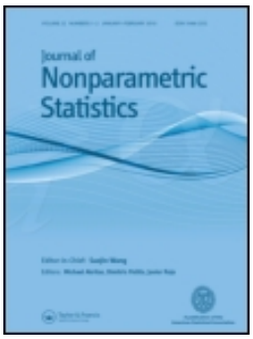

Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: www.tandfonline.com/journals/gnst20

Identification and estimation of nonlinear models using two samples with nonclassical measurement errors

Raymond J. Carroll, Xiaohong Chen & Yingyao Hu

To cite this article: Raymond J. Carroll, Xiaohong Chen & Yingyao Hu (2010) Identification and estimation of nonlinear models using two samples with nonclassical measurement errors, Journal of Nonparametric Statistics, 22:4, 379-399, DOI: 10.1080/10485250902874688

To link to this article: https://doi.org/10.1080/10485250902874688

Journal of Nonparametric Statistics
Vol. 22, No. 4, May 2010, 379–399

Taylor & Francis
Taylor & Francis Group

Identification and estimation of nonlinear models using two samples with nonclassical measurement errors

Raymond J. Carroll[a]*, Xiaohong Chen[b] and Yingyao Hu[c]

[a]*Department of Statistics, Texas A&M University, USA;* [b]*Department of Economics, Yale University, USA;* [c]*Department of Economics, Johns Hopkins University, USA*

(*Received 29 September 2008; final version received 27 February 2009*)

This paper considers identification and estimation of a general nonlinear errors-in-variables (EIV) model using two samples. Both samples consist of a dependent variable, some error-free covariates, and an error-prone covariate, for which the measurement error has unknown distribution and could be arbitrarily correlated with the latent true values, and neither sample contains an accurate measurement of the corresponding true variable. We assume that the regression model of interest – the conditional distribution of the dependent variable given the latent true covariate and the error-free covariates – is the same in both samples, but the distributions of the latent true covariates vary with observed error-free discrete covariates. We first show that the general latent nonlinear model is *nonparametrically* identified using the two samples when both could have nonclassical errors, without either instrumental variables or independence between the two samples. When the two samples are independent and the nonlinear regression model is parameterised, we propose sieve quasi maximum likelihood estimation (Q-MLE) for the parameter of interest, and establish its root-*n* consistency and asymptotic normality under possible misspecification, and its semiparametric efficiency under correct specification, with easily estimated standard errors. A Monte Carlo simulation and a data application are presented to show the power of the approach.

Keywords: data combination; measurement error; misspecified parametric latent model; nonclassical measurement error; nonlinear errors-in-variables model; nonparametric identification; sieve quasi likelihood

1. Introduction

Measurement error problems are frequently encountered by researchers conducting empirical studies in the social and natural sciences. A measurement error is called *classical* if it is independent of the latent true values; otherwise, it is called *nonclassical*. There have been many studies on identification and estimation of linear, nonlinear, and even nonparametric models with classical measurement errors, see, e.g. (Cheng and Van Ness 1999; Carroll, Ruppert, Stefanski, and Crainiceanu 2006) for detailed reviews. However, numerous validation studies in survey data sets indicate that the errors in self-reported variables, such as earnings, are typically correlated with the true values, and hence, are nonclassical (Bound, Brown, and Mathiowetz 2001).

*Corresponding author. Email: carroll@stat.tamu.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250902874688
http://www.informaworld.com

In this work, we study the identification and estimation of possibly nonclassical, nonlinear measurement error models when (a) there are no validation data, i.e. no data where the error-prone covariate is known exactly, (b) there is no knowledge of the measurement error distribution,; and (c) there is no instrumental variable. As far as we know, this is the first paper to allow identification and estimation in the absence of knowledge about the measurement error distribution, of an instrumental variable and of validation data. Of course, some assumptions must be made, and our assumptions include that (i) there are two data sets with the same distribution for the response given the true covariates; (ii) the measurement error is nondifferential; (iii) there is a discrete-valued covariate that is not exogenous and that the distribution of the error-prone covariate given the discrete covariate differs in the two data sets, and (iv) some linear integral operators are invertible.

1.1. *Identification*

In Section 2, we show that by combining two samples the distributions of the latent nonlinear regression model, the measurement error model, and the model for the error-prone covariate are all nonparametrically identified. We assume that each sample consists of a dependent variable (Y), some error-free covariates (W), and only one measurement of the error-ridden covariate (X). In both samples, the measurement error has an unknown distribution and could be arbitrarily correlated with the latent true value (X^*), and neither sample contains an accurate measurement of the corresponding true variable. We assume that the latent model of interest, $f_{Y|X^*,W}$, the conditional distribution of the dependent variable given the latent true covariate and the error-free covariates, is the same in both samples, but the marginal distributions of the latent true variables differ across some contrasting subsamples indexed by W, which may be different geographic areas, age groups, or other observed demographic characteristics.

