Big| W_i, Y_i, S_i, Z_i\right]$$

in forming $\mathcal{L}_\beta$ while keeping everything else unchanged.

3.7 Use of Instrumental Variables

A referee asked about the connection between our methodology and that of the instrumental variables work of Carroll, Ruppert, Tosteson, Crainiceanu, and Karagas (2004). Although there are many differences between our work and theirs, there is one interesting connection, showing that our methodology is also applicable to the instrumental variables problem. Suppose that we have an additive measurement error model $W = X + U$, where U is normally distributed with mean 0 and variance σ_u^2, which is treated as γ_{mem}. Suppose that in addition to the main data (Y, S, Z, W), we also observe an instrument, T_{instru}. Carroll et al. (2004) considered this problem under certain conditions; see their eqs. (11)–(13). Their model for the instrument is a general varying-coefficient model of the form

$$T_{\text{instru}} = \alpha_0(S, Z) + \alpha_1(S, Z)X + \nu,$$

where ν has mean 0 and is independent of everything else. Using kernel techniques, they developed an estimate $\widehat{\gamma}_{\text{mem}}$, and in their theorems 3 and 4 they described an asymptotic expansion $n^{1/2}(\widehat{\gamma}_{\text{mem}} - \gamma_{0,\text{mem}}) = n^{-1/2}\sum_{i=1}^{n}\epsilon_{ui} + o_p(1)$, where the ϵ_{ui}'s satisfy $E(\epsilon_{ui}|Z_i) = 0$. They did a construction such that there are random variables $\widehat{\epsilon}_{ui}$ with the property that the sample moments of the $\widehat{\epsilon}_{ui}$ are consistent estimates of the sample moments of the ϵ_{ui}. Writing $\epsilon_{ui} = \epsilon_{ui}(\gamma_{0,\text{mem}})$, the asymptotic distribution of the estimate of $\mathcal{B}$ can be obtained as in Section 3.6 by appending the "estimating function" $\mathcal{L}_{i,\gamma_{\text{mem}}}(\gamma_{\text{mem}}) = \epsilon_{ui}(\gamma_{0,\text{mem}}) + \gamma_{0,\text{mem}} - \gamma_{\text{mem}}$.

In certain cases, a more explicit construction is possible. Consider once again the partially linear model of Section 3.3. Suppose that Z and X are scalar, and also that X and Z are independent and $T_{\text{instru}} = \alpha_0 + \alpha_1 X + \nu$. Then, assuming that all of the covariances are nonzero, we can characterize σ_u^2 as

$$\sigma_u^2 = \text{var}(W) - \frac{\text{cov}(W, T_{\text{instru}})\,\text{cov}(Y, W)}{\text{cov}(Y, T_{\text{instru}})}.$$

Write $\mathcal{Y} = (W, Y, T_{\text{instru}}, W^2, WT_{\text{instru}}, YW, YT_{\text{instru}})^{\mathrm{T}}$. If we consider the ensemble of moments and treat it as γ_{mem}, that is,

$$\begin{aligned}\gamma_{\text{mem}} = \big\{&E(W), E(Y), E(T_{\text{instru}}), E(W^2),\\ &E(WT_{\text{instru}}), E(YW), E(YT_{\text{instru}})\big\}^{\mathrm{T}},\end{aligned}$$

then the estimating equation for γ_{mem} is $\mathcal{L}_{\gamma_{\text{mem}}}(\mathcal{Y}, \gamma_{\text{mem}}) = \mathcal{Y} - \gamma_{\text{mem}}$. As in Section 3.6, absorbing γ_{mem} into $\mathcal{B}$ and denoting the original $\mathcal{B}$ as β, the estimating function becomes

$$\mathcal{L}_{\mathcal{B}}(\mathcal{Y}, \mathcal{B}, \theta) = \{\mathcal{L}_\beta^{\mathrm{T}}(\mathcal{Y}, \beta, \theta, \gamma_{\text{mem}}), \mathcal{L}_\gamma^{\mathrm{T}}(\mathcal{Y}, \gamma_{\text{mem}})\}^{\mathrm{T}},$$

where the dependence of $\mathcal{L}_\beta$ on γ_{mem} is through σ_u^2, which is an explicit function of γ_{mem}.

4. SIMULATION STUDY

We illustrate the method using a small simulation study. Consider a partially quadratic logistic relation in the central model and normal additive measurement model, that is,

$$\text{logit}\{\Pr(Y = 1|X, Z)\} = \beta_1 X + \beta_2 X^2 + \theta(Z),$$

where $W = X + U$ and $U = \text{normal}(0, \sigma_u^2)$, with σ_u^2 known. In our simulation we set $\sigma_u = .4$, $\mathcal{B} = (\beta_1, \beta_2)^{\mathrm{T}} = (.7, .7)^{\mathrm{T}}$, and $\theta(z) = .5\cos(z) - 1$. We set the sample size $n = 500$ and ran 1,000 simulations. We generated X from a normal distribution $\text{normal}(\mu_x = -1, \sigma_x^2 = 1)$ independent of Z, and generated Z from a uniform distribution $\text{uniform}(0, \pi)$. Using this setup, we can gain some insight into the effect of measurement error, as follows. Let $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2) = .86$ be the reliability ratio, then the usual regression calibration approximation replaces $\beta_1 X + \beta_2 X^2$ by $E(\beta_1 X + \beta_2 X^2|W) = \text{constant} + \lambda\{\beta_1 + 2\beta_2(1-\lambda)\mu_x\}W + \beta_2\lambda^2 W^2$. This suggests that the naive method that ignores measurement error will have a limiting value for β_1 of $\lambda\{\beta_1 + 2\beta_2(1-\lambda)\mu_x\} \approx .43$ and a limiting value for β_2 of $\beta_2\lambda^2 \approx .52$, which are fairly close to what we found in our simulations.

