Simultaneous treatment of unspecified heteroskedastic model error distribution and mismeasured covariates for restricted moment models

Tanya P. Garcia[1]
Department of Epidemiology and Biostatistics, Texas A&M University

AND

Yanyuan Ma
Department of Statistics, Pennsylvania State University

Abstract

We develop consistent and efficient estimation of parameters in general regression models with mismeasured covariates. We assume the model error and covariate distributions are unspecified, and the measurement error distribution is a general parametric distribution with unknown variance-covariance. We construct root-n consistent, asymptotically normal and locally efficient estimators using the semiparametric efficient score. We do not estimate any unknown distribution or model error heteroskedasticity. Instead, we form the estimator under possibly incorrect working distribution models for the model error, error-prone covariate, or both. Empirical results demonstrate robustness to different incorrect working models in homoscedastic and heteroskedastic models with error-prone covariates.

Some Key Words: Influence function; Linear operator; Measurement error; Nuisance tangent space; Restricted moment model

Short title: Regression with Error

JEL Classification: C1

[1]Tanya P. Garcia is Corresponding Author, Department of Epidemiology and Biostatistics, Texas A&M University Health Science Center, TAMU 1266, College Station, TX 77843-1266. Phone: (979) 436-9396, Fax: (979) 436-9595; Email: tpgarcia@sph.tamhsc.edu.

1 Introduction

1.1 Motivating problem

Regression is arguably the most familiar topic in econometrics and statistics and has motivated a vast amount of literature. Many scientific phenomena can be modeled using a general regression model where a univariate response Y is related to covariates $X \in \mathbb{R}^k$ and $Z \in \mathbb{R}^s$ through

$$Y = m(X, Z; \beta) + \epsilon. \tag{1}$$

Here, m is known up to the parameter $\beta \in \mathbb{R}^p$, and the model error ϵ is only required to satisfy $E(\epsilon|X, Z) = 0$. With the conditional distribution of ϵ unspecified, this model is also known as a restricted moment model (RMM). A typical challenge with RMMs is that some covariates, say Z, are precisely measured, whereas others, say X, are mismeasured. In place of X_i, $i = 1, \ldots, n$, one instead observes ℓ surrogate replicates

$$W_{ij} = X_i + U_{ij}, \quad j = 1, \ldots, \ell, \tag{2}$$

where U_{ij}'s are independent, mean zero random variables with unknown variance-covariance $\Omega_U \in \mathbb{R}^{k \times k}$. The surrogacy assumption implies that Y_i and W_{ij}'s are conditionally independent given (X_i, Z_i). Lastly, we suppose the measurement error is classical so that X_i and U_{ij} are independent.

An example of this model is in the nutrition study of Flagg et al. (2000). There, a key interest is properly modeling the relationship between percent calories from fat (Y), race (Z), and saturated fat intake (X). Saturated fat intake is not known exactly and only an approximate version via two repeated measurements, $W_{\cdot 1}, W_{\cdot 2}$, is available from food frequency questionnaires. To handle the measurement error in this example and in any model characterized by (1) and (2), the goal of this paper is to estimate the model parameters β and Ω_U under the following general assumptions:

Assumption (i): the mean model $m(X, Z; \beta)$ is any linear or nonlinear function;

Assumption (ii): the model error ϵ may depend on (X, Z) (i.e., heteroskedasticity), and its conditional distribution $p_{\epsilon|X,Z}(\epsilon|x, z)$ is unspecified;

Assumption (iii): the conditional distribution of X given Z, $p_{X|Z}(x|z)$, and the distribution of Z, $p_Z(z)$, exist but are completely unspecified. Thus we have a modern functional measurement error model (Carroll et al., 2006, chap. 7.2);

Assumption (iv): the measurement error is classical and U_{ij}, $i = 1, \ldots, n; j = 1, \ldots, \ell$, has a general parametric distribution $p_{U_{ij}}(u; \Omega_U)$ with Ω_U unknown. This contrasts from the usual normality assumption for measurement error (Carroll et al., 1999, 2004).

1.2 Estimation challenges

Allowing $p_{\epsilon|X,Z}$, $p_{X|Z}$, and p_Z to be unspecified provides more modeling flexibility and reduces the chance of model misspecification. However, it also raises serious challenges. The unknown distributions cannot be ignored, and arbitrarily adopting models for $p_{\epsilon|X,Z}$ or $p_{X|Z}$ may cause bias. Estimating these distributions is also potentially difficult. For example, $p_{X|Z}$ is a model of unobserved variables. Its estimation would involve an inverse operation such as deconvolution (Stefanski and Carroll, 1990), which results in a very slow rate (Carroll and Hall, 1988; Fan, 1991). The estimation of $p_{\epsilon|X,Z}(\epsilon|x, z)$ is equally challenging because residuals are unobtainable in measurement error models even if model parameters were known. The unavailability of the residuals makes correctly estimating the model error's variance-covariance difficult. This is especially problematic when the model error is heteroskedastic, and a proper variance-covariance is needed to yield consistent model parameter estimates. Although methods exist to estimate the unknown variance-covariance, they are either approximate (Carroll and Wang, 2008) or complex (Delaigle and Hall, 2011).

1.3 Competing methods and features of our approach

A nonlinear, classical measurement error model with replicates has been treated by several authors. Extensive research has focused on measurement error problems with specific forms of $m(X, Z; \beta)$, ranging from polynomial regression (Chan and Mak, 1985; Cheng and

Schneeweiss, 1998; Cheng et al., 2000; Huang and Huwang, 2001) to generalized linear mixed models (Liang, 2009; Li and Liqun, 2012). For our purposes, we consider $m(X, Z; \beta)$ to be any linear or nonlinear form, which is a general assumption of many existing works. For example, Li (2002) used Kotlarski's identification (Rao, 1992, p. 21) to identify and consistently estimate model parameters for a general $m(X, Z; \beta)$ with two replicates W_{i1}, W_{i2}. Tsiatis and Ma (2004) developed a consistent, asymptotically normal estimator when $p_{\epsilon|X,Z}$ is known and parametric. Schennach (2004a) used properties of Fourier transforms, where the crux of her work lies in constructing moments of the unobserved X, and then forming estimators that can be written in terms of these moments. Schennach (2004b) developed an unbiased, Nadaraya-Watson based estimator to nonparametrically estimate $m(X, Z)$ in (1). Lastly, Hu and Schennach (2008) and Schennach and Hu (2013) used a sieve maximum likelihood estimator (MLE) which yields consistency and the former successfully handles heteroskedastic measurement error (i.e., U in (2) depends on X). For an overview on measurement error models, see Fuller (1987) for earlier results in linear models and Carroll et al. (2006) for modern approaches in linear and nonlinear models. The developed methodologies have all positively impacted the literature of regression with classical measurement error. Still, some limitations linger and it is these limitations that motivated this work.

In this paper, we propose to overcome two key limitations of existing methods: the direct estimation or knowledge of $p_{\epsilon|X,Z}$, and the inability to handle model error heteroskedasticity. In this regard, we develop a semiparametric estimator which avoids estimating $p_{X|Z}$ and $p_{\epsilon|X,Z}$. This is possible through deriving the semiparametric efficient score (Bickel et al., 1993; Tsiatis, 2006) which we reveal is robust to misspecification of the unknown distributions. Our approach involves adopting working parametric models for the unknown distributions. We show that if the working models are correct, then the estimator is semiparametric efficient; otherwise, the estimator is still root-n consistent and asymptotically normal. Lastly, our method does not require correctly estimating the model error's variance-covariance.

Not having to directly estimate $p_{\epsilon|X,Z}$ differs from the semiparametric Tsiatis and Ma (2004) method and the sieve MLE (Shen, 1997; Schennach and Hu, 2013). Tsiatis and Ma (2004) assume $p_{\epsilon|X,Z}$ is a known, parametric form. Unfortunately, in our own numerical studies (Section 4), we found that such an assumption is sensitive to misspecification of the

model error variance. With the sieve MLE, $p_{\epsilon|X,Z}$ and $p_{X|Z}$ are represented by increasingly rich parametric representations such as a truncated series of basis functions. The parameters in the truncated series and regression are then jointly estimated via MLE subject to constraints that ensure the estimated $p_{\epsilon|X,Z}$, $p_{X|Z}$ are valid densities and that $E(\epsilon|X,Z) = 0$. Sieve methods yield consistent estimators and are fairly straight-forward to implement, making the approach widely appreciated in the literature. However, compared to the sieve MLE, our approach bypasses the consistent estimation of $p_{\epsilon|X,Z}$, $p_{X|Z}$. In doing so, our method eliminates a step in the aim of constructing a consistent estimator and, as described next, flexibly handles potential heteroskedasticity in the model error.

Model error heteroskedasticity is a challenging problem, especially in a measurement error setting where residuals are unavailable to aid the appropriate modeling of variance-covariance structures. In bypassing the correct estimation of $p_{\epsilon|X,Z}$, our method implicitly handles misspecifications of the model error's variance structure. That is, knowledge of the model error being heteroskedastic or homoskedastic is not needed. In our own explorations of existing estimators to handle model error heteroskedasticity, we found some shortcomings.

The estimators of Li (2002) and Schennach and Hu (2013) both assume ϵ and (X, Z) are independent (i.e., homoskedastic model error). Consequently, ignoring the homoskedastic assumption naturally results in bias when the model error is truly heteroskedastic; see numerical studies in Section 4 for bias of the sieve estimator from Schennach and Hu (2013). The bias persists even when the number of terms in the sieve representations increases. As improvement, Hu and Schennach (2008) developed a different sieve estimator that successfully handles heteroskedastic measurement error (i.e., U in (2) depends on X). Unfortunately, when we extended their methodology to handle heteroskedastic model error, we encountered two difficulties. First, for the heteroskedastic sieve of $p_{\epsilon|X,Z}$ to be a valid density and have conditional mean zero, we require imposing twelve constraints (see Section S.8, Supplementary Material). Second, from our numerical studies (Section 4), we found that the heteroskedastic sieve estimator yielded biased estimates for the RMM models considered here. Given that the sieve approach has been widely successful in various regressions with errors-in-covariates, we were initially surprised by these results. However, we now believe the biasedness is a consequence of the complex computation that attempts a constrained

optimization subject to too many constraints.

Lastly, for a nonparametric regression with classical measurement error, Schennach (2004b) developed an unbiased, Nadaraya-Watson based estimator that can handle heteroskedastic model error. However, our situation is completely different in that we consider a semiparametric regression model (i.e., $m(X, Z; \beta)$), not a nonparametric one (i.e., $m(X, Z)$).

Thus, as far as we are aware, we believe our semiparametric approach provides advantages over existing methods in that it bypasses estimating $p_{\epsilon|X,Z}$ and $p_{X|Z}$, and simultaneously handles unspecified heteroskedastic model error and mismeasured covariates. It is important to note that our method is developed under specific assumptions in Section 1.1, among which require multiple proxy variables and classical measurement error (Assumption (iv)). Under Assumption (iv), we may easily estimate Ω_U in the measurement error distribution (Section 2.1) and thus, more easily identify estimating equations for β (Theorem 1). When this assumption no longer holds, the estimation procedure is more difficult: a more general method is needed to simultaneously estimate Ω_U and β. Work in this area has been explored; see Hu and Schennach (2008) and Chen et al. (2009) for developments in non-classical measurement error and estimation without available replicates (Chen et al., 2009).

The rest of the paper is as follows. Section 2 establishes identifiability results for the model parameters. Section 3 describes the main results for the semiparametric estimator, including theoretical properties, robustness to misspecifications of working distributions and its numerical implementation. We show the satisfying performance of the estimator through a simulation study in Section 4 and a data example in Section 5. Section 6 concludes the paper with a brief discussion. Technical proofs and additional simulation results are provided in the Supplementary Material. All computer codes are available upon request.

2 Identification

2.1 Identification of Ω_U

The identification of Ω_U is facilitated by the observed replicates. If replicates are unavailable, then validation data (Lee and Sepanski, 1995) or instrumental variables (Carroll et al., 2004)

can be used.

To identify Ω_U, we use the usual components of variance analysis (Carroll et al., 2006, chap. 4). Define $W_i = \sum_{j=1}^{\ell} W_{ij}/\ell$ and $V_i = \sum_{j=1}^{\ell} (W_{ij} - W_i)(W_{ij} - W_i)^T$. Then $\Omega_U = E(V_i)/(l-1)$, hence it is identifiable. In practice, we solve

$$\sum_{i=1}^{n} \left(\frac{V_i}{\ell - 1} - \Omega_U \right) = 0 \tag{3}$$

to obtain $\widehat{\Omega}_U$.

2.2 Identification of β

We demonstrate identifiability of β by casting the RMM with measurement error into a semiparametric framework. Let $\eta_1(x,z) \equiv p_{X|Z}(x|z)$, $\eta_2(\epsilon, x, z) \equiv p_{\epsilon|X,Z}(\epsilon|x,z)$, and $\eta_3(z) \equiv p_Z(z)$ denote infinite-dimensional nuisance parameters corresponding to the unknown distributions. Let W denote the average of the observed replicates and $p_{W|X,Z}(w|x,z;\alpha)$ denote its conditional distribution given (X,Z), with $\alpha = \text{vech}(\Omega_U)$ (i.e., the vectorized version of the upper block of Ω_U including its diagonal). Then, the probability density function of (Y,W,Z) is

$$\begin{aligned} &p_{Y,W,Z}(y,w,z;\beta,\alpha,\eta_1,\eta_2,\eta_3) \\ &= \int \eta_2\{y - m(x,z;\beta), x, z\} p_{W|X,Z}(w|x,z;\alpha)\eta_1(x,z)\eta_3(z)d\mu(x), \end{aligned} \tag{4}$$

where $d\mu(\cdot)$ denotes the dominating measure, which is the Lebesgue measure for continuous variables and the counting measure for discrete variables. The density of (Y,W,Z) contains both finite and infinite-dimensional parameters, hence the RMM with measurement error is a semiparametric model.

The identifiability of β in the RMM with measurement error is closely linked to the identifiability of β in the RMM without measurement error. To see this, assume to the contrary that the RMM without measurement error is identifiable, but that β in the RMM with measurement error is not. Then, there exist $\beta_0, \eta_1, \eta_2, \eta_3$ and $\beta^\dagger, \eta_1^\dagger, \eta_2^\dagger, \eta_3^\dagger$ where $\beta_0 \neq \beta^\dagger$,

but $\beta_0, \eta_1, \eta_2, \eta_3$ and $\beta^\dagger, \eta_1^\dagger, \eta_2^\dagger, \eta_3^\dagger$ yield the same data generation procedure:

$$\begin{aligned} p_{Y,W,Z}(y,w,z;\beta,\alpha,\eta_1,\eta_2,\eta_3) &= \int \eta_2\{y - m(x,z;\beta_0),x,z\}\eta_1(x,z)\eta_3(z)p_U(w-x;\alpha)dx \\ &= \int \eta_2^\dagger\{y - m(x,z;\beta^\dagger),x,z\}\eta_1^\dagger(x,z)\eta_3^\dagger(z)p_U(w-x;\alpha)dx. \end{aligned}$$

Here, $p_U(u;\alpha)$ denotes the measurement error distribution. Deconvolution then implies that for all (Y,X,Z), $\eta_2\{y-m(x,z;\beta_0),x,z\}\eta_1(x,z)\eta_3(z) = \eta_2^\dagger\{y-m(x,z;\beta^\dagger),x,z\}\eta_1^\dagger(x,z)\eta_3^\dagger(z)$. A similar argument to

$$p_{W,Z}(w,z;\alpha,\eta_1,\eta_3) = \int p_U(w-x;\alpha)\eta_1(x,z)\eta_3(z)dx = \int p_U(w-x;\alpha)\eta_1^\dagger(x,z)\eta_3^\dagger(z)dx,$$

yields $\eta_1(x,z)\eta_3(z) = \eta_1^\dagger(x,z)\eta_3^\dagger(z)$ for all (x,z). Together, these results imply that on the support of the probability density of (x,z),

$$\eta_2\{y - m(x,z;\beta_0),x,z\} = \eta_2^\dagger\{y - m(x,z;\beta^\dagger),x,z\} \tag{5}$$

for all (Y,X,Z). Hence, (5) implies that the conditional model error distributions under β_0 and under $\beta^\dagger$ are identical which makes the RMM without measurement error not identifiable. This contradicts our original assumption. Therefore, we have identifiability as long as we begin with an identifiable RMM without measurement error. Identifiability of the RMM without measurement error depends on the specific form of the mean model and is generally straight-forward to establish.

3 Methodology

3.1 Estimation of Ω_U and β

Estimation of Ω_U, equivalently $\alpha = \text{vech}(\Omega_U)$, follows directly from the solution to (3). Estimation of β builds upon the semiparametric results for an RMM without measurement error. For this latter case, Tsiatis (2006) demonstrated that consistent estimators are the

solutions to the linear estimating equation

$$\sum_{i=1}^{n} A(X_i, Z_i)\{Y_i - m(X_i, Z_i; \beta)\} = 0.$$

Here, $A(X, Z) \in \mathbb{R}^p$ is an arbitrary function that does not cause the above estimating equation to degenerate. If $A(X, Z) = \partial m(X, Z; \beta)/\partial\beta E(\epsilon^2|X, Z)^{-1}$, then the equation is named the optimal generalized estimating equation (optimal GEE; Liang and Zeger, 1986), and it yields the efficient estimator. See Section S.1 (Supplementary Material) for a brief overview of the semiparametric procedure and its application to the RMM without measurement error.

Applying the semiparametric procedure to the RMM *with* measurement error, we establish in Theorem 1 the condition that any consistent estimator for β must satisfy. A detailed derivation is given in Section S.2 (Supplementary Material).

Theorem 1 *For the RMM with measurement error, a consistent estimator for β is the solution to $\sum_{i=1}^{n} f(Y_i, W_i, Z_i; \beta) = 0$ where f is a p-dimensional function in*

$$\Lambda^{\perp} = [f(Y, W, Z) : E\{f(Y, W, Z)|Y, X, Z\} = g(X, Z)\epsilon].$$

Here, g is an arbitrary function of (X, Z) with finite variance, and

$$E\{f(Y, W, Z)|Y, X, Z\} = \int f(y, w, z) p_{W|X,Z}(w|x, z; \alpha) d\mu(w). \tag{6}$$

Theorem 1 states that to determine if a function $f(Y, W, Z; \beta)$ yields a consistent estimator for β, one must verify that f belongs to $\Lambda^{\perp}$. The verification involves computing the integral in (6) and checking that the result is of the form $g(X, Z)\epsilon$ for some function $g(X, Z)$. Note that the integration in (6) does not involve the unknown distributions η_1 or η_2. Instead, it only involves the distribution $p_{W|X,Z}(w|x, z; \alpha)$ which is completely known once α is estimated from (3). This observation means that even without knowing η_1 and η_2, one can verify if a function f belongs to $\Lambda^{\perp}$, and thus use it to form a consistent estimator for β.