There are currently three broad identification methods for general nonparametric nonlinear errors-in-variables (EIV) models. The first imposes parametric restrictions on measurement error distributions (Fan 1991; Liang, Härdle, and Carroll 1999; Hong and Tamer 2003). The second assumes the existence of instrumental variables (IV), such as a repeated measurement of the mismeasured covariates, that do not enter the latent model of interest but do contain information on how to recover features of latent true variables (Hausman, Ichimura, Newey, and Powell 1991; Li and Vuong 1998; Li 2002; Carroll, Ruppert, Crainiceanu, Tosteson, and Karagas 2004, Wang 2004; Zinde-Walsh 2007; Hu and Schennach 2008). The third approach is to combine two samples in which one sample also contains an accurate measurement of the latent true variable (Carroll and Wand 1991; Lee and Sepanski 1995; Chen, Hong, and Tamer 2005). Additional references and discussions about these existing methods can be found in a recent survey by Chen, Hong, and Nekipelov (2007).

Our identification strategy differs from all the existing ones. In particular, we do not require an IV excluded from the latent model of interest, and all the variables in our samples may be included in the model. Also, neither of our two samples contains an accurate measurement of the latent true variable.

1.2. *Estimation and inference*

Our identification result allows for fully nonparametric EIV models and also allows for two correlated samples. However, in most empirical applications, the latent models of interest are parametric nonlinear models, and the two samples are regarded as independent. In Section 3, we tackle the question of how to do practical estimation and inference in the case that the distribution of the response given the true predictors is specified parametrically (could be misspecified), but that the measurement error model and the model for the error-prone covariate are

nonparametric. We propose a two-sample sieve quasi-maximum likelihood estimator (Q-MLE) of all the unknown finite- and infinite-dimensional model parameters, and establish its consistency. Under possible misspecification of the latent parametric model, we establish the root-n asymptotic normality of the sieve Q-MLE of the finite-dimensional parameters, as well as its semiparametric efficiency under correct specification. Easily computed standard errors are also provided.

1.3. *Outline*

Section 2 establishes the nonparametric identification of the latent model of interest, $f_{Y|X^*,W}$, using two samples with (possibly) nonclassical errors. Section 3 presents the two-sample sieve Q-MLE for a possibly misspecified parametric latent model. Section 4 provides a Monte Carlo study, and Section 5 contains an empirical illustration. In Section 4, we describe the development of a device for checking the assumption that the regression model is the same in the two samples, based on the work of Huang, Stefanski, and Davidian (2006). We apply this method to the empirical example in Section 5, showing that the assumptions seem reasonable in the context. The appendix contains technical arguments.

A long version of this paper is available at http://www.stat.tamu.edu/ftp/pub/rjcarroll/2009.papers.directory/CCH_Long.pdf. It contains additional identification results for the case of discrete random variables, a second empirical example and proofs of the asymptotic normality of our sieve estimator.

2. Nonparametric identification

In this paper, $f_{A|B}$ denotes the conditional density of A given B, while f_A denotes the density of A. We assume the existence of two samples. One sample (sometimes called a primary sample) is a random sample from (X, W, Y), in which X is a mismeasured X^*, and another sample (sometimes called an auxiliary sample) is a random sample from (X_a, W_a, Y_a), in which X_a is a mismeasured X_a^*. These two samples could be correlated and could have different joint distributions.

We are interested in identifying a latent probability model: $f_{Y|X^*,W}(y|x^*, w)$, in which Y is a continuous dependent variable, X^* is an unobserved (latent) continuous regressor subject to a possibly nonclassical measurement error, X is observed in place of X^*, and W is an accurately measured discrete covariate, e.g. subpopulations with different demographic characteristics, such as marital status, race, gender, profession, and geographic location.

Suppose the supports of X, X^*, Y, and W are $\mathcal{X} \subseteq \mathbb{R}$, $\mathcal{X}^* \subseteq \mathbb{R}$, $\mathcal{Y} \subseteq \mathbb{R}$, and $\mathcal{W} = \{w_1, w_2, \ldots, w_J\}$ with $J \geq 2$, respectively. We assume the following:

ASSUMPTION 2.1 (i) *$f_{Y,X,X^*,W}(y, x, x^*, w)$ is positive, bounded on its support $\mathcal{Y} \times \mathcal{X} \times \mathcal{X}^* \times \mathcal{W}$, and is continuous in $(y, x, x^*) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{X}^*$*; (ii) *$f_{X|X^*,W,Y}(x|x^*, w, y) = f_{X|X^*}(x|x^*)$ on $\mathcal{X} \times \mathcal{X}^* \times \mathcal{W} \times \mathcal{Y}$.*

ASSUMPTION 2.2 (i) *$f_{Y_a,X_a,X_a^*,W_a}(y, x, x^*, w)$ is positive, bounded on its support $\mathcal{Y} \times \mathcal{X}_a \times \mathcal{X}^* \times \mathcal{W}$, and is continuous in $(y, x, x^*) \in \mathcal{Y} \times \mathcal{X}_a \times \mathcal{X}^*$*; (ii) *$f_{X_a|X_a^*,W_a,Y_a}(x|x^*, w, y) = f_{X_a|X_a^*}(x|x^*)$ on $\mathcal{X}_a \times \mathcal{X}^* \times \mathcal{W} \times \mathcal{Y}$.*

Assumptions 2.1(ii) and 2.2(ii) assume that measurement errors in both samples are *nondifferential*. In such cases, X and X_a are surrogates for X^* and X_a^*, respectively.