We implemented the proposed semiparametric estimator under two different posited models for $p(X|Z)$, the true normal model and a false model $p(X|Z) = \text{uniform}(-4, 2)$. The support of the uniform model is selected to practically agree with the support of the true $p(X)$, although the variance is quite different. For illustrative purposes, we selected the bandwidth as $h = \widehat{\sigma}_w n^{-1/3}$, where $\widehat{\sigma}_w$ is the estimated standard deviation of W. Under the true model, the estimator should yield an efficient estimator, whereas under the misspecified model, it should still give a consistent estimator. The standard error estimates for our method were computed as in Section 3.1, where the bandwidth involved in calculating $U(Z)$ is chosen as $h = .5\widehat{\sigma}_w n^{-1/5}$.

To compare other approaches with our method, we also implemented the naive estimator, a regression calibration estimator and a SIMEX estimator. The naive estimator simply fits the

partially quadratic logistic model ignoring measurement error and treating W as if it were X. The regression calibration estimators were adopted from Carroll et al. (2006, chap. 4), where we replaced X and X^2 by estimates of $E(X|W)$ and $E(X^2|W)$ and then run a standard partially quadratic logistic regression. To calculate $E(X|W)$ and $E(X^2|W)$, we used the following device: we computed $E(X^k|W)$ for $k = 1, 2$ under the correct normal model. For both the naive estimator and the regression calibration estimators, asymptotic standard errors were constructed as done by Severini and Staniswalis (1994).

We performed the SIMEX estimator following the description of Carroll et al. (2006, chap. 5), with the simulation step carried out at five inflated measurement error scales, spaced equally between σ_u and $2\sigma_u$, with 100 replicates for each error scale. We used the standard extrapolant functions: linear, quadratic, and rational linear. Note that SIMEX has two parts when the measurement error is additive with known variance as here. The first part is the actual estimation; in our case, studying some of the simulated datasets suggests that the linear extrapolant is to be preferred, and so we confine our attention to it. SIMEX also has a variance estimator; we used the linear extrapolant for it as well.

The simulation results are presented in the upper part of Table 1. The bias of the naive method is approximately in line with the regression calibration approximation. As expected from the theory, our methods provide relatively unbiased estimates whether computed with the correct model or with the incorrect model, and inference for $\mathcal{B}$ is also approximately correct in both cases, in the sense that the coverage probabilities were near the nominal 95%. As expected, the naive estimator is severely biased, and the regression calibration estimator and SIMEX estimator also have some bias. In a simulation not reported here with sample size $n = 1{,}000$, we observed that the bias, variance, and 95% confidence interval for our method improved significantly, whereas the naive, regression calibration, and SIMEX estimators all either deteriorated further or showed no sign of improvement.

The simulation was repeated with $\theta(Z) = .5\cos(2Z) - 1$, where the nonparametric function has higher frequency and its departure from linearity is more visible. Although the nonparametric component here is harder to estimate, the results show that the estimation of $\mathcal{B}$ is not very much influenced. The results are presented in the lower part of Table 1.

5. DATA ANALYSIS

We now analyze the Framingham data described in Section 1. The parameter ξ_{latent} in this problem contains the mean and variance of $\log(X - 50)$. We estimate the mean by the sample mean of the $W_i^* = \log(W_i - 50)$, whereas the variance estimate is the difference of the sample variance of the W_i^* and σ_u^2. In this scale the attenuation is .75, reflecting that roughly $1/4$ of the variability in measured log systolic blood pressure is due to measurement error. The actual blood pressure data do not closely follow a log-normal distribution (see Fig. 1, which gives a q–q plot of the W_i^*), so that, at least in principle, a likelihood analysis might suffer from model misspecification.

We applied our methods to the Framingham nonsmoker data under the models (1)–(2). We used various bandwidths, ranging from 25% to 100% of the standard deviation of age (Z), with no discernible effect on the estimate of β or its estimated variance for either the analysis that ignored measurement error or our analysis. Roughly, independent of the bandwidth, the analysis that ignored measurement error has a regression coefficient of .014 with an estimated standard error of .0040, and our analysis accounting for measurement error had an estimated regression coefficient of .019 with an estimated standard error of .0045. Given the attenuation in the log scale, these numbers are roughly what one would expect if the true X-data were log-normal, a happy coincidence.