Unfortunately, finding f that belongs to $\Lambda^{\perp}$ is not a trivial task. It is equivalent to the challenge of finding a corrected score which is only resolved for generalized linear models

(Nakamura, 1990). An approximate corrected score is possible using complex-variable computations and Monte Carlo averaging (Novick and Stefanski, 2002). In this work, we use a careful analytic derivation to construct f in $\Lambda^{\perp}$.

Let $\eta_1^*(x,z)$ and $\eta_2^*(\epsilon,x,z)$ be working models of η_1 and η_2, respectively. The working models may be completely different from the true models, denoted as η_{10},η_{20}, but we assume the support is the same. Throughout, let $E_*(\cdot)$ denote the expectation computed under η_1^*, η_2^*, and $E(\cdot)$ denote the expectation computed under η_{10},η_{20}. Define conjugate linear operators

$$\mathcal{K}_1\{h(Y,X,Z)\} = E_*\{h(Y,X,Z)|Y,W,Z\}, \quad \mathcal{K}_2\{f(Y,W,Z)\} = E\{f(Y,W,Z)|Y,X,Z\}.$$

It is important to note that $\mathcal{K}_2$ is independent of η_1^*,η_2^* as evident from (6); hence its definition is asterisk-free.

Using the projection theorem (Rudin, 1987), we demonstrate in Section S.3 (Supplementary Material) that a function in $\Lambda^{\perp}$ is $\mathcal{K}_1\{d^*(Y,X,Z)\}$ where $d^*(Y,X,Z)$ is a p-dimensional function that satisfies

$$\epsilon E_*(d^*\epsilon|X,Z) + \mathcal{K}_2\circ\mathcal{K}_1(d^*)E_*(\epsilon^2|X,Z) - \epsilon E_*\{\mathcal{K}_2\circ\mathcal{K}_1(d^*)\epsilon|X,Z\} = m'_\beta(X,Z;\beta)\epsilon. \quad (7)$$

Here, $\circ$ denotes the composite operation and $m'_\beta(X,Z;\beta)$ is $\partial m(X,Z;\beta)/\partial\beta$. To see that $\mathcal{K}_1\{d^*(Y,X,Z)\}$ indeed belongs to $\Lambda^{\perp}$, we can easily re-arrange (7) to show that $E[\mathcal{K}_1\{d^*(Y,X,Z)\}|Y,X,Z] = g(X,Z)\epsilon$ with $g(X,Z) = (m'_\beta(X,Z;\beta) - E_*[\{d^* - \mathcal{K}_2\circ\mathcal{K}_1(d^*)\}\epsilon|X,Z])E_*(\epsilon^2|X,Z)^{-1}$. It is worth noting that in the terminology of semiparametric theory, $\mathcal{K}_1\{d^*(Y,X,Z)\}$ is known as the locally efficient score vector

$$S^*_{\text{eff}}(Y,W,Z;\beta,\alpha,\eta_1^*,\eta_2^*) \equiv \mathcal{K}_1\{d^*(Y,X,Z)\}.$$

A few remarks are in order. First, equation (7) may admit more than one solution d^*. However, by the projection theorem (Rudin, 1987), even if d^* is not unique, $\mathcal{K}_1\{d^*(Y,X,Z)\}$ is unique; see Section S.3 (Supplementary Material). Hence differences in numerical procedures for obtaining d^* will not affect the final estimating equation which

is formed using $S^*_{\text{eff}}(Y, W, Z; \beta, \alpha, \eta_1^*, \eta_2^*) \equiv \mathcal{K}_1\{d^*(Y, X, Z)\}$. Second, to ensure that the parameter values are identified from the ensuing estimating equation, we require that $E\{S^*_{\text{eff}}(Y_i, W_i, Z_i; \beta, \alpha, \eta_1^*, \eta_2^*)\} = 0$ has unique root. Third, even if the unique root property holds at the population level, the estimating equation may still have multiple roots at the sample level. As far as we are aware, selecting among the multiple roots in estimating equations is a thorny issue; empirical knowledge for root selection is usually needed in practice. Lastly, because $\mathcal{K}_1\{d^*(Y, X, Z)\}$ is constructed to be an element of $\Lambda^\perp$ and all elements of $\Lambda^\perp$ yield consistent estimators for β (Theorem 1), the choice of η_1^*, η_2^* in forming $\mathcal{K}_1\{d^*(Y, X, Z)\}$ does not affect consistency. See Section 3.3 for a discussion of choosing η_1^*, η_2^* in practice. To the best of our knowledge, this is the only existing root-n consistent estimator for the RMM with measurement error that does not require estimating the unknown η_1, η_2.

3.2 Algorithm for estimating Ω_U and β

The algorithm for estimating Ω_U and β in model (1) and (2) is as follows.

1. Recall that $W_i = \sum_{j=1}^{\ell} W_{ij}/\ell$ and $V_i = \sum_{j=1}^{\ell}(W_{ij} - W_i)(W_{ij} - W_i)^T$. Solve for $\widehat{\Omega}_U$ as the root of (3) and form $\widehat{\alpha}_n = \text{vech}(\widehat{\Omega}_U)$.

2. Propose a working density model η_1^* for η_1.

3. Propose a working density model η_2^* for η_2 that satisfies $E_*(\epsilon|X, Z) = 0$.

4. Perform $\mathcal{K}_1, \mathcal{K}_2, E_*(\cdot|X, Z)$ under $p_{W|X,Z}(w|x, z; \widehat{\alpha}_n)$, η_1^*, and η_2^*. Solve for $d^*(Y, X, Z)$ from (7). When (7) admits more than one solution, pick one arbitrarily.

5. Form the score vector $S^*_{\text{eff}}(Y, W, Z; \beta, \widehat{\alpha}_n, \eta_1^*, \eta_2^*) = \mathcal{K}_1(d^*)$ by calculating $\mathcal{K}_1$ under η_1^* and $p_{W|X,Z}(w|x, z; \widehat{\alpha}_n)$. Even if (7) has multiple solutions, they will yield the same $\mathcal{K}_1(d^*)$ (Rudin, 1987).

6. Solve the estimating equation $\sum_{i=1}^{n} S^*_{\text{eff}}(Y_i, W_i, Z_i; \beta, \widehat{\alpha}_n, \eta_1^*, \eta_2^*) = 0$ for the estimator $\widehat{\beta}_n$.

In estimating β, we have treated α via a plug-in estimator obtained from Step 1. Alternatively, we can also augment α to β and simultaneously estimate both using the procedure from Step 2 on. That is, we may solve for $\widehat{\theta}_n = (\widehat{\alpha}_n^T, \widehat{\beta}_n^T)^T$ as the root of

$$\sum_{i=1}^{n} \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \eta_1, \eta_2) = 0. \tag{8}$$

Here, $\mathcal{S} = (\phi^T, f^T)^T$ with ϕ denoting the estimating equations in (3) corresponding to the α elements, and $f \in \Lambda^{\perp}$. In our algorithm, we set $f = S^*_{\text{eff}}$, and $\eta_1 = \eta_1^*, \eta_2 = \eta_2^*$. Solving for $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ simultaneously does not change the analysis.

The numerical implementation of the algorithm is given in Section 3.5. We now give some remarks regarding the algorithm.

3.3 Selection and impact of working models η_1^*, η_2^*

One flexibility of our algorithm is the ability to choose possibly incorrect, working models η_1^*, η_2^* for η_1, η_2 (Steps 2 and 3 of the algorithm). We now discuss a practical approach for selecting these working models and its impact on the consistency and efficiency of $\widehat{\beta}_n$.

Remark 1 *When either or both η_1^*, η_2^* are misspecified and the measurement error distribution is estimated as $p_{W|X,Z}(w|x,z;\widehat{\alpha}_n)$, the algorithm still provides a consistent estimator.*

To prove the consistency claim in Remark 1, we make the following regularity conditions, stated using the general $\mathcal{S} = (\phi^T, f^T)^T$ and $\theta = (\alpha^T, \beta^T)^T$ notation. We assume θ belongs to a domain of interest $\boldsymbol{\Theta}$ which is a compact set.

(*R1*) The estimating equation in (8) and its expectation $E\{\mathcal{S}(Y, W, V, Z; \theta, \eta_1, \eta_2)\}$ are sufficiently smooth in (θ, η_1, η_2) in a neighborhood of $(\theta_0, \eta_{10}, \eta_{20})$. This condition is needed so that the weak law of large numbers is valid.

(*R2*) The matrix $E\{\partial \mathcal{S}(Y, W, V, Z; \theta, \eta_1, \eta_2)/\partial \theta^T\}$ is invertible, bounded and smooth in (θ, η_1, η_2) in a neighborhood of $(\theta_0, \eta_{10}, \eta_{20})$. This assumption permits the re-arrangement of a Taylor expansion and hence, the applicability of the central limit theorem.

(*R3*) For $\eta_1 = \eta_1^*, \eta_2 = \eta_2^*$, the equation $E\{\mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \eta_1, \eta_2)\} = 0$ has a unique solution and $E\{\sup_{\theta \in \mathbf{\Theta}} |\mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \eta_1, \eta_2)|\} < \infty$ component wise. The unique solution requirement is commonly needed in semiparametric estimation and in parametric estimation, except when the objective function guarantees the unique root property such as when it is convex. While a globally unique root property is somewhat restrictive, one can instead require a unique root in a region of interest, so long as it is justifiable to consider parameters only in that region.

We show in Section S.4 (Supplementary Material) that under these regularity conditions, $\widehat{\theta}_n$ is a consistent estimator even when $\eta_1 = \eta_1^*, \eta_2 = \eta_2^*$ are misspecified.

From Remark 1, in terms of obtaining a consistent estimator, we are free to choose any working model η_1^* and η_2^*. Thus, for computational ease, we suggest using Gaussian models due to their simplicity. We also recommend choosing the support of η_1^*, η_2^* to be as large as that of the true distributions so as to maintain numerical stability. Of course, the true distributions are unknown, so the latter requirement may be achieved by choosing the support based on the observed data. For example, after centering the observed data (Y, W, Z), one may choose η_1^* to be a normal distribution with mean zero and variance equal to the sample variance of W. Likewise, one may choose η_2^* to be a normal distribution with mean zero and variance as estimated from the residual sum of squares after regressing Y on $m(W, Z; \beta)$.

Although any working models η_1^*, η_2^* maintain consistency, they affect efficiency in theory as we now describe.

Remark 2 *The choice of η_1^*, η_2^* only affects β, so we characterize the efficiency for β only with α fixed at the truth. When the working models are correct, i.e., $\eta_1^* = \eta_{10}$ and $\eta_2^* = \eta_{20}$, the algorithm gives the optimal estimator in that its estimation variance achieves the semiparametric efficiency bound (Tsiatis, 2006, chap. 4). Such results follow because, in this case, the resulting estimator solves the true efficient score estimating equation $\sum_{i=1}^n S_{\text{eff}}(Y_i, W_i, Z_i; \widehat{\beta}_n, \alpha_0, \eta_{10}, \eta_{20}) = 0$.*

Justification of Remark 2 follows from the principles of semiparametric theory (Tsiatis, 2006, chap. 4). From Remark 2, if the working models η_1^*, η_2^* are exactly the true models η_{10}, η_{20},

then the resulting estimator $\widehat{\beta}_n$ is most efficient. Of course knowing the true models η_{10}, η_{20} is rarely an option. Hence, some efficiency loss is expected since the working models will most likely differ from the truth. The incurring loss depends on the proposed working models, and can be theoretically characterized as follows.

Let S^*_{eff} be as in Step 5 of the algorithm which is constructed under the possibly misspecified working models η_1^*, η_2^*. Let $A_* = E\{\partial S^*_{\text{eff}}(Y, W, Z; \beta_0, \alpha_0, \eta_1^*, \eta_2^*)/\partial\beta\}$ and $B_* = \text{var}\{S^*_{\text{eff}}(Y, W, Z; \beta_0, \alpha_0, \eta_1^*, \eta_2^*)\}$, where β_0, α_0 denote the true parameter values. The asterisks in A_* and B_* are used to emphasize that S^*_{eff} depends on the working models. Finally, let A, B and S_{eff} be defined analogously to A_*, B_*, and S^*_{eff}, respectively, except with $\eta_1^* = \eta_{10}$ and $\eta_2^* = \eta_{20}$.

In Theorem 2 (see Section 3.4), we demonstrate that under working models η_1^*, η_2^*, the estimator $\widehat{\beta}_n$ is asymptotically normal with mean zero and variance-covariance $A_*^{-1}B_*(A_*^{-1})^T$. In comparison, under the true η_{10}, η_{20}, the asymptotic variance-covariance of $\widehat{\beta}_n$ is $A^{-1}B(A^{-1})^T$. Therefore, the theoretical efficiency loss of the estimator computed under misspecified working models and the truth is the difference $A_*^{-1}B_*(A_*^{-1})^T - A^{-1}B(A^{-1})^T$. This difference is identical to $E\{(A_*^{-1}S^*_{\text{eff}} - A^{-1}S_{\text{eff}})(A_*^{-1}S^*_{\text{eff}} - A^{-1}S_{\text{eff}})^T\}$ (see Section S.5 in Supplementary Material), which means that the efficiency loss is positive definite. The precise efficiency loss can thus be evaluated in each case. In our limited empirical studies (see Section 4.3.2), it has been observed that the loss is generally small, and the estimation variance is quite insensitive to the choice of the working models.

In summary, our procedure allows flexible working models η_1^*, η_2^* to construct consistent estimators and achieves local efficiency. This contrasts from existing methods in the literature, including that from Tsiatis and Ma (2004), which are highly sensitive to the variance misspecification of the model error. Moreover, in bypassing the estimation of η_1, η_2, our algorithm minimizes the unnecessary work in the process of estimating α and β.

3.4 Theoretical properties

We describe the theoretical properties of $\widehat{\theta}_n = (\widehat{\alpha}_n^T, \widehat{\beta}_n^T)^T$ under working models $\eta_1^*(x, z; \gamma_1)$ and $\eta_2^*(\epsilon, x, z; \gamma_2)$ where γ_1, γ_2 are finite-dimensional parameters. The parameters γ_1, γ_2 re-

flect the common practice of using parametric forms for the working models η_1^*, η_2^*. The true forms η_{10}, η_{20} may or may not belong to these working model families.

Let $\gamma = (\gamma_1^T, \gamma_2^T)^T$ belong to a compact set $\boldsymbol{\mathcal{G}}$ and $\widehat{\gamma}_n$ be an estimator of γ. We assume $\widehat{\gamma}_n$ is root-n consistent in the proposed working models, so $n^{1/2}(\widehat{\gamma}_n - \gamma^*)$ is bounded in probability for some constant γ^*. We now demonstrate that under $\eta_1^*(x, z; \gamma_1)$ and $\eta_2^*(\epsilon, x, z; \gamma_2)$, the estimator $\widehat{\theta}_n$ is asymptotically normal, and its efficiency does not depend on how efficiently we estimate γ.

To establish these results, we further make the following assumptions:

($R4$) The equation $E\{\mathcal{S}(Y_i, W_i, Z_i, V_i; \theta, \gamma^*)\} = 0$ has a unique solution. In addition, $E\{\sup_{\theta \in \boldsymbol{\Theta}, \gamma^* \in \boldsymbol{\mathcal{G}}} |\mathcal{S}(Y_i, W_i, Z_i, V_i; \theta, \gamma^*)|\} < \infty$ component wise, and the expectation of the squared l_2 norm of $\mathcal{S}$, i.e. $E\{\|\mathcal{S}(Y_i, W_i, Z_i, V_i; \theta_0, \gamma^*)\|^2\}$, is bounded. This condition is similar to condition ($R3$).

($R5$) $n^{-1} \sum_{i=1}^n \partial \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \gamma)/\partial \theta$ converges in probability to $E\{\partial \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \gamma)/\partial \theta\}$ uniformly in (θ, γ) in a neighborhood of (θ_0, γ^*).

($R6$) $n^{-1} \sum_{i=1}^n \partial \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \gamma)/\partial \gamma$ converges in probability to $E\{\partial \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \gamma)/\partial \gamma\}$ uniformly in (θ, γ) in a neighborhood of (θ_0, γ^*).

The last two conditions are very mild and are generally satisfied following the law of large lumbers and equicontinuity conditions.

Our first theoretical result shows that $\widehat{\theta}_n$ is asymptotically normal whether $\eta_1^*(x, z; \gamma_1)$ and $\eta_2^*(\epsilon, x, z; \gamma_2)$ contain the true η_{10}, η_{20} or not.

Theorem 2 *Let f be an arbitrary p-dimensional function belonging to $\Lambda^\perp$ in Theorem 1. Let $\eta_1^*(x, z; \gamma_1)$ and $\eta_2^*(\epsilon, x, z; \gamma_2)$ be working parametric models for η_1, η_2. Let $\gamma = (\gamma_1^T, \gamma_2^T)^T$ and $\widehat{\gamma}_n$ be its estimate such that for some constant γ^*, $n^{1/2}(\widehat{\gamma}_n - \gamma^*)$ is bounded in probability. Finally, let $\widehat{\theta}_n = (\widehat{\alpha}_n^T, \widehat{\beta}_n^T)^T$ and $\theta_0 = (\alpha_0^T, \beta_0^T)^T$ denote the truth. Under regularity conditions (*R1*)-(*R6*), the root $\widehat{\theta}_n$ of $\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \widehat{\gamma}_n) = 0$ is consistent and*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to \mathit{Normal}(0, \mathcal{V}_*)$$

in distribution as $n \to \infty$. *Here,* $\mathcal{V}_* = \mathcal{A}_*^{-1}\mathcal{B}_*(\mathcal{A}_*^{-1})^T$ *with* $\mathcal{A}_* = E\{\partial \mathcal{S}(Y, W, V, Z; \theta_0, \gamma^*)/\partial\theta^T\}$ *and* $\mathcal{B}_* = diag[var\{\phi(V; \alpha_0)\}, var\{f(Y, W, Z; \theta_0, \gamma^*)\}]$, *a block diagonal matrix.*

In Theorem 2, the arguments in $\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \widehat{\gamma}_n) = 0$ differ from those in (8) in that we have replaced η_1^*, η_2^* with $\widehat{\gamma}_n = (\widehat{\gamma}_{n1}^T, \widehat{\gamma}_{n2}^T)^T$ to emphasize our use of parametric working models for η_1^*, η_2^*. The results in Theorem 2 hold because we can express $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ as a summand of normalized, zero-mean random vectors based on Taylor expansion and the properties of $\Lambda^\perp$. Consequently, by our regularity assumptions and the central limit theorem, this normalized sum will converge in distribution to a multivariate normal with zero mean and variance-covariance $\mathcal{V}_*$; see Section S.6 (Supplementary Material) for complete details. In addition, the result in Theorem 2 is useful for performing inference on θ where $\mathcal{V}_*$ in practice is estimated by the sandwich estimator $\widehat{\mathcal{A}}_*^{-1}\widehat{\mathcal{B}}_*(\widehat{\mathcal{A}}_*^{-1})^T$. Here,

$$\begin{aligned}
\widehat{\mathcal{A}}_* &= n^{-1}\sum_{i=1}^n \partial \mathcal{S}(Y_i, W_i, V_i, Z_i; \widehat{\theta}_n, \widehat{\gamma}_n)/\partial\theta^T, \\
\widehat{\mathcal{B}}_* &= n^{-1}\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i; \widehat{\theta}_n, \widehat{\gamma}_n)\mathcal{S}^T(Y_i, W_i, V_i, Z_i; \widehat{\theta}_n, \widehat{\gamma}_n).
\end{aligned}$$

Remark 3 *The result in Theorem 2 applies to any function* $f \in \Lambda^\perp$. *In Section 3.1, we argued that a particular function in* $\Lambda^\perp$ *is* $S_{\text{eff}}^* = \mathcal{K}_1(d^*)$. *Thus, by Remark 1 and Theorem 2, when* $\mathcal{S} = (\phi^T, S_{\text{eff}}^{*\text{T}})^T$, *the resulting estimator* $\widehat{\theta}_n$ *from our proposed algorithm is consistent and asymptotically normal.*

Our second theoretical result demonstrates that the asymptotic efficiency of $\widehat{\theta}_n$ does not depend on how efficiently we estimate γ in the working parametric models. Specifically, consider the case when $\widehat{\theta}_n$ solves the estimating equation $\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \widehat{\gamma}_n) = 0$, and $\check{\theta}_n$ solves the estimating equation $\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \gamma^*) = 0$, where $\mathcal{S} = (\phi, f^T)^T$ and f belongs to $\Lambda^\perp$ in Theorem 1. Our previous results from Theorem 2 warrant that $\widehat{\theta}_n$ and $\check{\theta}_n$ are root-n consistent estimators and asymptotically normal. A stronger result, shown below, is that $\widehat{\theta}_n$ and $\check{\theta}_n$ also have the same asymptotic efficiency even though the former involves the estimated $\widehat{\gamma}_n$, and the latter only involves the constant γ^*. Thus, as long as we consistently estimate $\widehat{\gamma}_n$, then using either $\widehat{\gamma}_n$ or γ^* in the working parametric models yields

the same efficiency for $\widehat{\theta}_n$.