ASSUMPTION 2.3 (i) *$f_{Y_a|X_a^*,W_a}(y|x^*, w) = f_{Y|X^*,W}(y|x^*, w)$ on $\mathcal{Y} \times \mathcal{X}^* \times \mathcal{W}$*; (ii) *$f_{Y|X^*,W=w}$ changes with w.*

Assumption 2.3(i) assumes that the latent structural probability model is the same in both samples, which is a reasonable *stability* assumption.

ASSUMPTION 2.4 (i) $\int f_{X_a|X_a^*}(x|x^*)h(x^*)\,\mathrm{d}x^* = 0$ *for all* $x \in \mathcal{X}$ *for all bounded function* h *implies that* $h \equiv 0$; (ii) *Assumption* $A.1$ *in the appendix holds.*

Assumption 2.4(i) says that the measurement error distribution of X_a given X_a^* is not pathological, which is commonly imposed in general deconvolution problems (Bissantz, Hohage, Munk, and Ruymgaart 2007). Assumption 2.4(i) is the *bounded completeness* of the conditional density $f_{X_a^*|X_a}$ (Mattner 1993). When X_a and X_a^* are discrete, Assumption 2.4(i) requires that the support of X_a is not smaller than that of X_a^*.

Define

$$k_{X_a^*}^{ij}(x^*) \equiv \frac{f_{X_a^*|W_j}(x^*)f_{X^*|W_i}(x^*)}{f_{X^*|W_j}(x^*)f_{X_a^*|W_i}(x^*)} \quad \text{for } x^* \in \mathcal{X}^*. \tag{1}$$

ASSUMPTION 2.5 *For any* $x_1^* \neq x_2^*$, *there exist* $i, j \in \{1, 2, \ldots, J\}$, *such that* $k_{X_a^*}^{ij}(x_1^*) \neq k_{X_a^*}^{ij}(x_2^*)$ *and* $\sup_{x^* \in \mathcal{X}^*} k_{X_a^*}^{ij}(x^*) < \infty$.

Note that the subsets $W_1, W_2, \ldots, W_J \subset \mathcal{W}$ do not need to be collectively exhaustive. We may only consider those subsets in $\mathcal{W}$ in which Assumption 2.5 holds. Since the indices i, j are exchangeable, the condition $\sup_{x^* \in \mathcal{X}^*} k_{X_a^*}^{ij}(x^*) < \infty$ may be replaced by $\inf_{x^* \in \mathcal{X}^*} k_{X_a^*}^{ij}(x^*) > 0$. Assumption 2.5 implies that the two samples are not random samples from the same population, nor are they derived by splitting a single random sample into two. In addition, the distributions of X^* given W and X_a^* given W_a are not identical.

ASSUMPTION 2.6 *One of the following holds for all* $x^* \in \mathcal{X}^*$: (i) (*mean*) $\int x f_{X_a|X_a^*}(x|x^*)\,\mathrm{d}x = x^*$ *or* (ii) *(mode)* $\arg\max_x f_{X_a|X_a^*}(x|x^*) = x^*$ *or (iii) (quantile) there is an* $\gamma \in (0, 1)$ *such that* $\inf\{z : \int_{-\infty}^{z} f_{X_a|X_a^*}(x|x^*)\,\mathrm{d}x \geq \gamma\} = x^*$.

Assumption 2.6 says that the surrogate X_a is targeted for the true X_a^*. Specifically, either the mean, mode or a fixed quantile of the distribution of X_a given X_a^* is equal to X_a^*. This condition is not required in the primary data set.

We obtain the following nonparametric identification result.

THEOREM 2.1 *Suppose Assumptions* 2.1–2.6 *hold, and the Assumption* A.1 *in the appendix also holds. Then, the densities* $f_{X,W,Y}$ *and* f_{X_a,W_a,Y_a} *uniquely determine* $f_{Y|X^*,W}$, $f_{X|X^*}$, $f_{X_a|X_a^*}$, $f_{X^*|W_j}$ *and* $f_{X_a^*|W_j}$.

The identification theorem does not require that the two samples are independent of each other.