The estimated function $\widehat{\theta}(\cdot)$ was, as expected, more sensitive to the bandwidth, because of the sparsity of responses $Y = 1$. For example, if the bandwidth is half of the standard deviation of age, then there were only four coronary heart disease cases in the kernel window for the lowest age, leading to instability in the fitted function. Because of this, for illustration purposes here we used a larger bandwidth equal to the standard deviation of Z. In this problem, age (Z) is only very weakly correlated with blood pressure (X); see Figure 1, where the sample correlation between Z and W is .22. This suggests that the fitting functions $\theta(z)$ in age will not change shape much, which is what happens (Fig. 2). The difference in the two functions

Table 1. Mean, Estimated Standard Error (SE), Empirical SE, and 95% Coverage of Four Classes of Estimators

		$\beta_1 (= .7)$				$\beta_2 (= .7)$			
Estimator	*p(X)*	*Mean*	*Estimated SE*	*Empirical SE*	*95%*	*Mean*	*Estimated SE*	*Empirical SE*	*95%*
$\theta(Z) = .5\cos(Z) - 1$									
Naive		.3900	.1516	.1720	.4450	.4780	.0733	.0873	.2250
Regression calibration	True	.6292	.2124	.2263	.9020	.6431	.1089	.1174	.8570
SIMEX		.5811	.2225	.2260	.8860	.6298	.1128	.1214	.8510
Semiparametric	True	.7205	.2873	.2772	.9470	.7264	.1721	.1557	.9390
	False	.7221	.2940	.2802	.9460	.7270	.1772	.1572	.9410
$\theta(Z) = .5\cos(2Z) - 1$									
Naive		.3958	.1514	.1705	.4640	.4797	.0734	.0866	.2180
Regression calibration	True	.6364	.2113	.2244	.9190	.6453	.1081	.1164	.8540
SIMEX		.5883	.2222	.2239	.8970	.6321	.1127	.1204	.8420
Semiparametric	True	.7268	.2844	.2759	.9510	.7284	.1699	.1554	.9400
	False	.7287	.2929	.2786	.9480	.7292	.1761	.1568	.9350

NOTE: The first class contains the naive estimator. The second class contains the regression calibration estimator when $p(X)$ is posited to be the true model, that is, normal$(-1, 1)$. The third class contains the SIMEX estimator with linear extrapolation function. The fourth class contains two semiparametric estimators, where $p(X)$ is posited to be the true model [i.e., normal$(-1, 1)$] and a very badly misspecified model [i.e., uniform$(-4, 2)$]. The true θ is $\theta(Z) = .5\cos(Z) - 1$ (upper panel) and $\theta(Z) = .5\cos(2Z) - 1$ (lower panel). We generated 1,000 simulated datasets, with sample size $n = 500$.

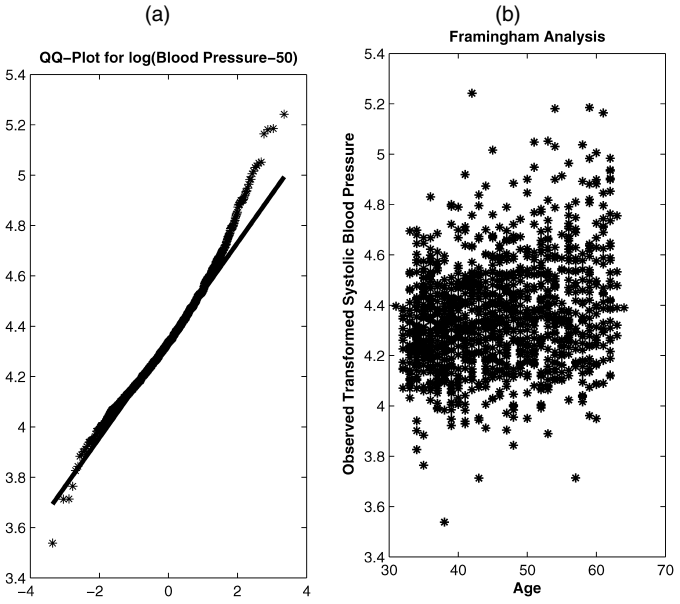

Figure 1. Framingham Heart Study Data. (a) q–q plot of the logarithm of systolic blood pressure. (b) Plot of W, observed systolic blood pressure, on Z, age, indicating a weak relationship between the two.

is easily explained. If we had changed the model to have logit $\beta_0+\beta_1 X+\theta(Z)$ with $\theta(Z)$ constrained to have mean 0, then the two functions almost overlap, so the difference seen in Figure 2 really reflects a difference in the intercept estimates.

6. DISCUSSION

We have constructed a locally efficient semiparametric estimator for a general class of semiparametric models with measurement errors, in which there are two infinite-dimensional nuisance functions, the distribution of the latent variable X and the nonparametric function $\theta(\cdot)$. The essential idea is to combine a parametric model estimator and a local kernel estimator through backfitting, rather than doing two projections as might be standard. The resulting backfitting-based estimator enjoys similar asymptotic properties as the corresponding parametric model estimator; that is, a consistent parametric model estimator will yield a consistent semiparametric model estimator, and an efficient parametric model estimator will yield an efficient semiparametric model estimator. When the estimator of Tsiatis and Ma (2004) is adopted as the parametric model estimator, the resulting semiparametric model estimator is guaranteed to be consistent, with its only efficiency loss caused by the possible misspecification of $p(X|S,Z)$.

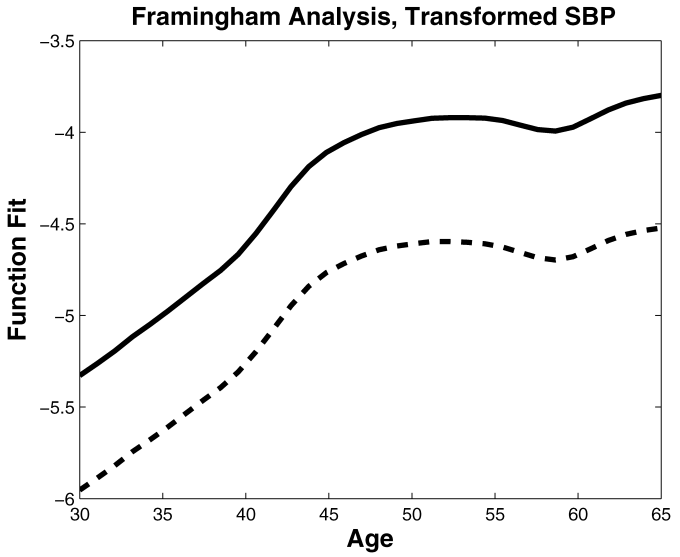

Figure 2. Framingham Data Function Fits. The solid line is the fit ignoring measurement error and the dashed line is our estimate.