Theorem 3 *Let the p-dimensional function f belong to $\Lambda^{\perp}$ in Theorem 1. Assume $\widehat{\gamma}_n$ is such that $n^{1/2}(\widehat{\gamma}_n - \gamma^*)$ is bounded in probability. Then, under regularity conditions (*R1*)-(*R6*), the efficiency of the estimator $\widehat{\theta}_n$ obtained as the root of $\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \widehat{\gamma}_n) = 0$ is asymptotically equivalent to the efficiency of the estimator $\check{\theta}_n$ obtained as the root of $\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta, \gamma^*) = 0$. Namely, both $n^{1/2}(\widehat{\theta}_n - \theta_0)$ and $n^{1/2}(\check{\theta}_n - \theta_0)$ are asymptotically normal with mean zero and variance-covariance $\mathcal{V}_*$ as in Theorem 2.*

The proof of Theorem 3 follows analogously to that of Theorem 2 in that $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ and $\sqrt{n}(\check{\theta}_n - \theta_0)$ can be expressed as the same summand of normalized, zero-mean random vectors via Taylor expansion; see Section S.7 (Supplementary Material). Thus, because the first order expansions of $\widehat{\theta}_n$ and $\check{\theta}_n$ are the same, it immediately follows from the regularity conditions and central limit theorem that both estimators are asymptotically normal with identical variance-covariance $\mathcal{V}_*$.

The results in Theorems 2 and 3 hold whether or not $\eta_1^*(x, z; \gamma_1)$ and $\eta_2^*(\epsilon, x, z; \gamma_2)$ contain the true distributions η_{10}, η_{20}. However, when the working parametric models do contain the true distributions, the resulting estimator $\widehat{\theta}_n$ is actually semiparametric efficient as noted below.

Remark 4 *A particularly interesting case is when f is the efficient score S_{eff}^* as in our algorithm. Since $S_{\text{eff}}^* \in \Lambda^{\perp}$, Theorem 3 tells us that if correct parametric models with parameters γ are used for $\eta_1(x, z), \eta_2(\epsilon, x, z)$, and root-n estimators can be found for the parameters γ, then it is as if $\eta_1(x, z)$, $\eta_2(\epsilon, x, z)$ were known precisely. In this case, we achieve optimal semiparametric efficiency. This is a stronger statement than Remark 2.*

In practice, a correct parametric model is certainly not easy to obtain. It requires good knowledge of $\eta_1(x, z)$ and $\eta_2(\epsilon, x, z)$, both of which are "invisible" due to the unobservable X's. Thus, if reducing estimation variability is important, one can propose a relatively large model for η_1 and η_2, and proceed with the locally efficient estimator. With richer models of η_1, η_2, the chance of achieving efficiency is increased.

3.5 Implementation of the algorithm

Steps 1-3 in our algorithm are easily handled by following the guidance in Section 3.3 for selecting η_1^*, η_2^*. We thus focus on the details for executing Steps 4-6.

Step 4 requires solving for $d^*(Y, XZ)$ from the ill-posed problem in (7). Although this ill-posed problem may at first appear challenging, we benefit from two aspects. First, solving for d^* is a "good" ill-posed problem in the sense that the ill-posedness is only because more than one solution may satisfy (7). This is beneficial since our objective is to find any one of these solutions. Second, what we really need for estimation and inference is not d^* itself, but a smoothed version of d^*, namely $\mathcal{K}_1(d^*) = E(d^*|Y, W, Z)$ which is unique and hence no longer an ill-posed problem. We now demonstrate how (7) can be solved analytically in some cases and numerically otherwise.

3.5.1 Analytic d^*

For some mean models, d^* may be computed analytically such as for the simple, linear RMM with two replicates:

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i, \quad W_{ij} = X_i + U_{ij}, \quad E(\epsilon_i|X_i) = 0,$$

for $i = 1, \ldots, n$, $j = 1, 2$. Here, U_{ij} is normally distributed with mean zero and unknown variance $2\sigma_U^2$.

Following our algorithm, solve for $\widehat{\sigma}_U^2$ from (3) and let $W_i = (W_{i1} + W_{i2})/2$. With $\eta_1 \equiv p_X(x)$ and $\eta_2 \equiv p_{\epsilon|X}(\epsilon|x)$, we suppose that (Y_i, W_i) are standardized so that it is reasonable to posit η_1^*, η_2^* as standard normals. Then, under η_1^*, η_2^*, an analytic solution to (7) is $d^* = (d_1^*, d_2^*)^T$ with

$$\begin{aligned} d_1^*(Y, X) &= Y - \beta_1 - \beta_2 X(1 + c_1^{-1}\widehat{\sigma}_U^2), \\ d_2^*(Y, X) &= c_2^{-1}\beta_2\widehat{\sigma}_U^2 \left\{c_1(1 - \beta_1^2) + \widehat{\sigma}_U^2 + 1 - c_1(Y - 2\beta_1)Y\right\} + \\ &\quad c_2^{-1}(c_1 + \widehat{\sigma}_U^2)X \left\{(2c_1 - 1)(Y - \beta_1) - \beta_2(c_1 + \widehat{\sigma}_U^2)X\right\}, \end{aligned}$$

and $c_1 = 1 + \beta_2^2\widehat{\sigma}_U^2$, $c_2 = c_1(1 + 2\widehat{\sigma}_U^2) - \widehat{\sigma}_U^2$.

Using the analytic d^* to form the score vector $S^*_{\text{eff}}(Y, W; \beta, \widehat{\sigma}^2_U, \eta^*_1, \eta^*_2) = \mathcal{K}_1(d^*)$ then yields that $\widehat{\beta}_n$ solves

$$0 = Cn^{-1} \sum_{i=1}^{n} \left\{ \begin{pmatrix} 1 & W_i \\ W_i & W_i^2 - \widehat{\sigma}^2_U \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - \begin{pmatrix} Y_i \\ Y_i W_i \end{pmatrix} \right\},$$

where $C = \text{diag}(c_1^{-1}, -c_2^{-1})$. Because C is non-singular, the above estimating equation is exactly the same explicit form previously given in Hall and Ma (2007). In other words, the estimator in Hall and Ma (2007) is a special case of our solution family corresponding to a natural choice of standard normals for the working models η^*_1, η^*_2.

3.5.2 Numerical d^*

For general mean models, d^* is computed numerically. The implementation below is provided in software available on the first-author's website. We propose solving for d^* by approximating it with a linear combination of basis functions. For ease of presentation, we demonstrate the procedure when X and Z are univariate; however, the method extends to the multivariate case.

In our approach, we approximate d^* by

$$d^*(Y, X, Z) = \sum_{j,k=1}^{q} c_{jk}(Z) g_k(Y) h_j(X),$$

where $c_{jk}(Z)$, $j, k = 1, \ldots, q$, is a p-dimensional vector of unknown coefficients, and $g_k(\cdot)$, $h_j(\cdot)$ are sets of real-valued basis functions (e.g., Hermite polynomials, Chebychev polynomials, Fourier series, B-splines, Legendre polynomials). The number of bases q is chosen to give accurate approximation and permit fast computation. The number of basis functions q is dependent on the true $d^*(Y, X, Z)$ function and on the type of basis functions. Empirically, we suggest to start from $q = 4$ and increase it until the result stabilizes.

With d^* as above, the goal then is to form (7) and solve for the coefficients $c_{jk}(Z)$,

$j, k = 1, \ldots, q$. To this end, (7) becomes

$$\begin{aligned} & \sum_{j,k=1}^{q} c_{jk}(Z)\epsilon h_j(X) E_*\{g_k(Y)\epsilon | X, Z\} \\ + & \sum_{j,k=1}^{q} c_{jk}(Z) g_k(Y) \mathcal{K}_2 \circ \mathcal{K}_1\{h_j(X)\} E_*(\epsilon^2 | X, Z) \\ - & \sum_{j,k=1}^{q} c_{jk}(Z)\epsilon E_*[g_k(Y)\mathcal{K}_2 \circ \mathcal{K}_1\{h_j(X)\}\epsilon | X, Z] = m'_\beta(X, Z; \beta)\epsilon. \end{aligned} \tag{9}$$

Under the working models η_1^* and η_2^*, we evaluate the expectations in (9) using discretization and quadrature integration (e.g., Hermite quadrature). Specifically, we discretize $\eta_1^*(x, z)$ at r points $x_1, \ldots, x_r$ across the support of X with weights given by $\eta_1^*(x, z) = \sum_{s=1}^{r} p_s(z) I(x = x_s)$ such that $\sum_{s=1}^{r} p_s(z) = 1$ for all z in the support of Z. Under this discretization, the terms in (9) are computed using the formulas

$$\begin{aligned} \mathcal{K}_1\{f_1(Y, X, Z)\} &= \frac{\sum_{s=1}^{r} f_1(Y, x_s, Z) p_{W|X,Z}(W | x_s, Z; \widehat{\alpha}_n) \eta_2^*\{Y - m(x_s, Z; \beta), x_s, Z\} p_s(Z)}{\sum_{s=1}^{r} p_{W|X,Z}(W | x_s, Z; \widehat{\alpha}_n) \eta_2^*\{Y - m(x_s, Z; \beta), x_s, Z\} p_s(Z)}, \\ \mathcal{K}_2\{f_2(Y, W, Z)\} &= \int f_2(Y, w, Z) p(w | X, Z; \widehat{\alpha}_n) d\mu(w), \\ E_*\{f_1(Y, X, Z) | X, Z\} &= \int f_1(y, X, Z) \eta_2^*\{y - m(X, Z; \beta), X, Z\} d\mu(y), \end{aligned}$$

for appropriate functions $f_1(Y, X, Z)$, $f_2(Y, W, Z)$. Finally, the integrals in $\mathcal{K}_2$ and $E_*(\cdot | X, Z)$ are evaluated using quadrature integration (Kress, 1999, chap. 12). It is important to note that our way of discretizing $\eta_1^*(x, z)$ simplifies the computation of $\mathcal{K}_1$ into a simple summation of functions evaluated at $x_1, \ldots, x_r$. Doing so avoids the complex task of estimating the unknown distribution $p_{X|Y,W,Z}(x | y, w, z)$. The number of discretization points r controls the integral approximation accuracy. Empirically, we suggest to use $r = 20$ and increase it until the results stabilize.

In the last step of solving for the coefficients of d^*, each term in (9) is evaluated at q^2 grid-points (y_m, x_ℓ, Z) for $m, \ell = 1, \ldots, q$, typically chosen as quadrature points. Doing so leads to p linear systems of size $q^2 \times q^2$, from which we may evaluate $c_{jk}(Z)$ at each observed Z. After obtaining the coefficients, we then verify that (7) is really solved by plugging in

the coefficients to d^*. The verification needs to be done only at the grid points (y_m, x_ℓ, Z) because d^* was only solved for at these grid points. By having to verify d^* only at these grid points rather than at all (Y, X, Z), we essentially bypass the functional nature of solving for d^*, which means solving for d^* is actually simpler than it appears.

After the coefficients of d^* are verified, we then do Step 5 and form

$$S^*_{\text{eff}}(Y, W, Z; \beta, \widehat{\alpha}_n, \eta_1^*, \eta_2^*) = \mathcal{K}_1(d^*) = \sum_{j,k=1}^{q} c_{jk}(Z) g_k(Y) \mathcal{K}_1\{h_j(X)\}$$

to construct $\sum_{i=1}^n S^*_{\text{eff}}(Y_i, W_i, Z_i; \beta, \widehat{\alpha}_n, \eta_1^*, \eta_2^*) = 0$. In Step 6, the estimator for β is then the root of the constructed estimating equation.

One possible concern about our proposed implementation is that the different numerical approximations may ultimately affect the efficiency of the proposed estimator. However, this is not the case. If d^* is constructed so that (7) is indeed satisfied, then $S^*_{\text{eff}} = \mathcal{K}_1(d^*)$ belongs to $\Lambda^\perp$ as stated in Section 3.1. Elements in $\Lambda^\perp$ lead to consistent estimators for β (see Theorem 1 and Remark 1) with efficiency affected only by the choice of η_1^*, η_2^*, not the approximation of d^* (see Remark 2 and the ensuing discussion). Therefore, a critical step is ensuring that the obtained d^* does indeed satisfy (7), which is exactly what we do. Therefore, solving for d^* is genuinely and completely a computational issue since no data is involved in the solution process. To ensure that (7) is properly solved, one may need to choose a rich class of basis functions, for example, combinations of polynomial bases, B-splines, or Fourier series. A full discussion of various methods to solve (7) is a well studied topic in numerical analysis and can be found in Kress (1999) and references therein.

4 Empirical Studies

We now demonstrate the performance of our method and compare its results to five competing methods.

4.1 Simulation design

We consider the RMM with measurement error

$$Y_i = \beta_2 \exp(-\beta_1 X_i^2) + \beta_3 Z_i + \epsilon_i, \quad W_{ij} = X_i + U_{ij}, \quad U_{ij} \sim \text{Normal}(0, 2\sigma_U^2),$$

for $i = 1, \ldots, n$ and $j = 1, 2$. The true model parameters are $(\sigma_U^T, \beta_1, \beta_2, \beta_3)^T = (0.05, 0.25, 0.7, 0.5)^T$. Results for other mean models are reported in the Supplementary Material (Section S.9).

The true distribution η_{10} of X is uniform on $[1.1 - \sqrt{0.9}, 1.1 + \sqrt{0.9}]$, and the true distribution η_{30} of Z is Bernoulli with parameter 0.5. To evaluate the robustness of our method, we set the model error distribution η_{20} to be either a uniform or t-distribution with 5 degrees of freedom (i.e., t_5 distribution), and its variance either homoskedastic or heteroskedastic. Specifically, we consider

Setting 1: Uniform distribution.

- Homoskedastic: η_{20} is uniform on $[-1, 1]$;
- Heteroskedastic: η_{20} is $(|X| + 1)\mathcal{U}$ where $\mathcal{U}$ is a uniform distribution on $[-1, 1]$.

Setting 2: t_5-distribution.

- Homoskedastic: η_{20} is $0.4t_5$;
- Heteroskedastic: η_{20} is $(0.4|X| + 0.5)t_5$.

4.2 Methods evaluated

For all settings, we generated 1000 data sets with sample size $n = 500$. Parameters $\sigma_U^2, \beta_1, \beta_2, \beta_3$ were estimated using six different methods.

4.2.1 Proposed method

We used our proposed method where we set working models η_1^*, η_2^* different from the true η_{10}, η_{20} both in terms of distributional form and variance structure. The differences are intended to demonstrate the robustness of our method when η_1^*, η_2^* differ from the truth.

In Settings 1 and 2, we let the working model η_1^* be Normal$(1.1, 0.9/3.5^2)$. In Setting 1, the working model η_2^* was Normal$(0, 0.9^2)$, and in Setting 2, η_2^* was Normal$(0, 1.7^2)$. While the working models have supports as large as the true distributions, the proposed η_2^* in no way accounts for the possible heteroskedasticity in η_{20}. Our approach was implemented following the procedure in Section 3.5 where d^* was computed numerically with $q = 7$ Hermite bases and $r = 20$ discretization points; all integrals were computed using Hermite quadrature.

4.2.2 Homoskedastic and heteroskedastic sieve estimator

The second and third method is a sieve MLE which either assumes homoskedastic or heteroskedastic model error. Specifically, the sieve MLE is the solution to

$$\arg\max_\beta \sup_{(f_1,f_2)} \frac{1}{n}\sum_{i=1}^{n} \ln \int f_1\{y_i - m(x, z_i; \beta)|x, z_i\} p_U(w_i - x; \widehat{\alpha}_n) f_2(x) d\mu(x), \qquad (10)$$

where f_1, f_2 are truncated series used to estimate the unknown distributions of $p_{\epsilon|X,Z}(\epsilon|x, z)$ and $p_{X|Z}(x|z)$, respectively. For our simulations, we have that $p_{X|Z}(x|z) = p_X(x)$ since X and Z are generated independently of each other; thus, f_2 is set to represent $p_X(x)$. Lastly, p_U corresponds to the normal distribution for $W_i = (W_{i1} + W_{i2})/2$ and $\widehat{\alpha}_n$ is the vectorized solution to (3).

We consider two different sieves for f_1. The first is a homoskedastic sieve where f_1 will estimate $p_\epsilon(\epsilon)$ and thus ignore any dependence between ϵ and (X, Z). The second is a heteroskedastic sieve where f_1 will estimate $p_{\epsilon|X,Z}(\epsilon|x, z)$ and thus account for any dependence between ϵ and (X, Z).