2.1. *A simple example*

More complex examples, both simulated and real data-based, are given in Sections 4 and 5. Suppose that W and W_a are binary. Suppose further that given (X^*, W), Y is normally distributed with mean $\beta_0 + \beta_1 X^*$ and variance σ_ϵ^2, and also in the auxiliary sample. This is Assumption 2.3.

Next suppose that measurement error is nondifferential (Assumptions 2.1(ii) and 2.2(ii)), and that given (X^*, W), the observed surrogate X has mean X^* and variance σ_u^2, while in the auxiliary sample, X_a given (X_a^*, W_a) has mean X_a^* and variance σ_{ua}^2, i.e. different measurement error variances. This satisfies Assumption 2.6 with mean 'targeting' in the auxiliary data set.

Finally, suppose that X^* given W has mean μ_x independent of W, with variance σ_x^2, but that X_a^* given W_a has mean $\alpha_0 + \alpha_1 W$ and variance σ_{xa}^2. This satisfies Assumption 2.5 if either $\alpha_1 \neq 0$ or $\sigma_x^2 \neq \sigma_{xa}^2$.

Theorem 2.1 now asserts that if the measurement error does not have a pathological distribution, Assumption 2.4, then all the parameters listed are identified from the observed data, as are the unspecified distributions. For example, if $\alpha_1 \neq 0$, $E(Y_a|W_a = 1) - E(Y_a|W_a = 0) = \beta_1\alpha_1$ and $E(X_a|W_a = 1) - E(X_a|W_a = 0) = \alpha_1$, so that β_1 is readily identified.

2.2. *What does nonparametric identification tell us?*

We believe that our identification result is of real practical importance, see below.

Under our assumptions, the point is that all aspects of the problem can be identified: regression model, measurement error model, and latent variable model. Crucially, this says that whatever one's favorite paradigm, be it parametric, semiparametric, or nonparametric, be it Bayesian or frequentist, consistent estimation is possible. If one were Bayesian, then our result suggests that inference will not depend crucially upon the prior.

This opens up many different avenues for the construction of estimators of the regression function. At the most parametric level, it assures us that fully parametric models are identified and likelihood inference can proceed in the usual fashion. The result also tells us that if the regression model is semiparametric, but the measurement error model and the latent variable model are parametric, then we can still expect consistent and efficient estimation from semiparametric profile approaches.

In Section 3, we pursue one of the many variants of estimation and inference that our identification result makes possible. Specifically, in our theory we consider the case that the regression model is specified parametrically, but the measurement error and latent variable models are nonparametric. However, the power of the identification result is that we can do many other things. For example, in one of the empirical illustrations described in Section 5.1, we consider a parametric mean regression function but with the distribution of the deviations from the mean modeled nonparametrically: the identification result says that this approach too is consistent.

3. Sieve quasi likelihood estimation

Our identification result is very general and does not require the two samples to be independent. Nevertheless, for many applications, it is reasonable to assume that there are two random samples $\{X_i, W_i, Y_i\}_{i=1}^n$ and $\{X_{aj}, W_{aj}, Y_{aj}\}_{j=1}^{n_a}$, that are mutually independent.

Theorem 2.1 says that the densities $f_{Y|X^*,W}$, $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$, and $f_{X_a^*|W_a}$ are nonparametrically identified under Assumptions 2.1–2.6. Nevertheless, in empirical studies, we typically have either a semiparametric or a parametric specification of the conditional density $f_{Y|X^*,W}$ as the model of interest. In this section, we treat the other densities $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$, and $f_{X_a^*|W_a}$ as unknown nuisance functions, but consider a parametrically specified conditional density of Y given (X^*, W):

$$\{g(y|x^*, w; \theta) : \theta \in \Theta\}, \quad \Theta \text{ a compact subset of } \mathbb{R}^{d_\theta}, \quad 1 \leq d_\theta < \infty.$$

Define

$$\theta_0 \equiv \arg\max_{\theta\in\Theta} \int \log\{g(y|x^*, w; \theta)\} f_{Y|X^*,W}(y|x^*, w)\, \mathrm{d}y.$$

The latent parametric model is *correctly specified* if $g(y|x^*, w; \theta_0) = f_{Y|X^*,W}(y|x^*, w)$, and θ_0 is called true parameter value; otherwise it is *misspecified*, and θ_0 is called the pseudo-true parameter.

Let $\alpha_0 \equiv (\theta_0^{\mathrm{T}}, f_{01}, f_{01a}, f_{02}, f_{02a})^{\mathrm{T}} \equiv (\theta_0^{\mathrm{T}}, f_{X|X^*}, f_{X_a|X_a^*}, f_{X^*|W}, f_{X_a^*|W_a})^{\mathrm{T}}$ denote the true parameter values, in which θ_0 is really 'pseudo-true' when the parametric model $g(y|x^*, w; \theta)$ is incorrectly specified for the unknown true density $f_{Y|X^*,W}$. We first introduce a sieve MLE estimator $\widehat{\alpha}$ for α_0, and in later subsections establish the asymptotic normality of $\widehat{\theta}$.