To simplify the implementation, we have proposed using a local constant estimator in the local kernel estimating equation. The results can, of course, be generalized to local-polynomial kernel methods without any change in the asymptotic theory. Other nonparametric estimators (e.g., regression splines, Fourier series) could be chosen; the asymptotic theory should not change, but it would need to be rederived.

Instead of backfitting, we could also use the efficient score to form an estimating equation, that is, replace $\mathcal{L}_{\mathcal{B}}(\bullet)$ by $\mathcal{L}_{\mathcal{B}}(\bullet)-\Psi_\theta(\bullet)U(Z)$. The advantage of this estimating equation is that we do not need undersmoothing of the kernel estimator—standard results suggest that the asymptotic theory would not change. However, as a practical matter, this approach is far more computationally complex than backfitting, because of the need for calculating a consistent version of $U(Z)$ at each stage of the process.

Although we constructed our method for a one-dimensional Z, the method can be further extended to higher-dimensional cases. One difficulty with such extension is the familiar curse of dimensionality. We believe that our basic approach can be used within the single-index model context; this will be taken up in a later article.

APPENDIX: TECHNICAL DETAILS

Define

$$\begin{aligned}&\mathcal{S}^*_{\text{eff}}(\mathcal{Y},\mathcal{B}_0,\gamma_{0,\text{mem}},\theta_0,\xi_{\text{latent}})\\&\quad=\{\mathcal{L}_{\mathcal{B}}(\mathcal{Y},\mathcal{B}_0,\gamma_{0,\text{mem}},\theta_0,\xi_{\text{latent}})^{\text{T}},\\&\qquad\Psi_\theta(\mathcal{Y},\mathcal{B}_0,\gamma_{0,\text{mem}},\alpha_0,\xi_{*,\text{latent}})^{\text{T}}\}^{\text{T}}.\end{aligned}$$

We need that

$$\begin{aligned}0&=E\left\{\frac{\partial\mathcal{L}_{\mathcal{B}}(\mathcal{Y},\mathcal{B}_0,\gamma_{0,\text{mem}},\alpha_0,\xi_{*,\text{latent}})}{\partial\xi_{\text{latent}}}\Big|S,Z\right\}\\&=E\left\{\frac{\partial\Psi_\theta(\mathcal{Y},\mathcal{B}_0,\gamma_{0,\text{mem}},\alpha_0,\xi_{*,\text{latent}})}{\partial\xi_{\text{latent}}}\Big|S,Z\right\}.\end{aligned}\tag{A.1}$$

To show (A.1), note that $\mathcal{S}^*_{\text{eff}}$ is a function in the nuisance tangent space orthogonal complement defined by Tsiatis and Ma (2004). Their results imply that $E\{\mathcal{S}^*_{\text{eff}}|X,S,Z\}=0$. Taking derivatives with respect to $\xi_{*,\text{latent}}$ and interchanging expectation and differentiation, we obtain

$$E\left\{\frac{\partial\mathcal{S}^*_{\text{eff}}}{\partial\xi^{\text{T}}_{*,\text{latent}}}\Big|X,S,Z\right\}+E\left[\mathcal{S}^*_{\text{eff}}\frac{\partial\log\{p(\mathcal{Y}|X,S,Z)\}}{\partial\xi^{\text{T}}_{\text{latent}}}\Big|X,S,Z\right]=0.\tag{A.2}$$

Because the conditional probability distribution function (pdf) $p(\mathcal{Y}|X,S,Z)$ is free of ξ_{latent}, the second term in (A.2) is 0, we have that $E(\partial\mathcal{S}^*_{\text{eff}}/\partial\xi^{\text{T}}_{*,\text{latent}}|S,Z)=E\{E(\partial\mathcal{S}^*_{\text{eff}}/\partial\xi^{\text{T}}_{*,\text{latent}}|X,S,Z)|S,Z\}=0$, as claimed.

Sketch of the Proof of Theorem 1

We provide only a sketch of the proof here. Precise conditions that justify our calculations and the general backfitting algorithm have been given by Claeskens and van Keilegom (2003) and Chen et al. (2003).

We assume the primitive conditions that Z has compact support and that its density function is positive on that support. We also assume that $\widehat{\theta}(z, \mathcal{B}, \gamma_{\text{mem}}, \xi_{\text{latent}})$ has the usual properties uniformly in z in neighborhoods of $(\mathcal{B}_0, \gamma_{0,\text{mem}}, \xi_{*,\text{latent}})$, and particularly that $\widehat{\theta}(z, \mathcal{B}_0, \gamma_{0,\text{mem}}, \xi_{*,\text{latent}}) = \theta_0(z) + o_p(n^{-1/4})$ uniformly in z; this follows because $nh^4 \to 0$.