For the homoskedastic sieve, we use the work of Schennach and Hu (2013), and use

$$\sqrt{f_1(\epsilon)} = \sum_{j=0}^{\kappa_\epsilon} \xi_j^\epsilon t_j(\epsilon),$$

where κ_ϵ is a smoothing parameter, $t_j(x) = (\sqrt{\sqrt{\pi} j! 2^j})^{-1} H_j(x) \exp(-x^2/2)$ and H_j are Hermite polynomials. To ensure that $f_1(\epsilon)$ is a valid density and that $E(\epsilon) = 0$, we require that $\sum_{j=0}^{\kappa_\epsilon} (\xi_j^\epsilon)^2 = 1$ and $\sum_{j=1}^{\kappa_\epsilon - 1} \sqrt{2(j+1)} \xi_j^\epsilon \xi_{j+1}^\epsilon = 0$. We expect that this homoskedastic f_1 will perform well when ϵ is in fact homoskedastic, but we do expect bias when ϵ is in fact

heteroskedastic.

For the heteroskedastic sieve, we extended the work of Hu and Schennach (2008), and use

$$\begin{aligned}\sqrt{f_1(\epsilon|x,z)} &= \left[a_{00} + a_{01}\cos\left\{\frac{\pi}{\ell_x}m(x,z;\beta)\right\} + a_{02}\cos\left\{\frac{2\pi}{\ell_x}m(x,z;\beta)\right\}\right] \\ &+ \sum_{k=1}^{3}\left[a_{k0} + a_{k1}\cos\left\{\frac{\pi}{\ell_x}m(x,z;\beta)\right\} + a_{k2}\cos\left\{\frac{2\pi}{\ell_x}m(x,z;\beta)\right\}\right]\cos\left(\frac{k\pi}{\ell_e}\epsilon\right) \\ &+ \sum_{k=1}^{3}\left[b_{k0} + b_{k1}\cos\left\{\frac{\pi}{\ell_x}m(x,z;\beta)\right\} + b_{k2}\cos\left\{\frac{2\pi}{\ell_x}m(x,z;\beta)\right\}\right]\sin\left(\frac{k\pi}{\ell_e}\epsilon\right).\end{aligned}$$

By construction $m(x,z;\beta) \in [0,\ell_x]$ and we simulated data such that $\epsilon \in [-\ell_e,\ell_e]$ for an appropriate choice of ℓ_e, so as to align with the assumptions of Hu and Schennach (2008). Finally, to ensure that $f_1(\epsilon|x,z)$ is a valid density and that $E(\epsilon|X,Z)=0$, we impose twelve constraints given in Section S.8 (Supplementary Material). It is important to note that our heteroskedastic sieve above differs from that in Hu and Schennach (2008) in two ways. First, we use a sieve to estimate $p_{\epsilon|X,Z}$ rather than $p_{U|X,Z}$ as in Hu and Schennach (2008) who considered heteroskedastic measurement error, not heteroskedastic model error. Second, we further require that f_1 is always non-negative while Hu and Schennach (2008) did not impose that in their numerical studies. In terms of performance, we expect that the heteroskedastic f_1 will perform well whether ϵ is homoskedastic or heteroskedastic, since homoskedasticity is a special case of heteroskedascity (i.e., $\sigma_\epsilon(x) \equiv \sigma_\epsilon$).

Lastly, regardless of the form for f_1, we let

$$\sqrt{f_2(x)} = \sum_{j=0}^{\kappa_x}\xi_j^x t_j(x),$$

where κ_x is a smoothing parameter, and $t_j(x)$ is the Hermite representation as defined for the homoskedastic f_1. To ensure that f_2 is a valid density we require that $\sum_{j=0}^{\kappa_x}(\xi_j^x)^2 = 1$.

The homoskedastic and heteroskedastic sieve MLE is then the solution to the optimization problem in (10) subject to all constraints stated above: three for the homoskedastic sieve MLE and thirteen for the heteroskedastic sieve MLE. The integral in (10) is evaluated using

Hermite quadrature. We set the smoothing parameters $\kappa_\epsilon = 6$ and $\kappa_x = 6$ as in Schennach and Hu (2013), but other values were considered and yielded similar results (not reported).

4.2.3 Homoskedastic and heteroskedastic Tsiatis-Ma estimator

The fourth and fifth methods are based on the work of Tsiatis and Ma (2004). The Tsiatis-Ma (TM) estimator also uses a working model η_1^*, but requires η_2^* to yield a correctly specified variance structure. To demonstrate this sensitivity, we applied the TM estimator assuming homoskedastic model errors (TM-Homoskedastic) and assuming heteroskedastic model errors (TM-Heteroskedastic).

For both TM-Homoskedastic and TM-Heteroskedastic estimators, we set the working model η_1^* as Normal$(1.1, 0.9/3.5^2)$. For the TM-Homoskedastic estimator, we let η_2^* be Normal$(0, 1/3)$ in Setting 1 and Normal$(0, 4/15)$ in Setting 2. The variances for η_2^* correspond to the true variances of η_{20} when η_{20} is homoskedastic. For the TM-Heteroskedastic estimator, we let η_2^* be Normal$\{0, (|x|+1)^2/3\}$ in Setting 1 and Normal$\{0, 5(0.4|x|+0.5)^2/3\}$ in Setting 2. The variances for η_2^* correspond to the true variances of η_{20} when η_{20} is heteroskedastic.

4.2.4 Naive estimator

The last method is the naive least squares estimator which is the solution to

$$\arg\max_\beta \sum_{i=1}^{n} \{y_i - m(w_i, z_i; \beta)\}^2.$$

The naive estimator ignores measurement error and falsely assumes X_i and $W_i = (W_{i1} + W_{i2})/2$ are the same.

4.3 Simulation results

4.3.1 Performance of methods compared

Results in Tables 1 and 2 show the bias, estimated variance, and estimated 95% coverage probabilities for the model parameter estimates based on all six methods. Overall, all esti-

mators consistently estimated the measurement error variance σ_U^2 and β_3 associated with the non-mismeasured covariate Z. Performances differed, however, for parameters β_1, β_2 which were affected by the mismeasured covariate X.

In general, compared to the other estimators, our estimator had smaller bias, estimated variances better matching the sample variances, and estimated coverage probabilities closer to the nominal 95% level. This performance was similar regardless of the true model error distribution and its variance structure, thus reflecting the proposed estimator's flexibility. The proposed estimator can yield valid estimates for an RMM with measurement error regardless of whether the true model error is homoskedastic or heteroskedastic. This is especially beneficial in practice since knowing the correct model error variance structure is almost impossible as residuals are not obtainable in measurement error models.

In comparison, the homoskedastic and heteroskedastic sieve MLE were, in some cases, sensitive to the model error's variance structure. When the model error was homoskedastic, the homoskedastic sieve MLE performed well and yielded unbiased estimates. Unfortunately, when applied to the heteroskedastic model error, this same estimator yielded biased estimates with bias up to 19 times larger than our proposed estimator. Increasing the number of smoothing parameters did not change the numerical results (a similar phenomenon was observed in Schennach and Hu (2013)), and it breaks the constrained optimization solver when the number becomes too large. The observed bias was expected, however, because the homoskedastic sieve MLE is not designed to handle heteroskedasticity. Instead, a more flexible sieve such as the heteroskedastic sieve estimator should actually be employed. Unfortunately, in our numerical studies, the heteroskedastic sieve MLE yielded biased estimates both when the model error was homoskedastic and heteroskedastic. We suspect the observed bias could be a result of the difficulty in solving a constrained optimization subject to too many constraints. When the model error is truly heteroskedastic, we further suspect that more specialized bases may be needed to properly account for the heteroskedasticity. Doing so, however, may be difficult as it would require estimating the model error's heteroskedasticity and defining a truncated series that can capture its form. For an RMM with measurement error, correctly determining the model error's variance-covariance is challenging, and is a step surpassed by our proposed estimator.

The TM-Homoskedastic and TM-Heteroskedastic estimators also heavily relied on the correctness of the model error variance. When the model error variance structure was correctly specified, the TM estimators had little bias and nearly perfect nominal 95% coverage probabilities. In this case, the TM estimator has one less nonparametric term than our proposed method, and thus performed well. In contrast, when the variance structure was incorrect, the TM estimators performed poorly compared to our proposed estimator. The poor performance was most notable when the data was generated with heteroskedastic model errors, and we applied the TM-Homoskedastic estimator. In this case, the TM-Homoskedastic estimator yielded estimates with bias up to 40 times larger than our proposed estimator.

Finally, the naive estimator had large bias and coverage probabilities less than the nominal 95%, indicating that the measurement error was significant enough and could not be ignored.

These results demonstrate that measurement error cannot be ignored and that methods that rely on knowing the model error variance structure will, unfortunately, yield biased estimates. Because our proposed estimator makes no assumptions about the model error's variance structure, our method does indicate more flexibility than existing methods, including the sieve MLE and Tsiatis-Ma method. Specifically, our proposed estimator provides consistent estimates even when the model error and covariate distributions are both misspecified. Similar results were observed for other mean models; see Supplementary Material (Section S.9).

4.3.2 Empirical impact of working models in proposed method

In Section 3.3, we discussed the theoretical impact of working models in our proposed method. We now evaluate the numerical impact. Specifically, we generated data as in Section 4.1, except with η_{10} as Normal$(0, 0.5^2)$ and η_{20} as Normal$(0, 0.4^2)$. We then evaluated our proposed method for four different cases of working models η_1^*, η_2^*:

Case 1: $\eta_1^* = \eta_{10}, \eta_2^* = \eta_{20}$.

Case 2: $\eta_1^* \neq \eta_{10}, \eta_2^* = \eta_{20}$ with η_1^* a t-distribution with 4 degrees of freedom.

Case 3: $\eta_1^* = \eta_{10}, \eta_2^* \neq \eta_{20}$ with η_2^* as Normal$\{0, (1 + |X|)^2/3^2\}$.

Case 4: $\eta_1^* \neq \eta_{10}, \eta_2^* \neq \eta_{20}$ with η_1^* a t-distribution with 4 degrees of freedom, and η_2^* as Normal$\{0, (1+|X|)^2/3^2\}$.

Results in Table 3 show that in all cases, the proposed estimator yields consistent estimates. As we progress from Case 2 to Case 4, the efficiency loss only slightly increases; for example, the estimated variance for $\widehat{\beta}_1$ is 0.0065 in Case 4 compared to an estimated variance of 0.0044 in Case 1. Similar results were observed for other regression models; see Supplementary Material (Section S.9). This small loss in efficiency and insensitivity to the choice of the working models was similarly observed in simpler models (see Tsiatis and Ma, 2004, Ma and Carroll, 2006 and Wang et al., 2009). Hence, for flexible choices of working models, our method yields consistent estimates and small efficiency loss when using incorrect working models.

5 A case study

Flagg et al. (2000) performed a study to evaluate the validity of a Nutrition Survey conducted by the American Cancer Society in 1992-1993. In the study, $n = 317$ male participants completed four 24 hour dietary recall interviews given over a one-year period. Interest lies in understanding the impact of saturated fat intake on percent calories from fat for different races (white vs. non-white). Saturated fat intake, however, is not known exactly and only a mismeasured version via two repeated measurements is available.

Let Y denote the percent calories from fat, X denote the log transformation of the true (unobserved) saturated fat intake, and Z denote race ($Z = 1$ refers to white). We let W_1 and W_2 be the centered, log-transformed saturated fat measurements. Through a QQ-plot in Figure 1, we find that $V = (W_1 - W_2)/2$ is acceptably normally distributed with some unknown variance σ_U^2. Normality was formally evaluated through a Pearson Chi-squared test where we used 10 to 20 bins for testing and obtained a p-value at least 0.63, thus assuring the normality assumption.

Because nutrition models usually assume percent calories from fat is related to saturated

fat intake through a linear regression, we use the model

$$Y_i = \beta_1 \exp(X_i) + \beta_2 + \beta_3 Z_i + \epsilon, \quad W_{ij} = X_i + U_{ij}, \quad U_{ij} \sim \text{Normal}(0, 2\sigma_U^2)$$

for $i = 1, \ldots, n; j = 1, 2$ and $E(\epsilon|X, Z) = 0$.

To estimate the model parameters, we used five methods: (i) The proposed method with working models η_1^* as Normal$(0, 0.56^2)$ and η_2^* as Normal$(0, 0.91^2)$. The variance for η_1^* is the sample variance of W, and the variance of η_2^* is the residual sum of squares after regressing Y on $\exp(W)$ and Z. (ii) The homoskedastic sieve MLE with smoothing parameters $\kappa_\epsilon = \kappa_x = 6$. (iii) The heteroskedastic sieve MLE with $\ell_x = \ell_e = \max_i |Y_i|$. (iv) The Tsiatis-Ma Homoskedastic estimator with η_1^*, η_2^* as in our proposed method. Unlike our method, the TM-Homoskedastic estimator assumes the specified η_2^* is correct. We did not use the TM-Heteroskedastic estimator because it is difficult to specify a heteroskedastic variance structure for an RMM with measurement error. (v) The naive estimator.

Parameter estimates for all methods are in Table 4. All methods yielded similar inference conclusions: among the male population, saturated fat intake is statistically significant in relation to percent calories from fat (e.g., proposed method yielded $\widehat{\beta}_1 = 1.59$, 95% CI: (1.23, 1.95)), whereas race is not (e.g., proposed method yielded $\widehat{\beta}_3 = -0.14$, 95% CI: (-0.35, 0.08)). Though inference conclusions were similar, the methods yielded different magnitudes of the parameter effects. For example, the proposed method indicated that a one unit increase in saturated fat is associated with an estimated increase of 1.59 units in the mean of percent calories. This is nearly twice as large as the naive estimates would conclude and at least 1.4 times as large as the homoskedastic sieve, heteroskedastic sieve or TM-Homoskedastic estimator would conclude. The contrast in these results indicate that measurement error cannot be ignored. Moreover, given that the Tsiatis-Ma and sieve MLE estimator exhibited sensitivity to misspecification of the model error variance, we would prefer to rely on the results from the proposed method which is insensitive to such misspecification. Therefore, our method indicates that saturated fat intake affects a male's percent calories from fat more than existing methods would indicate.

6 Discussion

We have developed root-n consistent estimators and provided inference tools for an RMM with errors in covariates where both the mean model and the measurement error model are in their general form. We showed that our method's consistency does not require independence between the covariates and the model error, nor require estimating the unobservable covariate distribution and model error distribution. This is advantageous over existing methods including the Tsiatis and Ma (2004) estimator and the sieve MLE which have shown numerical sensitivity to model error heteroskedasticity. The proposed estimator is derived via a semiparametric procedure different from that in Tsiatis and Ma (2004), and, to the best of our knowledge, the resulting root-n consistent estimator is the first known in its generality that is robust to various distribution misspecifications.

To identify and estimate Ω_U in Section 2.1, we used the average of repeated measures. An alternative is to directly use the repeated measures to perform estimation and inference. Based on our experience (Ma and Yin, 2008), there is generally not a definitive efficiency gain or loss with this approach relative to the average approach. However, more careful analysis will be needed to determine when one or the other is more efficient.

We assumed throughout that the measurement error distribution $p_{U_{ij}}(u;\Omega_U)$ is parametric with Ω_U unknown. We can relax this assumption to have a nonparametric measurement error distribution. In this case, still assuming X_i and U_{ij} are independent, Kotlarski's Theorem (Kotlarski, 1967) implies that the measurement error density is identifiable. From the repeated measures, a nonparametric kernel estimation of the measurement error density function $\widehat{p}_{U_{ij}}$ can be obtained, and operationally our estimation procedure can proceed with $p_{U_{ij}}$ replaced by $\widehat{p}_{U_{ij}}$. For such a plug-in procedure, we provide the following summary. (i) The identifiability of β (Section 2.2) still holds; (ii) Theorem 1 and the estimation procedure for β (Section 3.2) remain valid since they only required a consistent estimator for the measurement error density. (iii) The root-n consistency and asymptotic normality in Theorems 2 and 3 still hold, although the asymptotic variance will change and the proofs will need to be redone to take into account the additional nonparametric estimation. See Hall and Ma (2007) for details on how to incorporate a nonparametrically estimated error distribution in

a different model. (iv) The optimal efficiency bound in estimating β will decrease due to the the nonparametric estimation of $p_{U_{ij}}$.

Lastly, another extension of our method is to a conditional moment model where

$$E\{m(Y, X, Z; \beta)|X, Z\} = 0. \tag{11}$$

In this case, the proof of identification of β (Section 2.2) still holds since it does not require a particular form of the conditional density of Y conditional on X, Z. Our remaining estimation procedure, asymptotic properties, and implementation (Section 3) also remain intact except with ϵ replaced everywhere by $m(Y, X, Z; \beta)$ and $m'_\beta(X, Z; \beta)$ in the right hand side of equation (7) changed to $-E\{\partial m(Y, X, Z; \beta)/\partial\beta|X, Z\}$. To this end, even for general nonlinear and nonseparable regression models of the form $Y = f(X, Z, \epsilon, \beta_0)$, where the distribution of ϵ is unknown and may be subject to various restrictions, as long as we can construct moment conditions, i.e. finding $m(Y, X, Z; \beta_0)$ such that (11) holds, our general procedure is applicable. This extension is particularly useful in empirical economics where models can take a conditional or nonseparable forms.

Acknowledgments

This work was supported by the the National Institute Of Neurological Disorders And Stroke of the National Institutes of Health under Award Number K01NS099343, the Huntington's Disease Society of America Human Biology Project Fellowship, Texas A&M School of Public Health Research Enhancement and Development Initiative (REDI-23-202059-36000), and the National Science Foundation (DMS-1608540). We thank Yingyao Hu for providing code for the homoskedastic sieve estimator and for advising on the heteroskedastic sieve estimator. We thank Raymond J. Carroll for providing the nutrition data. We also thank the editor and two referees whose comments substantially improved the quality and presentation of the work.

Table 1: Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}^2_U, \widehat{\beta}^T)^T$ based on our proposed method (Semipar), homoskedastic sieve MLE (Sieve-Hom), heteroskedastic sieve MLE (Sieve-Het), Tsiatis-Ma homoskedastic estimator (TM-Hom), Tsiatis-Ma heteroskedastic estimator (TM-Het), and the naive estimator. Results based on 1000 simulations when $m(X, Z; \beta) = \beta_2 \exp(-\beta_1 X^2) + \beta_3 Z$, and true parameter values $(\sigma^2_{U,0}, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$.