3.1. *Sieve likelihood under possible misspecification*

Briefly, in the sieve method, we model the nonparametric densities for X given X^* and X^* given W via finite dimensional parametric representations, where this dimension increases with the sample size. A similar thing is done in the auxiliary sample. A good analogy is nonparametric regression, where the mean function is often modeled by a B-spline basis with the number of knots increasing with the sample size.

Of course, we need to impose some mild smoothness restrictions on the unknown densities. To do this, for concreteness we consider the widely used Hölder space of functions. Let $\xi = (\xi_1, \xi_2)^{\mathrm{T}} \in \mathbb{R}^2$, $a = (a_1, a_2)^{\mathrm{T}}$, and $\nabla^a h(\xi) \equiv (\partial^{a_1+a_2} h(\xi_1, \xi_2))/\partial\xi_1^{a_1}\partial\xi_2^{a_2}$ denote the $(a_1 + a_2)$th derivative. Let $\|\cdot\|_E$ denote the Euclidean norm. Let $\mathcal{V} \subseteq \mathbb{R}^2$ and $\underline{\gamma}$ be the largest integer satisfying $\gamma > \underline{\gamma}$. The Hölder space $\Lambda^\gamma(\mathcal{V})$ of order $\gamma > 0$ is a space of functions $h : \mathcal{V} \mapsto \mathbb{R}$, such that the first $\underline{\gamma}$ derivatives are continuous and bounded, and the $\underline{\gamma}$th derivative is Hölder continuous with the exponent $\gamma - \underline{\gamma} \in (0, 1]$. We define a Hölder ball as $\Lambda_c^\gamma(\mathcal{V}) \equiv \{h \in \Lambda^\gamma(\mathcal{V}) : \|h\|_{\Lambda^\gamma} \le c < \infty\}$, in which

$$\|h\|_{\Lambda^\gamma} \equiv \max_{a_1+a_2\le\underline{\gamma}} \sup_{\xi} |\nabla^a h(\xi)| + \max_{a_1+a_2=\underline{\gamma}} \sup_{\xi\neq\xi'} \frac{|\nabla^a h(\xi) - \nabla^a h(\xi')|}{(\|\xi - \xi'\|_E)^{\gamma-\underline{\gamma}}} < \infty.$$

The space of possible densities of X given X^* and of X_a given X_a^* are assumed to be in

$$\mathcal{F}_1 = \{f_1(\cdot|\cdot) \in \Lambda_c^{\gamma_1}(\mathcal{X} \times \mathcal{X}^*) : f_1(\cdot|x^*) \text{ is a positive density function for all } x^* \in \mathcal{X}^*\},$$

$$\mathcal{F}_{1a} = \left\{ \begin{array}{c} f_{1a}(\cdot|\cdot) \in \Lambda_c^{\gamma_{1a}}(\mathcal{X}_a \times \mathcal{X}^*) : \text{Assumption 2.6 holds,} \\ f_{1a}(\cdot|x^*) \text{ is a positive density function for all } x^* \in \mathcal{X}^* \end{array} \right\},$$

respectively. Also, the densities of X^* given W and of X_a^* given W_a are assumed to be in

$$\mathcal{F}_2 = \left\{ \begin{array}{c} f_2(\cdot|w), f_{2a}(\cdot|w) \in \Lambda_c^{\gamma_2}(\mathcal{X}^*) : \text{Assumption 2.5 holds,} \\ f_2(\cdot|w), f_{2a}(\cdot|w) \text{ are positive density functions for all } w \in \mathcal{W} \end{array} \right\}.$$

We introduce a dummy random variable S, with $S = 1$ indicating the primary sample and $S = 0$ indicating the auxiliary sample. Then we have the combined sample

$$\left\{Z_i^{\mathrm{T}} \equiv (S_i X_i, S_i W_i, S_i Y_i, S_i, (1 - S_i)X_i, (1 - S_i)W_i, (1 - S_i)Y_i)\right\}_{i=1}^{n+n_a}$$

such that $\{X_i, W_i, Y_i, S_i = 1\}_{i=1}^n$ is the primary sample and $\{X_i, W_i, Y_i, S_i = 0\}_{i=n+1}^{n+n_a}$ is the auxiliary sample. Denote $p \equiv \Pr(S_i = 1) \in (0, 1)$. Denote $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2$ as the parameter space. The log-joint likelihood for $\alpha \equiv (\theta^{\mathrm{T}}, f_1, f_{1a}, f_2, f_{2a})^{\mathrm{T}} \in \mathcal{A}$ is given by

$$\sum_{i=1}^{n+n_a} \{S_i \log[p \times f(X_i, W_i, Y_i|S_i = 1; \alpha)] + (1 - S_i) \log[(1 - p) \times f(X_i, W_i, Y_i|S_i = 0; \alpha)]\}$$