Using standard expansions and applying (A.1), it is easy to see that

$$\begin{aligned} &-\mathcal{F} n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) \\ &\quad = n^{-1/2} \sum_{i=1}^{n} \bigl[\mathcal{L}_{i\mathcal{B}}(\bullet) + \mathcal{L}_{i\mathcal{B}\theta}(\bullet)\{\widehat{\theta}(Z_i, \mathcal{B}_0) - \theta_0(Z_i)\}\bigr] \\ &\qquad + o_p(1). \end{aligned} \tag{A.3}$$

Note that $\mathcal{F}$ is not necessarily symmetric. Using standard results, we have the usual local estimating equation expansion

$$\begin{aligned} &\widehat{\theta}(z, \mathcal{B}_0, \gamma_{0,\text{mem}}, \xi_{*,\text{latent}}) - \theta_0(z) \\ &\quad = (h^2/2)\theta_0^{(2)}(z) - n^{-1} \sum_{i=1}^{n} K_h(Z_i - z)\Psi_{i\theta}(\bullet)/\{f_Z(z)\Omega(z)\} \\ &\qquad + o_p\bigl(n^{-1/2}\bigr), \end{aligned} \tag{A.4}$$

where $\Psi_{i\theta}(\bullet) = \mathcal{L}_\theta\{\mathcal{Y}_i, \mathcal{B}_0, \gamma_{0,\text{mem}}, \theta_0(Z_i), \xi_{*,\text{latent}}\}$. Consider the two terms on the right side of (A.4) and substitute each of them into (A.3). If we assume that $nh^4 \to 0$, then the first term disappears. The second term is easily seen to be $-n^{-1/2} \sum_{i=1}^{n} \Psi_{i\theta}(\bullet)\mathcal{U}(Z_i) + o_p(1)$, from which the result follows immediately.

Preliminaries Necessary for Theorem 2

To prove Theorem 2, we first need to establish some relevant geometric structures. Following Bickel, Klaassen, Ritov, and Wellner (1993), we view the space of all the mean-0 functions $f(\mathcal{Y})$ as a Hilbert space $\mathcal{H}$, with the inner product defined as the covariance between two functions. Let $\eta_0(X, S, Z)$ represent the true pdf, $p(X, S, Z)$. Here all the expectations are calculated under the true distribution $p_0\{\mathcal{Y}, \mathcal{B}_0, \gamma_{0,\text{mem}}, \theta_0(Z, \mathcal{B}_0, \gamma_{0,\text{mem}}), \eta_0(X, S, Z)\}$. We work within the framework where θ and η are unknown infinite-dimensional nuisance parameters. The model $p(\mathcal{Y}, \mathcal{B}, \gamma_{\text{mem}}, \theta, \eta)$ is semiparametric in the sense that it contains the infinite-dimensional nuisance parameters θ and η. A parametric submodel is defined as a parametric model that is contained in this semiparametric model and that contains the truth. The nuisance tangent space Λ is defined as the mean squared closure of the linear combination of the nuisance tangent spaces of all of the parametric submodels. The nuisance tangent space of a parametric model is defined as the mean squared closure of the linear combination of the derivative of the logarithm of the pdf with respect to the nuisance parameter. The orthogonal complement of Λ in the Hilbert space $\mathcal{H}$ is denoted as $\Lambda^\perp$. We also use the notion of a nuisance tangent space with respect to θ, denoted as Λ_θ. This means that concentrate on the nuisance parameter θ, treat other nuisance parameters as if they were known, and calculate the nuisance tangent space; its orthogonal complement in $\mathcal{H}$ is denoted as $\Lambda_\theta^\perp$. Similar calculations will apply to Λ_η and $\Lambda_\eta^\perp$. Using the definition of Λ, it is clear that $\Lambda = \Lambda_\eta + \Lambda_\theta$ and $\Lambda^\perp = \Lambda_\eta^\perp \cap \Lambda_\theta^\perp$. Using results concerning Λ_η and $\Lambda_\eta^\perp$ given by Tsiatis and Ma (2004), we know that Λ_η can be further decomposed into $\Lambda_{\eta 1} \oplus \Lambda_{\eta 2}$, where $\Lambda_{\eta 1} = [E\{h(X, S, Z)|\mathcal{Y}\} : E\{h(X, S, Z)|S, Z\} = 0]$, $\Lambda_{\eta 2} = [h(S, Z) : E\{h(S, Z)\} = 0]$, and $\Lambda_\eta^\perp = [h(\mathcal{Y}) : E\{h(\mathcal{Y})|X, S, Z\} = 0$ almost everywhere]. Because there are no constraints on $\theta(Z)$, it can be easily verified that $\Lambda_\theta = \mathcal{S}_\theta(\mathcal{Y})g(Z)$, where $\mathcal{S}_\theta(\mathcal{Y})$ is the score vector obtained by taking partial derivatives of $\log p(\mathcal{Y}, \mathcal{B}, \gamma_{\text{mem}}, \theta, \eta)$ with respect to θ and $g(Z)$ is an arbitrary function of Z. Therefore, $\Lambda_\theta^\perp = [h(\mathcal{Y}) : E\{h(\mathcal{Y})\mathcal{S}_\theta(\mathcal{Y})|Z\} = 0]$.

Once we derive Λ and $\Lambda^\perp$, we can derive the influence function and the asymptotic properties of the estimator by inspecting $\psi_\mathcal{B} = \Pi(\mathcal{L}_\mathcal{B}|\Lambda^\perp)$, the projection of $\mathcal{L}_\mathcal{B}$ onto $\Lambda^\perp$.