	Setting 1: $\eta_{20} \sim$Uniform				Setting 2: $\eta_{20} \sim t_5$			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
			η_{20}: Homoskedastic					
			Semipar					
bias	-0.0065	-0.0080	0.0011	5.1664×10^{-5}	-0.0056	-0.0059	0.0008	-9.2372×10^{-5}
var	0.0030	0.0031	0.0026	1.0255×10^{-5}	0.0024	0.0025	0.0022	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0030	0.0030	0.0027	1.0255×10^{-5}	0.0024	0.0024	0.0021	9.9839×10^{-6}
CI	0.9500	0.9390	0.9520	0.9490	0.9440	0.9370	0.9520	0.9320
			Sieve-Hom*					
bias	0.0066	0.0008	0.0021	5.1664×10^{-5}	0.0371	0.0235	0.0056	-9.2372×10^{-5}
var	0.0033	0.0030	0.0022	1.0255×10^{-5}	0.0046	0.0035	0.0023	1.0823×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
			Sieve-Het*					
bias	0.5022	0.8177	0.6900	9.6823×10^{-6}	0.7261	-0.1768	0.3083	-5.0634×10^{-5}
var	0.0450	0.0458	0.0795	1.0916×10^{-5}	0.0521	0.0581	0.0497	1.0109×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
			TM-Hom					
bias	0.0019	0.0019	-0.0000	-0.0002	0.0012	0.0004	-0.0013	-7.9103×10^{-5}
var	0.0035	0.0033	0.0027	1.0396×10^{-5}	0.0028	0.0026	0.0021	9.6008×10^{-6}
$\widehat{\text{var}}$	0.0035	0.0032	0.0027	9.8539×10^{-6}	0.0028	0.0026	0.0022	9.941×10^{-6}
CI	0.9460	0.9440	0.9470	0.9490	0.9450	0.9540	0.9610	0.9470
			TM-Het					
bias	-0.0144	-0.0185	0.0001	-0.0002	-0.0203	-0.0234	-0.0013	-7.9103×10^{-5}
var	0.0038	0.0034	0.0032	1.0396×10^{-5}	0.0026	0.0024	0.0023	9.6008×10^{-6}
$\widehat{\text{var}}$	0.0037	0.0032	0.0031	9.8539×10^{-6}	0.0026	0.0023	0.0023	9.941×10^{-6}
CI	0.9210	0.9300	0.9480	0.9490	0.9080	0.9160	0.9530	0.9470
			Naive					
bias	-0.0269	-0.0230	0.0027	5.1664×10^{-5}	-0.0255	-0.0206	0.0023	-9.2372×10^{-5}
var	0.0029	0.0030	0.0026	1.0255×10^{-5}	0.0023	0.0024	0.0022	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0030	0.0028	0.0027	1.0255×10^{-5}	0.0024	0.0023	0.0022	9.9839×10^{-6}
CI	0.8930	0.9130	0.9570	0.9490	0.8830	0.9190	0.9530	0.9320

*Estimated variances not available. The homoskedastic sieve MLE uses smoothing parameters $\kappa_\epsilon = \kappa_x = 6$, except for the uniform heteroskedastic setting which uses $\kappa_\epsilon = 5, \kappa_x = 6$. For the uniform heteroskedastic setting, the constrained optimization could not be solved for larger κ_ϵ values.

Table 2: Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}^2_U, \widehat{\beta}^T)^T$ based on our proposed method (Semipar), homoskedastic sieve MLE (Sieve-Hom), heteroskedastic sieve MLE (Sieve-Het), Tsiatis-Ma homoskedastic estimator (TM-Hom), Tsiatis-Ma heteroskedastic estimator (TM-Het), and the naive estimator. Results based on 1000 simulations when $m(X, Z; \beta) = \beta_2 \exp(-\beta_1 X^2) + \beta_3 Z$, and true parameter values $(\sigma^2_{U,0}, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$.

	Setting 1: $\eta_{20} \sim$Uniform				Setting 2: $\eta_{20} \sim t_5$			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
			η_{20}: Heteroskedastic					
				Semipar				
bias	0.0098	-0.0055	0.0013	2.9761×10^{-5}	0.0122	-0.0009	0.0008	-0.0001
var	0.0188	0.0101	0.0119	1.0161×10^{-5}	0.0160	0.0103	0.0124	1.0843×10^{-5}
$\widehat{\text{var}}$	0.0206	0.0100	0.0125	1.0022×10^{-5}	0.0195	0.0102	0.0124	9.9708×10^{-5}
CI	0.9610	0.9480	0.9550	0.9510	0.9510	0.9570	0.9520	0.9320
				Sieve-Hom*				
bias	0.1683	0.1008	-0.0757	-2.4913×10^{-5}	0.2437	0.1595	-0.0139	-0.0001
var	0.1686	0.2500	0.0518	9.9631×10^{-6}	0.0601	0.0410	0.0247	1.0492×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				Sieve-Het*				
bias	0.7334	0.6731	0.6527	9.6823×10^{-6}	0.7868	-0.2997	0.5669	-5.0634×10^{-5}
var	0.0423	0.0385	0.0589	1.0916×10^{-5}	0.0443	0.0844	0.0256	1.0109×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				TM-Hom				
bias	0.2931	0.2677	-0.0227	-0.0002	0.5032	0.5174	-0.0349	-0.0005
var	0.2640	0.1392	0.0123	1.037×10^{-5}	1.3260	1.6331	0.0119	8.9453×10^{-6}
$\widehat{\text{var}}$	0.1065	0.0692	0.0131	9.8513×10^{-6}	0.3768	0.2236	0.0135	9.7925×10^{-6}
CI	0.8110	0.6800	0.9580	0.9480	0.8950	0.7890	0.9600	0.9550
				TM-Het				
bias	0.0225	0.0100	0.0005	-0.0002	0.0134	0.0011	-0.0027	-9.3324×10^{-5}
var	0.0218	0.0096	0.0103	1.0409×10^{-5}	0.0195	0.0088	0.0096	9.5564×10^{-6}
$\widehat{\text{var}}$	0.0258	0.0098	0.0104	9.8583×10^{-6}	0.0215	0.0089	0.0099	9.9345×10^{-6}
CI	0.9500	0.9500	0.9540	0.9490	0.9520	0.9500	0.9590	0.9470
				Naive				
bias	0.0692	-0.0185	0.0033	2.9761×10^{-5}	0.0149	-0.0123	0.0028	-0.0001
var	3.0261	0.0102	0.0125	1.0161×10^{-5}	0.3106	0.0100	0.0127	1.0843×10^{-5}
$\widehat{\text{var}}$	1.5445	0.0102	0.0177	1.0022×10^{-5}	0.9245	0.0387	0.0792	9.9708×10^{-5}
CI	0.9390	0.9480	0.9550	0.9510	0.9310	0.9550	0.9550	0.9320

*Estimated variances not available. The homoskedastic sieve MLE uses smoothing parameters $\kappa_\epsilon = \kappa_x = 6$, except for the uniform heteroskedastic setting which uses $\kappa_\epsilon = 5, \kappa_x = 6$. For the uniform heteroskedastic setting, the constrained optimization could not be solved for larger κ_ϵ values.

Table 3: Evaluation of efficiency loss from proposed method when working models η_1^*, η_2^* may differ from the true η_{10}, η_{20}. Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}_U^2, \widehat{\beta}^T)^T$ with true parameter values $(\sigma_{U,0}^2, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$ and $m(X, Z; \beta) = \beta_2 \exp(-\beta_1 X^2) + \beta_3 Z$. Results based on 1000 simulations.

Setting		$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}_U^2$
$\eta_1^* = \eta_{10}, \eta_2^* = \eta_{20}$	bias	-0.0013	0.0009	-0.0017	-3.9892×10^{-5}
	var	0.0043	0.0010	0.0013	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0044	0.0010	0.0013	9.9257×10^{-6}
	CI	0.9500	0.9500	0.9430	0.9430
$\eta_1^* \neq \eta_{10}, \eta_2^* = \eta_{20}$	bias	-0.0001	0.0012	-0.0017	-3.9892×10^{-5}
	var	0.0046	0.0010	0.0013	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0047	0.0010	0.0013	9.9257×10^{-6}
	CI	0.9480	0.9500	0.9430	0.9430
$\eta_1^* = \eta_{10}, \eta_2^* \neq \eta_{20}$	bias	-0.0104	0.0007	-0.0063	-3.9892×10^{-5}
	var	0.0051	0.0012	0.0020	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0052	0.0012	0.0018	9.9257×10^{-6}
	CI	0.9380	0.9490	0.9410	0.9430
$\eta_1^* \neq \eta_{10}, \eta_2^* \neq \eta_{20}$	bias	-0.0081	0.0011	-0.0064	-3.9892×10^{-5}
	var	0.0064	0.0013	0.0021	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0065	0.0013	0.0019	9.9257×10^{-6}
	CI	0.9410	0.9470	0.9430	0.9430

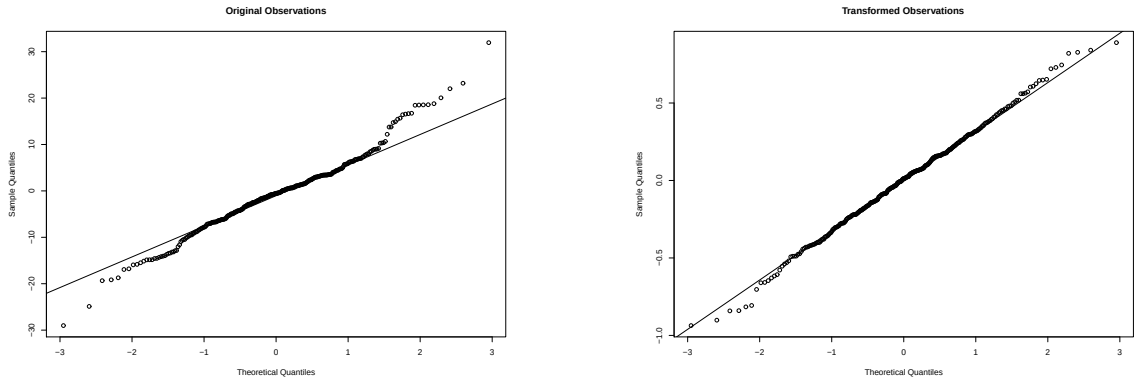

Figure 1: Nutrition study: quantile-quantile plots of the measurement error for the original first and third readings of the 24 hour recall surveys (top) and after the logarithm transform (bottom).

Table 4: Results from nutrition study when estimation is based on proposed method (Semipar), homoskedastic sieve MLE (Sieve-Hom), heteroskedastic sieve MLE (Sieve-Het), Tsiatis-Ma homoskedastic estimator (TM-Hom), and naive estimator. Parameter estimate (est), its estimated variance ($\widehat{\text{var}}$), and 95% confidence interval (CI).

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
		Semipar		
est	1.5926	-1.6611	-0.1364	0.1097
$\widehat{\text{var}}$	0.0338	0.0544	0.0117	0.0001
CI	(1.2322,1.9530)	(-2.1182, -1.2040)	(-0.3480, 0.0752)	(0.0923, 0.1271)
		Sieve-Hom*		
est	1.1397	-1.2449	-0.0401	0.1097
$\widehat{\text{var}}$	NA	NA	NA	NA
CI	NA	NA	NA	NA
		Sieve-Het*		
est	1.1407	1.1628	0.2398	0.1097
$\widehat{\text{var}}$	NA	NA	NA	NA
CI	NA	NA	NA	NA
		TM-Hom†		
est	1.2745	-1.3604	-0.0734	0.1097
$\widehat{\text{var}}$	0.0190	0.0242	0.0116	0.0001
CI	(1.0046, 1.5445)	(-1.6655, -1.0553)	(-0.2841, 0.1373)	(0.0923, 0.1271)
		Naive		
est	0.7110	-0.8044	-0.0351	0.1097
$\widehat{\text{var}}$	0.0065	0.0129	0.0104	0.0001
CI	(0.3506, 1.0714)	(-1.2615, -0.3473)	(-0.2467, 0.1766)	(0.0923, 0.1271)

*Estimated variances not available. The homoskedastic sieve MLE uses smoothing parameters $\kappa_\epsilon = \kappa_x = 6$. †We did not use the TM-Heteroskedastic estimator because it is difficult to specify a heteroskedastic variance structure for an RMM with measurement error.

References

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

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvoluting a density. *Journal of the American Statistical Association*, **83**, 1184-1186.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541-554.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D., Karagas, M. R. (2004). Nonlinear and nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, **99**, 736-750.

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

Carroll, R. J. and Wang, Y. (2008). Nonparametric variance estimation in the analysis of microarray data: a measurement error approach. *Biometrika*, **95**, 437-449.

Chan, L. K. and Mak, T. K. (1985). On the polynomial functional relationship. *Journal of the Royal Statistical Society, Series B*, **47**, 510-518.

Chen, X., Hu, Y. and Lewbel, A. (2009). Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information. *Statistica Sinica*, **19**, 949-968.

Cheng, C. L. and Schneeweiss, H. (1998). Polynomial regression with errors in the variables. *Journal of the Royal Statistical Society, Series B*, **60**, 189-199.

Cheng, C. L., Schneeweiss, H. and Thamerus, M. (2000). A small sample estimator for a polynomial regression with errors in the variables. *Journal of the Royal Statistical Society, Series B*, **62**, 699-709.

Delaigle, A. and Hall, P. (2011). Estimation of observation-error variance in errors-in-variables regression. *Statistica Sinica*, **21**, 1023-1063.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, **19**, 1257-1272.

Flagg, E., Coates, R., Calle, E., Potischman, N., and Thun, M. (2000). Validation of the American Cancer Society Cancer Prevention Study II Nutrition Survey Cohort Food Frequency Questionnaire. *Epidemiology*, **11**, 462-468.

Fuller, W. A. (1987). Measurement Error Models. New York: Wiley.

Hall, P. and Ma, Y. (2007). Measurement Error Models with Unknown Error Structure. *Journal of the Royal Statistical Society, Series B*, **69**, 429-446.

Huang, S. and Huwang, L. (2001). On the polynomial structural relationship. *The Canadian Journal of Statistics*, **29**, 495-512.

Hu, Y. and Schennach, S. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, **76**, 195-216.

Jennrich, R. I. (1969): Asymptotic Properties of Non-Linear Least Squares Estimators. *Annals of Mathematical Statistics*,**40**, 633-643.

Kotlarski, I. I. (1967). On characterizing the gamma and normal distribution. *Pacific Journal of Mathematics*, **20**, 69-76.

Kress, R. (1999). Linear integral equations (2nd edition). Berlin: Springer.

Lee, L. and Sepanski, J. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, **90**, 130-140.

Li, H. and Liqun, W. (2012). Consistent estimation in generalized linear mixed models with measurement error. *Journal of Biometrics and Biostatistics*, S7:007. doi:10.4172/2155-6180.S7-007.

Li, T. (2002). Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics*, **110**, 1-26.

Liang, H. (2009). Generalized partially linear mixed effects models incorporating mismeasured covariates. *Annals of the Institute of Statistical Mathematics*, **61**, 27-46.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

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. and Yin, G. (2008) Cure Rate Model with Mismeasured Covariates under Transformation *Journal of the American Statistical Association*, **103**, 743-756.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: methodology and application to generalized linear models. *Biometrika*, **77**, 127-137.

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

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.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. New York: Wiley.

Rao, P. (1992). Identifiability in Stochastic Models. New York: Academic Press.

Rudin, W. (1987). Real and complex analysis. Mathematics series. McGraw-Hill.

Schennach, S. (2004). Estimation of nonlinear models with measurement error. *Econometrica*, **72**, 33-75.

Schennach, S. (2004b). Nonparametric regression in the presence of measurement error. *Econometric Theory*, **20**, 1046-1093.

Schennach, S. and Hu, Y. (2013). Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, **108**, 177-186.

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics*, **25**, 2555-2591.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169-184.

Tsiatis, A. (2006). Semiparametric Theory and Missing Data. New York: Springer.

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

Wang, Y., Ma, Y., and Carroll, R. J. (2009). Variance estimation in the analysis of microarray data. *Journal of the Royal Statistical Society, Series B*, **71**, 725-745.

Supplement: Sketch of technical arguments and additional simulation results

S.1 Overview of semiparametric theory

For independent, identically distributed (i.i.d.) data O_i, an estimator $\widehat{\beta}_n$ for $\beta \in \mathbb{R}^p$ is regular, asymptotically linear (RAL) with influence function φ if

$$n^{1/2}(\widehat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(O_i; \beta_0) + o_P(1).$$

Here, β_0 denotes the truth; $\varphi(O_i; \beta_0)$ are i.i.d., mean zero random vectors of length p; and $o_P(1)$ converges in probability to zero as n tends to infinity. The asymptotic variance of $n^{1/2}\widehat{\beta}_n$ equals that of φ, so determining the most efficient RAL estimator is equivalent to identifying the influence function with smallest variance (i.e., efficient influence function denoted as φ_{eff}).

To identify the efficient influence function, a geometric approach is taken. First, we consider a Hilbert space $\mathcal{H}$ composed of mean zero, finite variance, p-dimensional functions. For our purposes, there are two subspaces of $\mathcal{H}$ that are of key interest. The first is the so-called nuisance tangent space Λ. This subspace is the mean-squared closure of elements of the form BS, where S is an arbitrary nuisance score vector and B is a conformable matrix with p rows. The second key subspace is the orthogonal complement of Λ, denoted as $\Lambda^{\perp}$. All the influence functions lie in $\Lambda^{\perp}$, including the efficient one φ_{eff}. To calculate φ_{eff}, one first computes the efficient score vector, $S_{\text{eff}}(O)$. The efficient score vector is defined as the orthogonal projection of the score vector $S_\beta(O)$ onto $\Lambda^{\perp}$, or, equivalently, is the residual of $S_\beta(O)$ after projecting it onto Λ. That is, $S_{\text{eff}} = \Pi(S_\beta|\Lambda^{\perp}) = S_\beta - \Pi(S_\beta|\Lambda)$, where $\Pi(\cdot|\cdot)$ denotes orthogonal projection. Then, the efficient influence function is the normalized version of S_{eff}: $\varphi_{\text{eff}} = [E\{S_{\text{eff}}(O)S_{\text{eff}}^T(O)\}]^{-1}S_{\text{eff}}(O)$ where the normalization ensures that $E(\varphi_{\text{eff}}S_\beta^T) = I_p$, the $p \times p$ identity matrix.

Thus, the general approach involves (1) specifying the Hilbert space $\mathcal{H}$ based on the probability distribution that generates the data; (2) characterizing the nuisance tangent

space Λ and its orthogonal complement $\Lambda^{\perp}$; and (3) orthogonally projecting the score vector S_β onto $\Lambda^{\perp}$ and normalizing the result. The efficient influence function can then be used to construct an estimating equation from which consistent estimators for β can be obtained.

S.1.1 Semiparametric results for the RMM without measurement error

For the RMM without measurement error, we observe data (Y, X, Z) which has probability density $p_{Y,X,Z}(Y, X, Z) = \eta_1(x, z)\eta_2\{y - m(x, z; \beta), x, z\}\eta_3(z)$ such that $\int \epsilon\eta_2(\epsilon, x, z)d\epsilon = 0$ for all x, z. Interest lies in estimating the p-dimensional parameter β in the presence of the unknown distributions η_1, η_2, and η_3. Applying semiparametric theory, the nuisance tangent space, its orthogonal complement and the efficient influence function are given in the proposition below. For a detailed derivation, see Tsiatis (2006, chap. 4).