$$= n \log(p) + n_a \log\{(1 - p)\} + \sum_{i=1}^{n+n_a} \ell(Z_i; \alpha),$$

in which

$$\ell(Z_i;\alpha) \equiv S_i \ell_p(Z_i;\theta, f_1, f_2) + (1-S_i)\ell_a(Z_i; f_{1a}, f_{2a}),$$

$$\ell_p(Z_i;\theta, f_1, f_2) = \log\left\{\int f_1(X_i|x^*) g(Y_i|x^*, W_i;\theta) f_2(x^*|W_i)\,\mathrm{d}x^*\right\} + \log f_W(W_i),$$

$$\ell_a(Z_i; f_{1a}, f_{2a}) = \log\left\{\int f_{1a}(X_i|x_a^*) g(Y_i|x_a^*, W_i;\theta) f_{2a}(x_a^*|W_i)\,\mathrm{d}x_a^*\right\} + \log\{f_{W_a}(W_i)\}.$$

Let $E(\cdot)$ denote the expectation with respect to the underlying true data generating process for Z_i. To stress that our combined data set consists of two samples, sometimes we let $Z_{pi} = (X_i, W_i, Y_i)^{\mathrm{T}}$ denote the ith observation in the primary data set, and $Z_{aj} = (X_{aj}, W_{aj}, Y_{aj})^{\mathrm{T}}$ denote jth observation in the auxiliary data set. Then

$$\alpha_0 = \arg\sup_{\alpha\in\mathcal{A}} E[\ell(Z_i;\alpha)] = \arg\sup_{\alpha\in\mathcal{A}}[pE\{\ell_p(Z_{pi};\theta, f_1, f_2)\} + (1-p)E\{\ell_a(Z_{aj}; f_{1a}, f_{2a})\}].$$

Let $\mathcal{A}_n = \Theta \times \mathcal{F}_1^n \times \mathcal{F}_{1a}^n \times \mathcal{F}_2^n \times \mathcal{F}_2^n$ be a sieve space for $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2$, which is a sequence of approximating spaces that are dense in $\mathcal{A}$ under some pseudo-metric. The two-sample sieve quasi- MLE $\widehat{\alpha}_n = (\widehat{\theta}^{\mathrm{T}}, \widehat{f}_1, \widehat{f}_{1a}, \widehat{f}_2, \widehat{f}_{2a})^{\mathrm{T}} \in \mathcal{A}_n$ for $\alpha_0 \in \mathcal{A}$ is defined as

$$\widehat{\alpha}_n = \arg\max_{\alpha\in\mathcal{A}_n} \sum_{i=1}^{n+n_a} \ell(Z_i;\alpha) = \arg\max_{\alpha\in\mathcal{A}_n} \left[\sum_{i=1}^{n} \ell_p(Z_{pi};\theta, f_1, f_2) + \sum_{j=1}^{n_a} \ell_a(Z_{aj}; f_{1a}, f_{2a})\right].$$

We shall use finite-dimensional sieve spaces since they are easier to implement. For $j = 1, 1a, 2$, let $p_j^{k_{j,n}}(\cdot)$ be a $k_{j,n} \times 1-$vector of known basis functions, such as power series, splines, Fourier series, wavelets, Hermite polynomials, etc. In the simulation study and real data examples we have used linear sieves to directly approximate unknown densities:

$$\mathcal{F}_1^n = \left\{f_1(x|x^*) = p_1^{k_{1,n}}(x, x^*)^{\mathrm{T}}\beta_1 \in \mathcal{F}_1\right\}, \quad \mathcal{F}_{1a}^n = \left\{f_{1a}(x_a|x_a^*) = p_{1a}^{k_{1a,n}}(x_a, x_a^*)^{\mathrm{T}}\beta_{1a} \in \mathcal{F}_{1a}\right\},$$

$$\mathcal{F}_2^n = \left\{f_2(x^*|w) = \sum_{j=1}^{J} I(w = w_j) p_2^{k_{2,n}}(x^*)^{\mathrm{T}}\beta_{2j} \in \mathcal{F}_2\right\},$$

as well as linear sieves to approximate square root of densities:

$$\mathcal{F}_1^n = \{f_1(x|x^*) = [p_1^{k_{1,n}}(x, x^*)^{\mathrm{T}}\beta_1]^2 \in \mathcal{F}_1\},$$

$$\mathcal{F}_{1a}^n = \left\{f_{1a}(x_a|x_a^*) = [p_{1a}^{k_{1a,n}}(x_a, x_a^*)^{\mathrm{T}}\beta_{1a}]^2 \in \mathcal{F}_{1a}\right\},$$

$$\mathcal{F}_2^n = \left\{f_2(x^*|w) = \left[\sum_{j=1}^{J} I(w = w_j) p_2^{k_{2,n}}(x^*)^{\mathrm{T}}\beta_{2j}\right]^2 \in \mathcal{F}_2\right\}.$$

The results of our simulation study and real data examples are not sensitive to these different choices of sieves spaces. See Section 3.5 for detailed discussion of implementation.