Proof of Theorem 2

Local efficiency implies two properties: (a) If the model that we assume for $p(X|S, Z)$ is correct, then the estimator is efficient; and (b) if the model for $p(X|S, Z)$ is incorrect, then the estimator is consistent. We focus on property (a) first. Decompose $\mathcal{L}_\mathcal{B}$ into $\mathcal{L}_\mathcal{B} = \psi_\mathcal{B} + \phi_r$, where $\psi_\mathcal{B}$ is the projection of $\mathcal{L}_\mathcal{B}$ onto $\Lambda^\perp$ [i.e., $\psi_\mathcal{B} = \Pi(\mathcal{L}_\mathcal{B}|\Lambda^\perp)$] and ϕ_r is the residual of the projection (i.e., $\phi_r \in \Lambda$). The right side of (11) can then be written as

$$n^{-1/2} \sum_{i=1}^{n} \{\psi_{i\mathcal{B}}(\bullet) + \phi_{ir}(\bullet) - \Psi_{i,\theta}(\bullet)\mathcal{U}(Z_i)\} + o_p(1).$$

Because Ψ_θ is constructed through projecting $\mathcal{S}_\theta$, the score vector with respect to θ, onto $\Lambda_\eta^\perp$, it follows that $\Psi_\theta(\bullet) = \mathcal{S}_\theta - \Pi(\mathcal{S}_\theta|\Lambda_\eta)$. Using the construction of Λ_θ, it follows that $\mathcal{S}_\theta\mathcal{U}(Z) \in \Lambda_\theta \subset \Lambda$. Write $\Pi(\mathcal{S}_\theta|\Lambda_\eta)$ as $\Pi(\mathcal{S}_\theta|\Lambda_{\eta 1}) \oplus \Pi(\mathcal{S}_\theta|\Lambda_{\eta 2})$.

Due to the description of $\Lambda_{\eta 2}$, for an arbitrary mean-0 function $f(\mathcal{Y})$, we have $\Pi(f|\Lambda_{\eta 2}) = E(f|S, Z)$; in particular, for $\mathcal{S}_\theta$, we have $\Pi(\mathcal{S}_\theta|\Lambda_{\eta 2}) = E(\mathcal{S}_\theta|S, Z)$. Moreover, $E\{E(\mathcal{S}_\theta|S, Z)\mathcal{U}(Z)\} = E\{E(\mathcal{S}_\theta|Z)\mathcal{U}(Z)\} = 0$, where the last equality holds because $E(\mathcal{S}_\theta|Z) = E\{\partial \log p(\mathcal{Y}|Z)/\partial\theta|Z\} = 0$. Thus we have that $\Pi(\mathcal{S}_\theta|\Lambda_{\eta 2})\mathcal{U}(Z) \in \Lambda_{\eta 2}$. Obviously, $\Pi(\mathcal{S}_\theta|\Lambda_{\eta 1})\mathcal{U}(Z) \in \Lambda_{\eta 1}$, so $\Psi_\theta(\bullet)\mathcal{U}(Z) \in \Lambda$. Thus we can rewrite (11) as

$$n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) = -n^{-1/2} \sum_{i=1}^{n} \{\mathcal{F}^{-1}\psi_{i\mathcal{B}} + \mathcal{F}^{-1}r_i\} + o_p(1),$$

where $\mathcal{F}^{-1}r \in \Lambda$. For any regular asymptotic linear semiparametric estimator $\widehat{\mathcal{B}}$, we have that $n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) = n^{-1/2} \sum_{i=1}^{n} \widetilde{\psi}_{i\mathcal{B}} + o_p(1)$ for some influence function $\widetilde{\psi}_\mathcal{B} \in \Lambda^\perp$ (Newey 1990, after thm. 2.1). Hence we have that

$$n^{-1/2} \sum_{i=1}^{n} (\widetilde{\psi}_{i\mathcal{B}} + \mathcal{F}^{-1}\psi_{i\mathcal{B}} + \mathcal{F}^{-1}r_i) = o_p(1). \tag{A.5}$$

Note that each individual term in (A.5) has mean 0 because they are functions in $\mathcal{H}$; hence the left side of (A.5) has the same order as the standard deviation of $\widetilde{\psi}_\mathcal{B} + \mathcal{F}^{-1}\psi_\mathcal{B} + \mathcal{F}^{-1}r$. Because $\widetilde{\psi}_\mathcal{B} + \mathcal{F}^{-1}\psi_\mathcal{B} \in \Lambda^\perp$ (the sum of two elements of $\Lambda^\perp$ is still in $\Lambda^\perp$) and $\mathcal{F}^{-1}r \in \Lambda$, we have $\widetilde{\psi}_\mathcal{B} + \mathcal{F}^{-1}\psi_\mathcal{B} \perp \mathcal{F}^{-1}r$. Note that the variance is the square of the length of a function in the Hilbert space $\mathcal{H}$; hence the variance of $\widetilde{\psi}_\mathcal{B} + \mathcal{F}^{-1}\psi_\mathcal{B} + \mathcal{F}^{-1}r$ is the sum of the variance of $\widetilde{\psi}_\mathcal{B} + \mathcal{F}^{-1}\psi_\mathcal{B}$ and the variance of $\mathcal{F}^{-1}r$. Obviously, both variances need to be $o_p(1)$ to satisfy (A.5), which implies that $\widetilde{\psi}_\mathcal{B} + \mathcal{F}^{-1}\psi_\mathcal{B} = o_p(1)$ almost everywhere and $\mathcal{F}^{-1}r = o_p(1)$ almost everywhere. This shows that $n^{-1/2} \sum_{i=1}^{n} \mathcal{F}^{-1}r_i = o_p(1)$. Hence (11) reduces to

$$n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) = -n^{-1/2} \sum_{i=1}^{n} \mathcal{F}^{-1}\psi_{i\mathcal{B}} + o_p(1). \tag{A.6}$$