Proposition 1 *For the RMM, the Hilbert space is* $\mathcal{H}^F = \{h(Y, X, Z) : E(h) = 0, var(h) < \infty\}$. *The nuisance tangent space is given by* $\Lambda^F = \Lambda_1^F \oplus \Lambda_2^F \oplus \Lambda_3^F$, *where* $\Lambda_1^F = [h_1(X, Z) : E\{h_1(X, Z)|Z\} = 0]$, $\Lambda_2^F = [h_2(Y, X, Z) : E\{h_2(Y, X, Z)|X, Z\} = 0, E\{\epsilon h_2(Y, X, Z)|X, Z\} = 0]$, *and* $\Lambda_3^F = [h_3(Z) : E\{h_3(Z)\} = 0]$ *(Newey, 1990; Tsiatis, 2006). Equivalently, the nuisance space can be written as*

$$\Lambda^F = \{h(Y, X, Z) : E(h\epsilon|X, Z) = 0, E(h) = 0, var(h) < \infty\}.$$

Consequently, the orthogonal complement to the nuisance tangent space is

$$\Lambda^{\perp F} = \{g(X, Z)\epsilon\} = [g(X, Z)\{Y - m(X, Z; \beta)\}].$$

Here, $g(X, Z)$ *is an arbitrary function such that* $E(g^T g) < \infty$. *The score vector with respect to* β *is* $S_\beta^F = -m'_\beta(X, Z; \beta_0)\partial log\eta_2(\epsilon, X, Z)/\partial\epsilon$ *where* $m'_\beta(X, Z; \beta_0)$ *denotes* $\partial m(X, Z; \beta_0)/\partial\beta$ *and* β_0 *denotes the truth. The efficient score vector is* $S_{\text{eff}}^F = m'_\beta(X, Z; \beta_0)E(\epsilon^2|X, Z)^{-1}\epsilon$ *and after normalization, the efficient influence function is* $\varphi_{\text{eff}}^F = E\{E(\epsilon^2|X, Z)^{-1}m'_\beta m'^T_\beta\}m'_\beta E(\epsilon^2|X, Z)^{-1}\epsilon$.

S.2 Proof of Theorem 1

We use the semiparametric approach of Section S.1 (Supplementary Material) to construct the set $\Lambda^{\perp}$. In the terminology of semiparametric theory, $\Lambda^{\perp}$ is the orthogonal complement to the nuisance tangent space and contains all consistent, semiparametric estimators for β.

Following the steps in Section S.1 (Supplementary Material), the Hilbert space is $\mathcal{H} = \{f(Y,W,Z) : E(f) = 0, \text{var}(f) < \infty\}$. To derive the nuisance tangent space, we first observe that the the nuisance score vectors for the RMM with measurement error and without are closely intertwined. Let $S^F_{\eta_i}(Y,X,Z)$ denote the nuisance score vectors for the RMM without measurement error, and let $S_{\eta_i}(Y,W,Z)$ denote the nuisance score vectors for the RMM with measurement error, $i = 1,2,3$. Then, based on the model in (4), we have that $S_{\eta_i}(Y,W,Z) = E\{S^F_{\eta_i}(Y,X,Z)|Y,W,Z\}$. Given this relationship between the nuisance score vectors for the RMM with and without measurement error, we have, by a result in Rao (1973, p. 330), that the nuisance tangent space for the RMM with measurement error is

$$\begin{aligned}\Lambda &= E(\Lambda^F|Y,W,Z) \\ &= [E\{h(Y,X,Z)|Y,W,Z\} : E(h\epsilon|X,Z) = 0, E(h) = 0, \text{var}(h) < \infty].\end{aligned}$$

To verify that $\Lambda^{\perp}$ in Theorem 1 is indeed the orthogonal complement to Λ, consider any function $f(Y,W,Z)$ such that $E\{f(Y,W,Z)|Y,X,Z\} = g(X,Z)\epsilon$ for some $g(X,Z)$. Then, the inner product of f and an arbitrary element $E\{h(Y,X,Z)|Y,W,Z\} \in \Lambda$ is:

$$\begin{aligned}& E[E\{h^T(Y,X,Z)|Y,W,Z\}f(Y,W,Z)] = E\{h^T(Y,X,Z)f(Y,W,Z)\} \\ = \;& E[h^T(Y,X,Z)E\{f(Y,W,Z)|Y,X,Z\}] = E\{h^T(Y,X,Z)g(X,Z)\epsilon\} = 0.\end{aligned}$$

The last equality holds because $E\{h(Y,X,Z)\epsilon|X,Z\} = 0$ by properties of Λ. Therefore, since f is orthogonal to an arbitrary element in Λ, f must belong to $\Lambda^{\perp}$.

Conversely, we now show that any mean zero function $f(Y,W,Z) \in \Lambda^{\perp}$ must satisfy $E\{f(Y,W,Z)|Y,X,Z\} = g(X,Z)\epsilon$. To see this, let $k(Y,X,Z) = E(f|Y,X,Z)$ and $r(Y,X,Z) = k(Y,X,Z) - g(X,Z)\epsilon$ where the function $g(X,Z) = E(k\epsilon|X,Z)/E(\epsilon^2|X,Z)$. It suffices to show that $r = 0$ since from such a result, we immediately have that

$E\{f(Y,W,Z)|Y,X,Z\} = g(X,Z)\epsilon$.

Observe that $E(r|Y,W,Z)$ is in Λ because $E(r\epsilon|X,Z) = 0$ and $E(r) = 0$ which holds since $E(f) = 0$ and $E(\epsilon|X,Z) = 0$. Therefore, the inner product of $f \in \Lambda^{\perp}$ and $E(r|Y,W,Z) \in \Lambda$ is zero, implying that

$$\begin{aligned} 0 &= E[f^T(Y,W,Z)E\{r(Y,X,Z)|Y,W,Z\}] = E\{f^T(Y,W,Z)r(Y,X,Z)\} \\ &= E[E\{f^T(Y,W,Z)|Y,X,Z\}r(Y,X,Z)] = E\{k^T(Y,X,Z)r(Y,X,Z)\} \\ &= E(r^Tr) + E[g^T(X,Z)E\{\epsilon r(Y,X,Z)|X,Z\}] = E(r^Tr). \end{aligned}$$

The last equality holds because $E(r\epsilon|X,Z) = 0$. Thus, by properties of Hilbert spaces, whenever $E(r^Tr) = 0$, we must have $r = 0$. Therefore, from having $r = 0$, we have shown that any element $f \in \Lambda^{\perp}$ satisfies $E(f|Y,X,Z) = g(X,Z)\epsilon$ for any $g(X,Z)$. Hence, $\Lambda^{\perp}$ is as claimed.

S.3 Construction of an element in $\Lambda^{\perp}$

From semiparametric theory (see Section S.1 in Supplementary Material), an element of $\Lambda^{\perp}$ is the p-dimensional efficient score vector, $S^*_{\text{eff}}(Y,W,Z;\beta_0,\alpha_0,\eta_1^*,\eta_2^*,\eta_3)$. The efficient score vector is the result of projecting the score vector S^*_{β} onto $\Lambda^{\perp}$. In the justification below, we will show that S^*_{eff} in fact equals $\mathcal{K}_1(d^*)$. Thus, because $\mathcal{K}_1(d^*)$ is the result of projecting S^*_{β} onto $\Lambda^{\perp}$, we have that although d^* is not unique, $\mathcal{K}_1(d^*)$ is unique by the projection theorem (Rudin, 1987).

We first establish some preliminary results. First, $\mathcal{K}_1$ and $\mathcal{K}_2$ are conjugate of each other because

$$\begin{aligned} &< \mathcal{K}_2\{f(Y,W,Z)\}, g(Y,X,Z) >= E_*[E\{f^T(Y,W,Z)|Y,X,Z\}g(Y,X,Z)] \\ &= E_*\{f^T(Y,W,Z)g(Y,X,Z)\} = E_*[f^T(Y,W,Z)E_*\{g(Y,X,Z)|Y,W,Z\}] \\ &=< f, \mathcal{K}_1\{g(Y,X,Z)\} > . \end{aligned}$$

Second, by the results in Proposition 1 (Section S.1.1, Supplementary Material) and the

results in Section S.2 (Supplementary Material), we have

$$\Lambda^* = \mathcal{K}_1(\Lambda^{F,*}), \quad \mathcal{K}_2(\Lambda^{\perp}) \subset \Lambda^{\perp F}. \tag{S.1}$$

Throughout, spaces and elements superscripted with ^{F,*} are from the RMM without measurement error when we use working models η_1^*, η_2^* for η_1, η_2, respectively.

We now demonstrate that $\mathcal{K}_1(d^*) = S_{\text{eff}}^*$ where S_{eff}^* is the result of projecting the score vector S_β^* onto $\Lambda^{\perp}$. By Theorem 3.5 in Tsiatis (2006),

$$\begin{aligned} S_{\text{eff}}^{F,*}(Y,X,Z) &= S_\beta^{F,*}(Y,X,Z) - \Pi\{S_\beta^*(Y,X,Z)|\Lambda^{F,*}\}, \\ S_{\text{eff}}^{*}(Y,W,Z) &= S_\beta^{*}(Y,W,Z) - \Pi\{S_\beta^*(Y,W,Z)|\Lambda^{*}\} \end{aligned} \tag{S.2}$$

where $\Pi(\cdot|\cdot)$ denotes projection. Using (S.2), we now proceed to write $S_{\text{eff}}^*(Y,W,Z)$ in a more explicit form.

First, the score vector $S_\beta^* = \partial \log p_{Y,W,Z}(y,w,z;\beta_0,\alpha_0,\eta_1^*,\eta_2^*,\eta_3)/\partial\beta$ is easily shown to satisfy $S_\beta^* = \mathcal{K}_1\{S_\beta^{F,*}(Y,X,Z)\}$. However, by (S.2), $\mathcal{K}_1\{S_\beta^{F,*}(Y,X,Z)\} = \mathcal{K}_1\{S_{\text{eff}}^{F,*} + \Pi(S_\beta^{F,*}|\Lambda^{F,*})\}$. Second, since $\Lambda^* = \mathcal{K}_1(\Lambda^{F,*})$ by (S.1), we have that $\Pi(S_\beta^*|\Lambda^*) = \mathcal{K}_1(a^{F,*})$ for some $a^{F,*}(Y,X,Z)$ in $\Lambda^{F,*}$. Together, these results imply that the efficient score vector satisfies

$$\begin{aligned} S_{\text{eff}}^*(Y,W,Z) &= S_\beta^*(Y,W,Z) - \Pi\{S_\beta^*(Y,W,Z)|\Lambda^*\} \\ &= \mathcal{K}_1\{S_{\text{eff}}^{F,*} + \Pi(S_\beta^{F,*}|\Lambda^{F,*})\} - \mathcal{K}_1(a^{F,*}) = \mathcal{K}_1\{d^*(Y,X,Z)\}, \end{aligned}$$

where $d^*(Y,X,Z) = S_{\text{eff}}^{F,*}(Y,X,Z) - b^{F,*}(Y,X,Z)$ and $b^{F,*} = a^{F,*} - \Pi(S_\beta^{F,*}|\Lambda^{F,*})$. Having expressed $S_{\text{eff}}^* = \mathcal{K}_1(d^*)$, we derive the properties of d^* so that it may be solved explicitly.

Projecting d^* onto $\Lambda^{\perp F}$ gives

$$\Pi(d^*|\Lambda^{\perp F}) = \Pi(S_{\text{eff}}^{F,*} - b^{F,*}|\Lambda^{\perp F}) = S_{\text{eff}}^{F,*}(Y,X,Z). \tag{S.3}$$

This is because $b^{F,*} \in \Lambda^{F,*}$, and $\Lambda^{F,*}$ and $\Lambda^{\perp F}$ are orthogonal, so that $\Pi(b^{F,*}|\Lambda^{\perp F}) = 0$. Second, $\mathcal{K}_2 \circ \mathcal{K}_1(d^*) = \mathcal{K}_2(S_{\text{eff}}^*)$ is an element of $\Lambda^{\perp F}$ since $S_{\text{eff}}^* \in \Lambda^{\perp}$ and $\mathcal{K}_2(\Lambda^{\perp}) \subset \Lambda^{\perp F}$ by

(S.1). Hence, because $\mathcal{K}_2 \circ \mathcal{K}_1(d^*) \in \Lambda^{\perp F}$, it follows that

$$\Pi\{\mathcal{K}_2 \circ \mathcal{K}_1(d^*)|\Lambda^{F,*}\} = 0.$$

Combining (S.3) and the above display demonstrates that there exists a function d^* such that $S^*_{\text{eff}} = \mathcal{K}_1(d^*)$ and

$$\Pi(d^*|\Lambda^{\perp F}) + \Pi\{\mathcal{K}_2 \circ \mathcal{K}_1(d^*)|\Lambda^{F,*}\} = S^{F,*}_{\text{eff}}(Y, X, Z),$$

which simplifies into expression (7).

To complete the demonstration, we show that any d^* satisfying (7) yields S^*_{eff}. To this end, if d^* satisfies (7), then $\Pi(d^*|\Lambda^{\perp F}) = S^{F,*}_{\text{eff}}$ and $\Pi\{\mathcal{K}_2 \circ \mathcal{K}_1(d^*)|\Lambda^{F,*}\} = 0$. Hence,

$$d^*(Y, X, Z) = S^{F,*}_{\text{eff}}(Y, X, Z) + a^{F,*}(Y, X, Z) = S^{F,*}_{\beta}(Y, X, Z) + b^{F,*}(Y, X, Z)$$

for some $a^{F,*}, b^{F,*} \in \Lambda^{F,*}$. Therefore,

$$\mathcal{K}_1(d^*) \;\; = \;\; \mathcal{K}_1(S^{F,*}_{\beta} + b^{F,*}) = S^*_{\beta} + b^*(Y, W, Z) = S^*_{\text{eff}} + a^*(Y, W, Z), \tag{S.4}$$

for some $a^*, b^* \in \Lambda^*$. The second equality above holds because $\mathcal{K}_1(S^{F,*}_{\beta}) = S^*_{\beta}$ and $\mathcal{K}_1(\Lambda^{F,*}) = \Lambda^*$. The third equality holds by (S.2). Up to now, we have shown that $\mathcal{K}_1(d^*) = S^*_{\text{eff}} + a^*(Y, W, Z)$. The argument will be complete once $a^*(Y, W, Z)$ is shown to be exactly zero.

Having d^* satisfy (7) means $\Pi\{\mathcal{K}_2 \circ \mathcal{K}_1(d^*)|\Lambda^{F,*}\} = 0$, and so $\mathcal{K}_2 \circ \mathcal{K}_1(d^*) \in \Lambda^{\perp F}$. From (S.4), this implies that $\mathcal{K}_2 \circ \mathcal{K}_1(d^*) = \mathcal{K}_2(S^*_{\text{eff}}) + \mathcal{K}_2\{a^*(Y, W, Z)\} \in \Lambda^{\perp F}$. The inner product of $\mathcal{K}_2\{a^*(Y, W, Z)\} \in \Lambda^{\perp F}$ and any $b^{F,*} \in \Lambda^{F,*}$ must be zero. Hence,

$$0 =< \mathcal{K}_2\{a^*(Y, W, Z)\}, b^{F,*}(Y, X, Z) >=< a^*(Y, W, Z), \mathcal{K}_1\{b^{F,*}(Y, X, Z)\} >,$$

where the latter equality holds from the conjugacy of $\mathcal{K}_1$ and $\mathcal{K}_2$. But the inner product above implies that $a^* \in \Lambda^{\perp}$ since $\mathcal{K}_1(b^{F,*}) \in \Lambda^*$. Having a^* in Λ^* and in $\Lambda^{\perp}$ implies that $a^* = 0$. Thus, d^* satisfying (7) requires $\mathcal{K}_1(d^*) = S^*_{\text{eff}}$.

Our argument thus shows that an element of $\Lambda^{\perp}$ is $\mathcal{K}_1(d^*) = S^*_{\text{eff}}$ where S^*_{eff} is the result of projecting the score vector S^*_β onto $\Lambda^{\perp}$.

S.4 Consistency of $(\widehat{\alpha}_n^T, \widehat{\beta}_n^T)^T$ even with possibly misspecified η_1^*,η_2^*

Let $\mathcal{S}(Y, W, V, Z; \alpha, \beta, \eta_1^*, \eta_2^*, \eta_3) = \{\phi^T(V; \alpha), S^{*\text{T}}_{\text{eff}}(Y, W, Z; \alpha, \beta, \eta_1^*, \eta_2^*, \eta_3)\}^T$, where ϕ is the estimating equation given (3) and $\alpha = vech(\Omega_U)$, the vectorized form of Ω_U. We prove consistency by first showing that $E(\mathcal{S}) = 0$ (i.e., $E(\phi) = 0$ and $E(S^*_{\text{eff}}) = 0$) where the expectation is computed under the true distribution.

First, the usual components of variance analysis (Carroll et al., 2006) establishes that $E(\phi) = 0$.

Second, we constructed $S^*_{\text{eff}} = \mathcal{K}_1(d^*)$ with d^* satisfying (7). A simple rearrangement of (7) shows that $\mathcal{K}_2 \circ \mathcal{K}_1(d^*) = E\{\mathcal{K}_1(d^*)|Y, X, Z\} = g(X, Z)\epsilon$ where $g(X, Z) = (m'_\beta(X, Z; \beta) - E_*[\{d^* - \mathcal{K}_2 \circ \mathcal{K}_1(d^*)\}\epsilon|X, Z])E_*(\epsilon^2|X, Z)^{-1}$. Hence, because $E(\epsilon|X, Z) = 0$ by assumption, we have that

$$\begin{aligned} E(S^*_{\text{eff}}) &= E\{\mathcal{K}_1(d^*)\} = E[E\{\mathcal{K}_1(d^*)|Y, X, Z\}] = E\{g(X, Z)\epsilon\} \\ &= E\{g(X, Z)E(\epsilon|X, Z)\} = 0. \end{aligned}$$

We have shown that $E\{\mathcal{S}(Y, W, V, Z, \alpha, \beta, \eta_1^*, \eta_2^*, \eta_3)\} = 0$ at the true parameter value α_0, β_0. Under the smoothness condition (*R1*) and unique root condition (*R3*), we have that for any (α_n, β_n) such that $E\{\mathcal{S}(Y, W, V, Z, \alpha_n, \beta_n, \eta_1^*, \eta_2^*, \eta_3)\} \to 0$ in probability, it implies $\alpha_n \to \alpha_0$ and $\beta_n \to \beta_0$ in probability. On the other hand, since $\widehat{\alpha}_n, \widehat{\beta}_n$ solves the estimating equation, we thus have that $n^{-1}\sum_{i=1}^n \mathcal{S}(Y_i, W_i, V_i, Z_i, \widehat{\alpha}_n, \widehat{\beta}_n, \eta_1^*, \eta_2^*, \eta_3) = 0$. Condition (*R3*) and the uniform law of large numbers (Jennrich, 1969) then imply $E\{\mathcal{S}(Y_i, W_i, V_i, Z_i, \widehat{\alpha}_n, \widehat{\beta}_n, \eta_1^*, \eta_2^*, \eta_3)\} = o_p(1)$. This leads to $\widehat{\alpha}_n \to \alpha$ and $\widehat{\beta}_n \to \beta$ in probability, i.e. our estimator is indeed consistent.