Consistency of the nonparametric components: Here we impose some conditions that imply consistency of the sieve estimator $\widehat{\alpha}_n = (\widehat{\theta}^{\mathrm{T}}, \widehat{f}_1, \widehat{f}_{1a}, \widehat{f}_2, \widehat{f}_{2a})^{\mathrm{T}}$.

ASSUMPTION 3.1 (i) *All the assumptions in Theorem* 2.1 *hold*; (ii) $f_{X|X^*}(\cdot|\cdot) \in \mathcal{F}_1$ *with* $\gamma_1 > 1$; (iii) $f_{X_a|X_a^*}(\cdot|\cdot) \in \mathcal{F}_{1a}$ *with* $\gamma_{1a} > 1$; (iv) $f_{X^*|W}(\cdot|w)$, $f_{X_a^*|W_a}(\cdot|w) \in \mathcal{F}_2$ *with* $\gamma_2 > 1/2$ *for all* $w \in \mathcal{W}$.

ASSUMPTION 3.2 (i) $\{X_i, W_i, Y_i\}_{i=1}^n$ *and* $\{X_{aj}, W_{aj}, Y_{aj}\}_{j=1}^{n_a}$ *are i.i.d and independent of each other.* $\lim_{n\to\infty} n/(n+n_a) = p \in (0,1)$; (ii) $g(y|x^*, w; \theta)$ *is continuous in* $\theta \in \Theta$, *and* Θ *is a compact subset of* $\mathbb{R}^{d_\theta}$; (iii) $\theta_0 \in \Theta$ *is the unique maximiser of* $\int [\log g(y|x^*, w; \theta)] f_{Y|X^*,W}(y|x^*, w)\,\mathrm{d}y$ *over* $\theta \in \Theta$.

Define a norm on $\mathcal{A}$ as: $\|\alpha\|_s = \|\theta\|_E + \|f_1\|_{\infty,\omega_1} + \|f_{1a}\|_{\infty,\omega_{1a}} + \|f_2\|_{\infty,\omega_2} + \|f_{2a}\|_{\infty,\omega_2}$ in which $\|h\|_{\infty,\omega_j} \equiv \sup_\xi |h(\xi)\omega_j(\xi)|$ with $\omega_j(\xi) = (1 + \|\xi\|_E^2)^{-\varsigma_j/2}$, $\varsigma_j > 0$ for $j = 1, 1a, 2$. We assume each of $\mathcal{X}$, $\mathcal{X}_a$, $\mathcal{X}^*$ is $\mathbb{R}$, and

ASSUMPTION 3.3 (i) $-\infty < E[\ell(Z_i; \alpha_0)] < \infty$, $E[\ell(Z_i; \alpha)]$ *is upper semicontinuous on* $\mathcal{A}$ *under the metric* $\|\cdot\|_s$; (ii) *there is a finite* $\kappa > 0$ *and a random variable* $U(Z_i)$ *with* $E\{U(Z_i)\} < \infty$ *such that* $\sup_{\alpha\in\mathcal{A}_n:\|\alpha-\alpha_0\|_s\le\delta} |\ell(Z_i; \alpha) - \ell(Z_i; \alpha_0)| \le \delta^\kappa U(Z_i)$.

ASSUMPTION 3.4 (i) $p_2^{k_{2,n}}(\cdot)$ *is a* $k_{2,n} \times 1$ *– vector of spline wavelet basis functions on* $\mathbb{R}$, *and for* $j = 1, 1a$, $p_j^{k_{j,n}}(\cdot,\cdot)$ *is a* $k_{j,n} \times 1$ *– vector of tensor product of spline wavelet basis functions on* $\mathbb{R}^2$; (ii) $\min\{k_{1,n}, k_{1a,n}, k_{2,n}\} \to \infty$ *and* $\max\{k_{1,n}, k_{1a,n}, k_{2,n}\}/n \to 0$.

The following consistency lemma is a direct application of (Chen 2007, Theorem 3.1 (or Remark 3.3)); hence, we omit its proof.

LEMMA 3.1 *Under Assumptions* 3.1–3.4, *we have* $\|\widehat{\alpha}_n - \alpha_0\|_s = o_p(1)$.

3.2. *Asymptotic normality under possible misspecification*

We show that the two-sample sieve quasi MLE $\widehat{\theta}_n$ is asymptotically normally distributed around θ_0, regardless of whether the latent parametric model $g(y|x^*, w; \theta_0)$ is correctly specified or not. The technical details are given in a longer version of the paper available from the first author.