Note that $\psi_\mathcal{B}$ is constructed by projecting the score $S_\mathcal{B}$ onto $\Lambda^\perp$, so this in fact guarantees that $\psi_\mathcal{B}$ is the efficient score, that is, $\psi_\mathcal{B} = \Pi\{\Pi(\mathcal{S}_\mathcal{B}|\Lambda_\eta^\perp)|\Lambda^\perp\} = \Pi(\mathcal{S}_\mathcal{B}|\Lambda^\perp)$.

We have thus demonstrated that the backfitting estimator has the same influence function as the one corresponding to the efficient score; that is, the backfitting estimator is exactly the efficient estimator. Hence property (a) is shown.

When the model for $p(X|S,Z)$ is incorrect, from Theorem 1, we still have $E\{\mathcal{L}_{\mathcal{B}}(\bullet) - \Psi_\theta(\bullet)\mathcal{U}(Z)\} = 0$. Hence the estimator remains consistent, and property (b) is shown.

Proof of Corollary 1

Comparing (A.6) and (11), we know that $\mathcal{F}^{-1}\psi_{\mathcal{B}}(\bullet) = \mathcal{F}^{-1} \times \mathcal{L}_{\mathcal{B}}(\bullet) - \mathcal{F}^{-1}\Psi_\theta(\bullet)\mathcal{U}(Z)$. Under the correct model for $p(X|S,Z)$, $-\mathcal{F}^{-1}\psi_{\mathcal{B}}(\bullet)$ is the efficient influence function, so $E(-\mathcal{F}^{-1}\psi_{\mathcal{B}} \times S_{\mathcal{B}}^{\mathrm{T}}) = I$. Because $\psi_{\mathcal{B}} = \Pi(S_{\mathcal{B}}|\Lambda^{\perp})$, where $S_{\mathcal{B}}$ is the score function with respect to $\mathcal{B}$, we obtain $E(-\mathcal{F}^{-1}\psi_{\mathcal{B}}\psi_{\mathcal{B}}^{\mathrm{T}}) = I$. Therefore, we have $\mathcal{F} = -\mathrm{cov}\{\psi_{\mathcal{B}}(\bullet)\} = -\mathrm{cov}\{\mathcal{L}_{\mathcal{B}}(\bullet) - \Psi_\theta(\bullet)\mathcal{U}(Z)\} = -\Sigma$. Thus, under a correctly specified model, $p(X|S,Z)$, $\mathcal{F} = -\Sigma$ is a symmetric matrix, and the variance $\mathcal{F}^{-1}\Sigma\mathcal{F}^{-\mathrm{T}}$ of the estimator $\widehat{\mathcal{B}}$ simplifies to Σ^{-1}.

Expressions for *a* and *b* in the Partially Linear Model Example

Calculating the conditional density of $p(X|\delta)$ under normality, we obtain

$$a = (\Omega^{-1} + A^{\mathrm{T}}B^{-1}A)^{-1}(\Omega^{-1}\mu - A^{\mathrm{T}}B^{-1}\alpha)$$

and

$$b = (\Omega^{-1} + A^{\mathrm{T}}B^{-1}A)^{-1}A^{\mathrm{T}}B^{-1},$$

where $\alpha = \theta\xi_{\mathrm{latent}}\beta/\sigma_\epsilon^2$, $A = I + \xi_{\mathrm{latent}}\beta\beta^{\mathrm{T}}/\sigma_\epsilon^2$, and $B = \xi_{\mathrm{latent}} + \xi_{\mathrm{latent}}\beta\beta^{\mathrm{T}}\xi_{\mathrm{latent}}/\sigma_\epsilon^2$.

Proof of Equivalence in the Partially Linear Model Example

We calculate $U(Z)$ first. For simplicity, denote $c = 1/(1 + \beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta/\sigma_\epsilon^2)$. We have $\Psi_{\theta\theta} = -c$. Because the first component of $\mathcal{L}_\beta$ is $A_1 = \Psi_\theta(a + b\delta)$, we have $A_{1\theta} = -c(a + b\delta) + \Psi_\theta a_\theta$, where we use the fact that b and δ do not depend on θ. The partial derivative of the second component A_2 with respect to θ is $A_{2\theta} = -2\Psi_\theta$. Thus $E(A_{1\theta}|Z) = -c\{a + bE(\delta|Z)\}$ and $E(A_{2\theta}|Z) = 0$. Hence $U(Z) = [\{a + bE(\delta|Z)\}^{\mathrm{T}}, 0]^{\mathrm{T}}$.