S.5 Theoretical Efficiency Loss under Misspecified η_1^*, η_2^*

Using properties of score vectors and semiparametric theory, we show that the theoretical efficiency loss under misspecified η_1^*, η_2^* satisfies

$$A_*^{-1}B_*(A_*^{-1})^T - A^{-1}B(A^{-1})^T = E\{(A_*^{-1}S_{\text{eff}}^* - A^{-1}S_{\text{eff}})(A_*^{-1}S_{\text{eff}}^* - A^{-1}S_{\text{eff}})^T\}.$$

Because

$$\begin{aligned} E\{(A_*^{-1}S_{\text{eff}}^* - A^{-1}S_{\text{eff}})(A_*^{-1}S_{\text{eff}}^* - A^{-1}S_{\text{eff}})^T\} &= A_*^{-1}B_*(A_*^{-1})^T \\ -A_*^{-1}E(S_{\text{eff}}^*S_{\text{eff}}^{*T})(A^{-1})^T - A^{-1}E(S_{\text{eff}}S_{\text{eff}}^{*T})(A_*^{-1})^T &+ A^{-1}B(A^{-1})^T, \end{aligned}$$

it suffices to show that $A = -B$ and $E(S_{\text{eff}}^*S_{\text{eff}}^{*T}) = -A_*$.

We have that $A = -B$ because S_{eff} is a score vector. This simplification further implies that $A^{-1}B(A^{-1})^T = B^{-1}$.

To show $E(S_{\text{eff}}^*S_{\text{eff}}^{*T}) = -A_*$, we use the fact that $S_{\text{eff}} = S_\beta + a$, where S_β is the score vector with respect to β, and a is an element of Λ, the space orthogonal to $\Lambda^\perp$ in Theorem 1 (Tsiatis, 2006, Thm. 3.5). Because we constructed S_{eff}^* to be an element of $\Lambda^\perp$, this means,

$$E(S_{\text{eff}}^*S_{\text{eff}}^{*T}) = E(S_{\text{eff}}^*S_\beta^T) + E(S_{\text{eff}}^*a^T) = E(S_{\text{eff}}^*S_\beta^T).$$

The last equality holds because $S_{\text{eff}}^* \in \Lambda^\perp$ and $a \in \Lambda$. Also, in Section S.4 (Supplementary Material), we showed that $E(S_{\text{eff}}^*) = 0$ which means

$$0 = \int S_{\text{eff}}^*(Y, W, Z; \beta, \alpha_0, \eta_1^*, \eta_2^*)p_{Y,W,Z}(Y, W, Z; \beta, \alpha_0, \eta_{10}, \eta_{20})d\mu(Y, W, Z).$$

Taking derivative with respect to β, we obtain

$$\begin{aligned} 0 &= \int \frac{\partial S_{\text{eff}}^*}{\partial \beta^T}p_{Y,W,Z}(Y, W, Z; \beta, \alpha_0, \eta_{10}, \eta_{20})d\mu(Y, W, Z) \\ &\quad + \int S_{\text{eff}}^*S_\beta^T p_{Y,W,Z}(Y, W, Z; \beta, \alpha_0, \eta_{10}, \eta_{20})d\mu(Y, W, Z) \\ &= A_* + E(S_{\text{eff}}^*S_\beta^T) = A_* + E(S_{\text{eff}}^*S_{\text{eff}}^{*T}). \end{aligned}$$

The last equality holds by the definition of A_* and our earlier argument that $E(S^*_{\text{eff}}S^T_{\text{eff}}) = E(S^*_{\text{eff}}S^T_\beta)$. Hence, we have shown that $E(S^*_{\text{eff}}S^T_{\text{eff}}) = -A_*$.

Therefore, using the above results, we have

$$\begin{aligned}
& E\{(A_*^{-1}S^*_{\text{eff}} - A^{-1}S_{\text{eff}})(A_*^{-1}S^*_{\text{eff}} - A^{-1}S_{\text{eff}})^T\} \\
= \; & A_*^{-1}B_*(A_*^{-1})^T - A_*^{-1}E(S^*_{\text{eff}}S^T_{\text{eff}})(A^{-1})^T - A^{-1}E(S_{\text{eff}}S^{*T}_{\text{eff}})(A_*^{-1})^T \\
& + A^{-1}B(A^{-1})^T \\
= \; & A_*^{-1}B_*(A_*^{-1})^T + A_*^{-1}A_*(A^{-1})^T + A^{-1}A_*^T(A_*^{-1})^T - A^{-1}A(A^{-1})^T \\
= \; & A_*^{-1}B_*(A_*^{-1})^T + (A^{-1})^T + A^{-1} - (A^{-1})^T = A_*^{-1}B_*(A_*^{-1})^T + A^{-1} \\
= \; & A_*^{-1}B_*(A_*^{-1})^T - B^{-1} = A_*^{-1}B_*(A_*^{-1})^T - A^{-1}B(A^{-1})^T.
\end{aligned}$$

S.6 Proof of Theorem 2

We first demonstrate that $E\{\partial\mathcal{S}(Y_i, W_i, V_i, Z_i; \theta_0, \gamma^*)/\partial\gamma^T\} = 0$. First, $E\{\partial\phi(V_1; \alpha_0)/\partial\gamma^T\} = 0$ trivially. Second, our construction of f ensures that $f(Y, W, Z; \theta_0, \gamma) \in \Lambda^\perp$ for all γ. Therefore, at any γ,

$$\int f(Y, W, Z; \theta_0, \gamma) p_{Y,W,Z}(Y, W, Z; \theta_0) d\mu(Y, W, Z) = 0. \tag{S.5}$$

Taking derivative of (S.5) with respect to γ, we further obtain

$$\int \frac{\partial f(Y, W, Z; \theta_0, \gamma)}{\partial\gamma^T} p_{Y,W,Z}(Y, W, Z; \theta_0) d\mu(Y, W, Z) = 0$$

for all γ. Evaluating the above equation at γ^* gives $E\left\{\partial f(Y, W, Z; \theta_0, \gamma^*)/\partial\gamma^T\right\} = 0$. We have thus shown that $E\{\partial\mathcal{S}(Y_i, W_i, V_i, Z_i; \theta_0, \gamma^*)/\partial\gamma^T\} = 0$.

Since $\widehat{\theta}_n$ solves the estimating equation, we have that $n^{-1}\sum_{i=1}^n S(Y_i, W_i, V_i, Z_i, \widehat{\theta}_n, \widehat{\gamma}_n) = 0$. Condition $(R4)$ and the uniform law of large numbers (Jennrich, 1969) imply that $E\{\mathcal{S}(Y_i, W_i, V_i, Z_i, \widehat{\theta}_n, \widehat{\gamma}_n)\} = o_p(1)$, and so $E\{\mathcal{S}(Y_i, W_i, V_i, Z_i, \widehat{\theta}_n, \gamma^*)\} = o_p(1)$ since $\widehat{\gamma}_n$ is defined as the root-n consistent estimator γ^*. On the other hand, assuming a unique root $(R3)$, assuming that $\mathcal{S}(Y, W, V, Z, \theta_n, \gamma^*)$ is a smooth function of θ_n and because

$E\{\mathcal{S}(Y,W,V,Z,\theta_0,\gamma^*)\}=0$, we have that for any θ_n such that $E\{\mathcal{S}(Y,W,V,Z,\theta_n,\gamma^*)\}\to 0$, it implies $\theta_n\to\theta_0$ in probability. This leads to $\widehat{\theta}_n\to\theta_0$ in probability; i.e. the estimator $\widehat{\theta}_n$ is indeed consistent.

Now, consider the Taylor expansion

$$\begin{aligned}
0 &= n^{-1/2}\sum_{i=1}^{n}\mathcal{S}(Y_i,W_i,V_iZ_i;\widehat{\theta}_n,\widehat{\gamma}_n)\\
&= n^{-1/2}\sum_{i=1}^{n}\mathcal{S}(Y_i,W_i,V_i,Z_i;\theta_0,\gamma^*)\\
&\quad +n^{-1}\sum_{i=1}^{n}\frac{\partial\mathcal{S}(Y_i,W_i,V_i,Z_i;\theta_n^\dagger,\gamma_n^\dagger)}{\partial\theta^T}\sqrt{n}(\widehat{\theta}_n-\theta_0)\\
&\quad +n^{-1}\sum_{i=1}^{n}\frac{\partial\mathcal{S}(Y_i,W_i,V_i,Z_i;\theta_n^\dagger,\gamma_n^\dagger)}{\partial\gamma^T}\sqrt{n}(\widehat{\gamma}_n-\gamma^*),
\end{aligned}$$

where $\theta_n^\dagger$ lies on the line connecting $\widehat{\theta}_n$ and θ_0, and $\gamma_n^\dagger$ lies on the line connecting $\widehat{\gamma}_n$ and γ^*. Condition $(R4)$ and the central limit theorem ensure that the first term after the second equality converges in distribution to a zero mean finite variance normal random variable, hence it is of order at most $O_p(1)$. The root-n consistency of $\widehat{\gamma}_n$ ensures that the third term after the second equality is also of order $O_p(1)$. Thus, the second term after the second equality is also of order $O_p(1)$. Since $E\{\partial\mathcal{S}(Y_i,W_i,V_i,Z_i;\theta_n^\dagger,\gamma_n^\dagger)/\partial\theta^T\}$ is nonsingular and bounded, we obtain that $\widehat{\theta}_n-\theta$ is of order $O_p(n^{-1/2})$.

Therefore, our Taylor expansion above becomes

$$\begin{aligned}
0 &= n^{-1/2}\sum_{i=1}^{n}\mathcal{S}(Y_i,W_i,V_i,Z_i;\theta_0,\gamma^*)\\
&\quad +n^{1/2}\left[E\left\{\frac{\partial\mathcal{S}(Y_1,W_1,V_1,Z_1;\theta_0,\gamma^*)}{\partial\theta^T}\right\}+o_p(1)\right](\widehat{\theta}_n-\theta_0)\\
&\quad +n^{1/2}\left[E\left\{\frac{\partial\mathcal{S}(Y_1,W_1,V_1,Z_1;\theta_0,\gamma^*)}{\partial\gamma^T}\right\}+o_p(1)\right](\widehat{\gamma}_n-\gamma^*)\\
&= n^{-1/2}\sum_{i=1}^{n}\mathcal{S}(Y_i,W_i,V_i,Z_i;\theta_0,\gamma^*)\\
&\quad +n^{1/2}\left[E\left\{\frac{\partial\mathcal{S}(Y_1,W_1,V_1,Z_1;\theta_0,\gamma^*)}{\partial\theta^T}\right\}+o_p(1)\right](\widehat{\theta}_n-\theta_0)+o_p(1) \qquad \text{(S.6)}
\end{aligned}$$

The second equality holds because we showed that $E\{\partial S(Y_1, W_1, V_1, Z_1; \theta_0; \gamma^*)/\partial \gamma^T\} = 0$.

From re-arranging (S.6) and given that $E\{\partial \mathcal{S}(Y, W, V, Z; \theta_0, \gamma^*)/\partial \theta^T\}$ is invertible

$$n^{1/2}(\widehat{\theta}_n - \theta_0) = n^{-1/2} \sum_{i=1}^{n} \left[-E\{\partial \mathcal{S}(Y, W, V, Z; \theta_0, \gamma^*)/\partial \theta^T\}\right]^{-1} \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta_0, \gamma^*) + o_p(1).$$

Using the central limit theorem, the above display implies that as $n \to \infty$,

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \to \text{Normal}\{0, \mathcal{A}_*^{-1}\mathcal{B}_*(\mathcal{A}_*^{-1})^T\}$$

in distribution.

S.7 Proof of Theorem 3

The result holds since

$$\begin{aligned}
0 &= n^{-1/2} \sum_{i=1}^{n} \mathcal{S}(Y_i, W_i, V_i, Z_i; \check{\theta}_n, \gamma^*) \\
&= n^{-1/2} \sum_{i=1}^{n} \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta_0, \gamma^*) \\
&\quad + n^{-1/2} \sum_{i=1}^{n} \frac{\partial \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta_n^\dagger, \gamma^*)}{\partial \theta^T} (\check{\theta}_n - \theta_0) \\
&= n^{-1/2} \sum_{i=1}^{n} \mathcal{S}(Y_i, W_i, V_i, Z_i; \theta_0, \gamma^*) \\
&\quad + n^{1/2} \left[E\left\{ \frac{\partial \mathcal{S}(Y_1, W_1, V_1, Z_1; \theta_0, \gamma^*)}{\partial \theta^T} \right\} + o_p(1) \right] (\check{\theta}_n - \theta_0),
\end{aligned}$$

where $\theta_n^\dagger$ lies on the line connecting $\check{\theta}_n$ and θ_0. The above is the same first-order expansion for $\widehat{\theta}_n$ as in Section S.6 (Supplementary Material). Hence, from this first-order expansion and the asymptotic results of Theorem 2, Theorem 3 holds.

S.8 Constraints for heteroskedastic sieve estimator

Let $c_j(x,z) = \cos\left\{\frac{j\pi}{\ell_x} m(x,z;\beta)\right\}$, and $c_k(\epsilon) = \cos\left(\frac{k\pi}{\ell_e}\epsilon\right)$, and $s_k(\epsilon) = \sin\left(\frac{k\pi}{\ell_e}\epsilon\right)$. Then, using standard integration techniques, we have that

$$\begin{aligned}\int f_1(\epsilon|x,z)d\epsilon &= 2\ell_e\{a_{00}+a_{01}c_1(x,z)+a_{02}c_2(x,z)\}^2 + \ell_e\sum_{k=1}^{3}\{a_{k0}+a_{k1}c_1(x,z)+a_{k2}c_2(x,z)\}^2 \\ &\quad +\ell_e\sum_{k=1}^{3}\{b_{k0}+b_{k1}c_1(x,z)+b_{k2}c_2(x,z)\}^2\end{aligned}$$

and

$$\begin{aligned}&\int \epsilon f_1(\epsilon|x)d\epsilon = \\ &-4\frac{\ell_e^2}{\pi}\{a_{00}+a_{01}c_1(x,z)+a_{02}c_2(x,z)\}\sum_{k=1}^{3}\{b_{k0}+b_{k1}c_1(x,z)+b_{k2}c_2(x,z)\}\frac{(-1)^k}{k\pi} \\ -\ &2\frac{\ell_e^2}{\pi}\sum_{k=1}^{3}\sum_{j\neq k}\left[\{a_{k0}+a_{k1}c_1(x,z)+a_{k2}c_2(x,z)\}\right. \\ &\left.\times\{b_{j0}+b_{j1}c_1(x,z)+b_{j2}c_2(x,z)\}\left\{\frac{(-1)^{|j-k|}}{(j-k)}+\frac{(-1)^{|j+k|}}{(j+k)}\right\}\right] \\ -\ &\frac{\ell_e^2}{\pi}\sum_{k=1}^{3}\{a_{k0}+a_{k1}c_1(x,z)+a_{k2}c_2(x,z)\}\{b_{k0}+b_{k1}c_1(x,z)+b_{k2}c_2(x,z)\}\frac{1}{k}.\end{aligned}$$

To ensure $\int f_1(\epsilon|x,z) = 1$ and $\int \epsilon f_1(\epsilon|x,z) = 0$, we thus require

$$\begin{aligned}&2a_{00}^2+\sum_{k=1}^{3}(a_{k0}^2+b_{k0}^2) = 1/\ell_e; \quad 2a_{01}a_{02}+\sum_{k=1}^{3}(a_{k1}a_{k2}+b_{k1}b_{k2}) = 0; \\ &\sum_{k=1}^{3}\left[4\frac{(-1)^k a_{00}b_{k0}+a_{k0}b_{k0}}{k}+2\sum_{j\neq k}\left\{\frac{(-1)^{|j-k|}}{(j-k)}+\frac{(-1)^{|j+k|}}{(j+k)}\right\}a_{k0}b_{j0}\right] = 0; \\ &\sum_{k=1}^{3}\left[4\frac{(-1)^k(a_{01}b_{k2}+a_{02}b_{k1})+a_{k1}b_{k2}+a_{k2}b_{k1}}{k}\right. \\ &\left.+2\sum_{j\neq k}\left\{\frac{(-1)^{|j-k|}}{(j-k)}+\frac{(-1)^{|j+k|}}{(j+k)}\right\}(a_{k1}b_{j2}+a_{k2}b_{j1})\right] = 0;\end{aligned}$$

and for $\ell = 1, 2$,

$$
\begin{aligned}
&2a_{00}a_{0\ell} + \sum_{k=1}^{3}(a_{k0}a_{k\ell} + b_{k0}b_{k\ell}) = 0; \quad 2a_{0\ell}^2 + \sum_{k=1}^{3}(a_{k\ell}^2 + b_{k\ell}^2) = 0; \\
&\sum_{k=1}^{3}\left[4\frac{(-1)^k(a_{00}b_{k\ell} + a_{0\ell}b_{k0}) + a_{k0}b_{k\ell} + a_{k\ell}b_{k0}}{k}\right. \\
&\left.+2\sum_{j\neq k}\left\{\frac{(-1)^{|j-k|}}{(j-k)} + \frac{(-1)^{|j+k|}}{(j+k)}\right\}(a_{k0}b_{j\ell} + a_{k\ell}b_{j0})\right] = 0; \\
&\sum_{k=1}^{3}\left[4\frac{(-1)^k a_{0\ell}b_{k\ell} + a_{k\ell}b_{k\ell}}{k} + 2\sum_{j\neq k}\left\{\frac{(-1)^{|j-k|}}{(j-k)} + \frac{(-1)^{|j+k|}}{(j+k)}\right\}a_{k\ell}b_{j\ell}\right] = 0.
\end{aligned}
$$

S.9 Additional simulation results

We considered two additional mean models:

$$
\begin{aligned}
m(X, Z; \beta) &= \beta_1 + \beta_2 X + \beta_3 Z; \\
m(X, Z; \beta) &= \beta_2 \exp(-\beta_1 X) + \beta_3 Z.
\end{aligned} \tag{S.7}
$$

The simulation design was similar to that described in Section 4.1. Results in Tables S.1-S.4 demonstrate that the proposed estimator results in the least amount of bias compared to the competing methods regardless of the true model error variance structure. In comparison, the homoskedastic sieve MLE was biased when the model error was heteroskedastic, the heteroskedastic sieve MLE was biased regardless of the model error variance structure, and the Tsiatis-Ma estimators were biased when the assumption of the model error variance was incorrect.