We now inspect $\mathcal{L}_\beta - \Psi_\theta U(Z)$. It consists of a first component, $\widetilde{A}_1 = A_1 - \Psi_\theta\{a + bE(\delta|Z)\} = \Psi_\theta b\{\delta - E(\delta|Z)\}$, and a second component, A_2. The corresponding estimator is

$$\sum_{i=1}^{n}\{Y_i - W_i^{\mathrm{T}}\beta - \theta(Z_i)\}\{\delta_i - E(\delta|Z_i)\} = 0, \qquad \text{(A.7)}$$

$$\sum_{i=1}^{n}\{Y_i - W_i^{\mathrm{T}}\beta - \theta(Z_i)\}^2 - n\sigma_\epsilon^2 - n\beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta = 0, \qquad \text{(A.8)}$$

and

$$\sum_{i=1}^{n}K_h(Z_i - z)\{Y_i - W_i^{\mathrm{T}}\beta - \theta(z)\} = 0. \qquad \text{(A.9)}$$

Denote $\widetilde{Y}_i = Y_i - E(Y_i|Z_i)$ and $\widetilde{W}_i = W_i - E(W_i|Z_i)$; then $Y_i - W_i^{\mathrm{T}}\beta - \theta(Z_i) = \widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta$. We next show that the set of equations (A.7), (A.8), and (A.9) is equivalent to the set of equations (A.10), (A.8), and (A.9), with

$$\sum_{i=1}^{n}\widetilde{W}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta) + n\xi_{\mathrm{latent}}\beta = 0. \qquad \text{(A.10)}$$

Because $\delta = W + Y\xi_{\mathrm{latent}}\beta/\sigma_\epsilon^2$, (A.7) is equivalent to

$$0 = \sum_{i=1}^{n}\sigma_\epsilon^2\widetilde{W}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta) + \sum_{i=1}^{n}\widetilde{Y}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)\xi_{\mathrm{latent}}\beta. \qquad \text{(A.11)}$$

Multiplying β^{T} from the left side on (A.11), in conjunction with (A.8), we obtain

$$\begin{aligned}0 &= \sum_{i=1}^{n}\widetilde{W}_i^{\mathrm{T}}\beta(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)\sigma_\epsilon^2 + \sum_{i=1}^{n}\widetilde{Y}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)\beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta\\ &= -\sum_{i=1}^{n}(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)^2\sigma_\epsilon^2 + \sum_{i=1}^{n}\widetilde{Y}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)(\sigma_\epsilon^2 + \beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta)\\ &= -n(\sigma_\epsilon^2 + \beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta)\sigma_\epsilon^2 + \sum_{i=1}^{n}\widetilde{Y}_i(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)(\sigma_\epsilon^2 + \beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta).\end{aligned}$$

Because $\sigma_\epsilon^2 + \beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta$ is positive, we obtain

$$\sum_{i=1}^{n}(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)\widetilde{Y}_i = n\sigma_\epsilon^2. \qquad \text{(A.12)}$$

Similarly, multiplying β^{T} from the left side of (A.10), in conjunction with (A.8), we obtain

$$\begin{aligned}0 &= \sum_{i=1}^{n}(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)\widetilde{W}_i^{\mathrm{T}}\beta + n\beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta\\ &= -\sum_{i=1}^{n}(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)^2 + \sum_{i=1}^{n}(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)\widetilde{Y}_i + n\beta^{\mathrm{T}}\xi_{\mathrm{latent}}\beta\\ &= \sum_{i=1}^{n}(\widetilde{Y}_i - \widetilde{W}_i^{\mathrm{T}}\beta)\widetilde{Y}_i - n\sigma_\epsilon^2,\end{aligned}$$

which also leads to (A.12). As long as (A.12) holds, (A.11) and (A.10) are equivalent, and hence the two sets of equations are equivalent.

Finally, we can easily verify that (A.10), (A.8), and (A.9) leads to the estimator given in (5), (9), and (8) of Liang et al. (1999).

[Received June 2005. Revised January 2006.]

REFERENCES

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Carroll, R. J., Ruppert, D., Crainiceanu, C., and Stefanski, L. A. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.), London: CRC Press.

Carroll, R. J., Ruppert, D., Tosteson, T., Crainiceanu, C., and Karagas, M. (2004), "Nonlinear and Nonparametric Regression and Instrumental Variables," *Journal of the American Statistical Association*, 99, 736–750.

Carroll, R. J., Ruppert, D., and Welsh, A. (1998), "Local Estimating Equations," *Journal of the American Statistical Association*, 93, 214–227.

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 van Keilegom, I. (2003), "Bootstrap Confidence Bands for Regression Functions and Their Derivatives," *The Annals of Statistics*, 31, 1852–1884.

Cook, J., and Stefanski, L. A. (1995), "A Simulation Extrapolation Method for Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Gustafson, P. (2004), *Measurement Error and Misclassification in Statistics and Epidemiology*, New York: Chapman & Hall/CRC.

Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B., and Kjelsberg, M. O. (1986), "Overall and Coronary Heart Disease Mortality Rates in Relation to Major Risk Factors in 325,348 Men Screened for MRFIT," *American Heart Journal*, 112, 825–836.

Kress, R. (1999), *Liner Integral Equations*, New York: Springer-Verlag.

Liang, H., Härdle, W., and Carroll, R. J. (1999), "Estimation in a Semiparametric Partially Linear Errors-in-Variables Model," *The Annals of Statistics*, 27, 1519–1535.

Liang, H., Wang, S., Robins, J. M., and Carroll, R. J. (2004), "Estimation in Partially Linear Models With Missing Covariates," *Journal of the American Statistical Association*, 99, 357–367.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Stefanski, L. A., and Carroll, R. J. (1987), "Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models." *Biometrika*, 74, 703–716.

Tsiatis, A. A., and Ma, Y. (2004), "Locally Efficient Semiparametric Estimators for Functional Measurement Error Models," *Biometrika*, 91, 835–848.