As in the main text, the proposed method was minimally affected by misspecification of the working models. Under the same four cases described in Section 4.3.2, we report in Tables S.5 and S.6 the results from our proposed method when the working models differed from the truth. Interestingly, in Case 2, when only η_1^* was misspecified, the estimation results were similar to Case 1 when η_1^* was the truth. This was observed for both mean models in (S.7). Otherwise, even when η_1^*, η_2^* were different from the true η_{10}, η_{20}, the efficiency of the estimates was only slightly less than when the working models were the truth.

Table S.1: Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}^2_U, \widehat{\beta}^T)^T$ based on our proposed method (Semipar), homoskedastic sieve MLE (Sieve-Hom), heteroskedastic sieve MLE (Sieve-Het), Tsiatis-Ma homoskedastic estimator (TM-Hom), Tsiatis-Ma heteroskedastic estimator (TM-Het), and the naive estimator. Results based on 1000 simulations when $m(X, Z; \beta) = \beta_1 + \beta_2 X + \beta_3 Z$, and true parameter values $(\sigma^2_{U,0}, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$.

	Setting 1: $\eta_{20} \sim$Uniform				Setting 2: $\eta_{20} \sim t_5$			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
				η_{20}: Homoskedastic				
				Semipar				
bias	-0.0047	0.0021	0.0024	5.1664×10^{-5}	0.0011	-0.0012	0.0010	-9.2372×10^{-5}
var	0.0049	0.0028	0.0028	1.0255×10^{-5}	0.0041	0.0024	0.0022	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0048	0.0028	0.0029	1.0255×10^{-5}	0.0039	0.0023	0.0023	9.9839×10^{-6}
CI	0.9530	0.9480	0.9550	0.9490	0.9460	0.9460	0.9620	0.9320
				Sieve-Hom*				
bias	0.0080	-0.0091	0.0022	5.1664×10^{-5}	-0.0093	0.0086	0.0015	-9.2372×10^{-5}
var	0.0044	0.0025	0.0024	1.0255×10^{-5}	0.0041	0.0024	0.0021	1.0823×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				Sieve-Het*				
bias	0.0835	0.7104	0.6000	9.6823×10^{-6}	-0.2467	0.0278	-0.0373	-5.0634×10^{-5}
var	0.1159	0.0383	0.0550	1.0916×10^{-5}	0.0547	0.0239	0.0127	$1.0109\text{e}\times10^{-5}$
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				TM-Hom				
bias	-0.0007	0.0006	0.0008	-0.0002	-0.0018	0.0019	-0.0015	-7.9103×10^{-5}
var	0.0050	0.0029	0.0027	1.0396×10^{-5}	0.0043	0.0024	0.0023	9.6008×10^{-6}
$\widehat{\text{var}}$	0.0047	0.0027	0.0029	9.8539×10^{-6}	0.0040	0.0023	0.0023	9.941×10^{-6}
CI	0.9410	0.9430	0.9590	0.9490	0.9380	0.9390	0.9510	0.9470
				TM-Het				
bias	0.0729	-0.0560	0.0006	-0.0002	0.0851	-0.0673	-0.0015	-7.9103×10^{-5}
var	0.0050	0.0029	0.0032	1.0396×10^{-5}	0.0037	0.0020	0.0025	9.6008×10^{-6}
$\widehat{\text{var}}$	0.0047	0.0028	0.0032	9.8539×10^{-6}	0.0034	0.0019	0.0025	9.941×10^{-6}
CI	0.8000	0.8120	0.9500	0.9490	0.6730	0.6590	0.9470	0.9470
				Naive				
bias	0.1067	-0.0991	0.0021	5.1664×10^{-5}	0.1111	-0.1013	0.0010	-9.2372×10^{-5}
var	0.0038	0.0020	0.0027	1.0255×10^{-5}	0.0032	0.0017	0.0022	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0038	0.0020	0.0028	1.0255×10^{-5}	0.0031	0.0016	0.0023	9.9839×10^{-6}
CI	0.5860	0.4030	0.9520	0.9490	0.4780	0.2910	0.9590	0.9320

*Estimated variances not available. The homoskedastic sieve MLE uses smoothing parameters $\kappa_\epsilon = \kappa_x = 6$.

Table S.2: Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}^2_U, \widehat{\beta}^T)^T$ based on our proposed method (Semipar), homoskedastic sieve MLE (Sieve-Hom), heteroskedastic sieve MLE (Sieve-Het), Tsiatis-Ma homoskedastic estimator (TM-Hom), Tsiatis-Ma heteroskedastic estimator (TM-Het), and the naive estimator. Results based on 1000 simulations when $m(X, Z; \beta) = \beta_1 + \beta_2 X + \beta_3 Z$, and true parameter values $(\sigma^2_{U,0}, \beta^T_0)^T = (0.05, 0.25, 0.7, 0.5)^T$.

	Setting 1: $\eta_{20} \sim$Uniform				Setting 2: $\eta_{20} \sim t_5$			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
			η_{20}: Heteroskedastic					
				Semipar				
bias	-0.0061	0.0011	0.0049	5.1664×10^{-5}	0.0032	-0.0046	0.0036	-9.2372×10^{-5}
var	0.0164	0.0131	0.0121	1.0255×10^{-5}	0.0173	0.0133	0.0123	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0160	0.0129	0.0127	1.0255×10^{-5}	0.0163	0.0127	0.0126	9.9839×10^{-6}
CI	0.9430	0.9510	0.9520	0.9490	0.9520	0.9490	0.9510	0.9320
				Sieve-Hom*				
bias	0.2443	-0.2347	0.0035	4.6779×10^{-5}	-0.2980	0.2476	0.0003	-9.413×10^{-5}
var	0.1356	0.1202	0.0105	1.0231×10^{-5}	0.0784	0.0635	0.0230	1.0818×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				Sieve-Het*				
bias	-0.1375	0.7656	0.5292	9.6823×10^{-6}	-0.2762	0.2761	0.0594	-5.0634×10^{-5}
var	0.0703	0.0634	0.0528	1.0916×10^{-5}	0.0876	0.1461	0.1086	1.0109×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				TM-Hom				
bias	-0.8754	0.7947	0.0027	-0.0002	-1.3688	1.2453	-0.0051	-8.4908×10^{-5}
var	0.0740	0.0601	0.0142	1.0396×10^{-5}	0.5476	0.4367	0.0200	9.5992×10^{-6}
$\widehat{\text{var}}$	0.0594	0.0491	0.0151	9.8539×10^{-6}	0.4999	0.4047	0.0205	9.939×10^{-6}
CI	0.0560	0.0710	0.9560	0.9490	0.0620	0.0660	0.9690	0.9470
				TM-Het				
bias	-0.0116	0.0114	0.0008	-0.0002	0.0112	-0.0091	-0.0029	-7.9103×10^{-5}
var	0.0144	0.0115	0.0104	1.0396×10^{-5}	0.0134	0.0111	0.0099	9.6008×10^{-6}
$\widehat{\text{var}}$	0.0138	0.0112	0.0106	9.8539×10^{-6}	0.0129	0.0110	0.0101	9.941×10^{-6}
CI	0.9410	0.9460	0.9570	0.9490	0.9430	0.9420	0.9530	0.9470
				Naive				
bias	0.1052	-0.1000	0.0046	5.1664×10^{-5}	0.1126	-0.1042	0.0036	-9.2372×10^{-5}
var	0.0128	0.0096	0.0120	1.0255×10^{-5}	0.0138	0.0097	0.0124	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0126	0.0094	0.0127	1.0255×10^{-5}	0.0128	0.0093	0.0126	9.9839×10^{-6}
CI	0.8470	0.8180	0.9550	0.9490	0.8190	0.8060	0.9540	0.9320

*Estimated variances not available. The homoskedastic sieve MLE uses smoothing parameters $\kappa_\epsilon = \kappa_x = 6$.

Table S.3: Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}^2_U, \widehat{\beta}^T)^T$ based on our proposed method (Semipar), homoskedastic sieve MLE (Sieve-Hom), heteroskedastic sieve MLE (Sieve-Het), Tsiatis-Ma homoskedastic estimator (TM-Hom), Tsiatis-Ma heteroskedastic estimator (TM-Het), and the naive estimator. Results based on 1000 simulations when $m(X, Z; \beta) = \beta_2 \exp(-\beta_1 X) + \beta_3 Z$, and true parameter values $(\sigma^2_{U,0}, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$.

	Setting 1: $\eta_{20} \sim$Uniform				Setting 2: $\eta_{20} \sim t_5$			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
				η_{20}: Homoskedastic				
				Semipar				
bias	0.0009	0.0000	0.0018	5.1664×10^{-5}	0.0036	0.0028	0.0014	-9.2372×10^{-5}
var	0.0097	0.0062	0.0026	1.0255×10^{-5}	0.0078	0.0050	0.0022	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0098	0.0061	0.0027	1.0255×10^{-5}	0.0078	0.0048	0.0021	9.9839×10^{-6}
CI	0.9560	0.9420	0.9550	0.9490	0.9550	0.9490	0.9510	0.9320
				Sieve-Hom*				
bias	-0.0055	-0.0048	0.0016	5.1664×10^{-5}	0.0366	0.0214	0.0098	-9.2372×10^{-5}
var	0.0076	0.0050	0.0022	1.0255×10^{-5}	0.0113	0.0068	0.0025	1.0823×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				Sieve-Het*				
bias	0.5870	0.7665	0.5211	9.6823×10^{-6}	0.7287	-0.1489	0.3844	-5.0634×10^{-5}
var	0.0466	0.0839	0.0695	1.0916×10^{-5}	0.0486	0.0574	0.0285	1.0109×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				TM-Hom				
bias	0.0053	0.0113	-0.0018	-0.0002	0.0113	0.0139	-0.0019	-0.0001
var	0.0095	0.0056	0.0025	1.0422×10^{-5}	0.0390	0.0248	0.0024	9.6642×10^{-6}
$\widehat{\text{var}}$	0.0094	0.0060	0.0027	9.8246×10^{-6}	0.0122	0.0114	0.0021	9.9276×10^{-6}
CI	0.9530	0.9590	0.9550	0.9450	0.9490	0.9540	0.9580	0.9470
				TM-Het				
bias	-0.0097	-0.0033	-0.0015	-0.0002	-0.0216	-0.0146	-0.0016	-0.0001
var	0.0094	0.0053	0.0031	1.0521×10^{-5}	0.0068	0.0039	0.0023	9.3885×10^{-6}
$\widehat{\text{var}}$	0.0100	0.0060	0.0031	9.8237×10^{-6}	0.0067	0.0040	0.0023	9.9091×10^{-6}
CI	0.9490	0.9610	0.9530	0.9450	0.9400	0.9420	0.9530	0.9500
				Naive				
bias	-0.0370	-0.0283	0.0018	5.1664×10^{-5}	-0.0346	-0.0256	0.0014	-9.2372×10^{-5}
var	0.0068	0.0045	0.0026	1.0255×10^{-5}	0.0054	0.0036	0.0021	1.0823×10^{-5}
$\widehat{\text{var}}$	0.0069	0.0044	0.0027	1.0255×10^{-5}	0.0055	0.0035	0.0021	9.9839×10^{-6}
CI	0.9290	0.9240	0.9520	0.9490	0.9150	0.9120	0.9520	0.9320

*Estimated variances not available. The homoskedastic sieve MLE uses smoothing parameters $\kappa_\epsilon = \kappa_x = 6$.

Table S.4: Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}_U^2, \widehat{\beta}^T)^T$ based on our proposed method (Semipar), homoskedastic sieve MLE (Sieve-Hom), heteroskedastic sieve MLE (Sieve-Het), Tsiatis-Ma homoskedastic estimator (TM-Hom), Tsiatis-Ma heteroskedastic estimator (TM-Het), and the naive estimator. Results based on 1000 simulations when $m(X, Z; \beta) = \beta_2 \exp(-\beta_1 X) + \beta_3 Z$, and true parameter values $(\sigma_{U,0}^2, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$.

	Setting 1: $\eta_{20} \sim$Uniform				Setting 2: $\eta_{20} \sim t_5$			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}_U^2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}_U^2$
				η_{20}: Heteroskedastic				
				Semipar				
bias	0.0061	0.0036	0.0029	2.1036×10^{-5}	0.0212	0.0150	0.0034	-8.0252×10^{-5}
var	0.0521	0.0214	0.0117	1.0162×10^{-5}	0.0525	0.0224	0.0125	1.082×10^{-5}
$\widehat{\text{var}}$	0.0512	0.0203	0.0126	1.0013×10^{-5}	0.0549	0.0221	0.0125	9.9871×10^{-6}
CI	0.9190	0.9280	0.9590	0.9500	0.9370	0.9420	0.9510	0.9320
				Sieve-Hom*				
bias	0.0810	0.0887	-0.0056	3.9759×10^{-5}	0.3347	0.1942	0.0117	-9.0088×10^{-5}
var	0.3353	0.1259	0.0123	1.0232×10^{-5}	0.1545	0.0891	0.0263	1.0827×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				Sieve-Het*				
bias	0.6686	0.6244	0.5111	9.6823×10^{-6}	0.7224	-0.2329	0.5513	-5.0634×10^{-5}
var	0.0470	0.0973	0.0591	1.0916×10^{-5}	0.0409	0.0494	0.0447	1.0109×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA	NA	NA	NA	NA
CI	NA	NA	NA	NA	NA	NA	NA	NA
				TM-Hom				
bias	0.3116	0.2645	-0.0003	-0.0002	0.4474	0.4003	-0.0016	-0.0004
var	0.3049	0.1760	0.0136	1.0185×10^{-5}	0.3095	0.3174	0.0143	9.1678×10^{-6}
$\widehat{\text{var}}$	0.2551	0.1282	0.0129	9.8516×10^{-6}	0.6010	0.5999	0.0137	9.8078×10^{-6}
CI	0.8890	0.9250	0.9500	0.9470	0.9240	0.9370	0.9500	0.9500
				TM-Het				
bias	0.0344	0.0366	-0.0041	-0.0002	0.0218	0.0272	-0.0068	-0.0001
var	0.0541	0.0164	0.0100	1.049×10^{-5}	0.0684	0.0296	0.0094	9.4827×10^{-6}
$\widehat{\text{var}}$	0.0496	0.0211	0.0104	9.8435×10^{-6}	0.0440	0.0188	0.0099	9.9245×10^{-6}
CI	0.9660	0.9670	0.9590	0.9440	0.9590	0.9560	0.9600	0.9500
				Naive				
bias	-0.0234	-0.0304	0.0035	2.1036×10^{-5}	-0.0273	-0.0219	0.0027	-8.0252×10^{-5}
var	0.2092	0.0150	0.0122	1.0162×10^{-5}	0.0335	0.0150	0.0124	1.082×10^{-5}
$\widehat{\text{var}}$	0.0615	0.0142	0.0132	1.0013×10^{-5}	0.0358	0.0146	0.0124	9.9871×10^{-6}
CI	0.9560	0.9470	0.9590	0.9500	0.9480	0.9450	0.9520	0.9320

*Estimated variances not available. The homoskedastic sieve MLE uses smoothing parameters $\kappa_\epsilon = \kappa_x = 6$.

Table S.5: Evaluation of efficiency loss from proposed method when working models η_1^*, η_2^* may differ from the true η_{10}, η_{20}. Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}_U^2, \widehat{\beta}^T)^T$ with true parameter values $(\sigma_{U,0}^2, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$ and $m(X, Z; \beta) = \beta_1 + \beta_2 X + \beta_3 Z$. Results based on 1000 simulations.

Setting		$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}_U^2$
$\eta_1^* = \eta_{10}, \eta_2^* = \eta_{20}$	bias	0.0012	-0.0008	-0.0014	-3.9892×10^{-5}
	var	0.0008	0.0008	0.0016	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0007	0.0008	0.0015	9.9257×10^{-6}
	CI	0.9390	0.9430	0.9330	0.9430
$\eta_1^* \neq \eta_{10}, \eta_2^* = \eta_{20}$	bias	0.0012	-0.0008	-0.0014	-3.9892×10^{-5}
	var	0.0008	0.0008	0.0016	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0007	0.0008	0.0015	9.9257×10^{-6}
	CI	0.9390	0.9430	0.9330	0.9430
$\eta_1^* = \eta_{10}, \eta_2^* \neq \eta_{20}$	bias	-0.0003	0.0029	-0.0031	-3.9892×10^{-5}
	var	0.0018	0.0010	0.0036	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0017	0.0010	0.0033	9.9257×10^{-6}
	CI	0.9460	0.9440	0.9510	0.9430
$\eta_1^* \neq \eta_{10}, \eta_2^* \neq \eta_{20}$	bias	-0.0007	0.0036	-0.0027	-3.9892×10^{-5}
	var	0.0025	0.0011	0.0048	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0023	0.0010	0.0044	9.9257×10^{-6}
	CI	0.9510	0.9470	0.9490	0.9430

Table S.6: Evaluation of efficiency loss from proposed method when working models η_1^*, η_2^* may differ from the true η_{10}, η_{20}. Bias, empirical sample variances (var), averaged estimated variances ($\widehat{\text{var}}$), and estimated 95% coverage probabilities (CI) for $(\widehat{\sigma}_U^2, \widehat{\beta}^T)^T$ with true parameter values $(\sigma_{U,0}^2, \beta_0^T)^T = (0.05, 0.25, 0.7, 0.5)^T$ and $m(X, Z; \beta) = \beta_2 \exp(-\beta_1 X) + \beta_3 Z$. Results based on 1000 simulations.

Setting		$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}_U^2$
$\eta_1^* = \eta_{10}, \eta_2^* = \eta_{20}$	bias	0.0015	0.0012	-0.0017	-3.9892×10^{-5}
	var	0.0015	0.0007	0.0013	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0014	0.0007	0.0013	9.9257×10^{-6}
	CI	0.9460	0.9420	0.9490	0.9430
$\eta_1^* \neq \eta_{10}, \eta_2^* = \eta_{20}$	bias	0.0015	0.0012	-0.0017	-3.9892×10^{-5}
	var	0.0015	0.0007	0.0013	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0014	0.0007	0.0013	9.9257×10^{-6}
	CI	0.9480	0.9420	0.9490	0.9430
$\eta_1^* = \eta_{10}, \eta_2^* \neq \eta_{20}$	bias	-0.0020	0.0025	-0.0064	-3.9892×10^{-5}
	var	0.0016	0.0010	0.0019	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0016	0.0009	0.0018	9.9257×10^{-6}
	CI	0.9440	0.9430	0.9440	0.9430
$\eta_1^* \neq \eta_{10}, \eta_2^* \neq \eta_{20}$	bias	-0.0017	0.0025	-0.0066	-3.9892×10^{-5}
	var	0.0017	0.0010	0.0021	1.0103×10^{-5}
	$\widehat{\text{var}}$	0.0017	0.0010	0.0019	9.9257×10^{-6}
	CI	0.9470	0.9470	0.9420	0.9430