THEOREM 3.1 *Suppose that Assumptions of Lemma* 3.1, *and Assumptions* A.2–A.10 *in the Appendix hold. Then* $\sqrt{n+n_a}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} N(0, V_*^{-1} I_* V_*^{-1})$, *with* V_* *defined in Equation* (A9) *and* I_* *given by Equation* (A11) *in the appendix.*

3.3. *Semiparametric efficiency under correct specification*

When $g(y|x^*, w; \theta_0)$ correctly specifies the true unknown conditional density $f_{Y|X^*,W}(y|x^*, w)$, then $I_* = V_*$ becomes the semiparametric information bound for θ_0, and our above estimator $\widehat{\theta}_n$ becomes semiparametrically efficient for θ_0. Specifically, by combining our Theorem 3.1 and (Shen 1997, Theorem 4), we immediately obtain the following:

THEOREM 3.2 *Suppose that* $g(y|x^*, w; \theta_0) = f_{Y|X^*,W}(y|x^*, w)$, *and that all the assumptions of Theorem 3.1 hold. Then the sieve MLE* $\widehat{\theta}_n$ *is semiparametrically efficient, and* $\sqrt{n+n_a}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} N(0, I_*^{-1})$.

3.4. *Estimation of standard error and confidence region*

There are two ways of consistently estimating the asymptotic covariance matrix of the two-sample sieve Q-MLE $\widehat{\theta}_n$. The first is to use the definitions of V_* and I_* (see Equations (A9) and (A11) in the appendix), but to replace expectations by sample averages and to replace unknown parameters by

their sieve estimates. This is asymptotically equivalent to using the sieve Q-MLE approximation as if it were a parametric model. Applying Theorem 5.1 of (Ai and Chen 2007) gives a consistent estimate of the asymptotic variance of $\widehat{\theta}_n$. Alternatively, applying Theorem B of (Chen, Linton, and Van Keilegom 2003), we know that the standard bootstrap also provides consistent estimates of confidence regions. We implemented both in the real data example.

3.5. *Computation*

There are many ways to compute the sieve estimators, and many ways to parameterise them. In our original implementations in the simulation study and the real data examples, we simply took a finite-dimensional linear sieve basis to directly approximate the densities without imposing constraints. In this version, for the simulation and the first real-data example, we use the linear sieve to approximate squared root of densities and we impose all the constraints. It is nice to see that the results are virtually the same for our sieve MLEs using these two kinds of sieves.

Here is a parameterisation that adheres to the ideas that densities are positive and integrate to one, and also that there is mean targeting. With W discrete, to estimate $f_{X^*|W}(x^*|W)$ we used the approximation $f_{X^*|W}(x^*|W=w) = \{\sum_{k=1}^{k_{2,n}} \gamma_k(w) q_k(x^*)\}^2$, where $q_k(x^*)$ is an orthonormal series with $\int q_k(x) q_j(x)\,\mathrm{d}x = \delta_{jk}$, the Dirac delta function. This result is a density function as long as $\sum_{k=1}^{k_{2,n}} \gamma_k^2(w) = 1$, a restriction that is easily handled. A similar form is used for $f_{X_a^*|W}(x_a^*|W)$.

To estimate $f_{X,X^*}(x, x^*)$, we used the approximation $f_{X,X^*}(x, x^*) = \{\sum_{j=1}^{J_n} \sum_{k=1}^{K_n} \gamma_{jk} p_j(x - x^*) q_k(x^*)\}^2$, where again the series $\{p_j(\cdot)\}$ and $\{q_k(\cdot)\}$ are orthonormal. The result is easily seen to be a density function if $\sum_{j=1}^{J_n} \sum_{k=1}^{K_n} \gamma_{jk}^2 = 1$. This means that $f_{X^*}(x^*) = \sum_{j=1}^{J_n} \sum_{k=1}^{K_n} \sum_{\ell=1}^{K_n} \gamma_{jk}\gamma_{j\ell} q_k(x^*) q_\ell(x^*)$, from which $f_{X|X^*}(x|x^*)$ is readily derived.

More difficult is the targeting Assumption 2.6, that is applied to (X_a, X_a^*). Here we use the same form for the density as for that of (X, X^*), and we consider mean targeting, i.e. $E(X_a|X_a^*) = X_a^*$. Let the Hermite orthogonal series be defined as $H_0(x) = 1$, $H_1(x) = 2x$ and $H_{n=1}(x) = 2xH)n(x) - 2nH_{n-1}(x)$, and define $p_n(x) = H_n(x)\exp(-x^2/2)(2^n n!\pi^{1/2})^{-1/2}$. Then $\{p_n(\cdot)\}$ is an orthogonal series of the type required, with the property that $\int x p_n(s) p_m(x)\,\mathrm{d}x = \{(n+1)/2\}^{1/2} I(m = n+1) + (n/2)^{1/2} I(m = n-1)$. Let $Q = \{q_1(x_a^*), \ldots, q_{K_n}(x_a^*)\}^\top$, $P = \{p_1(x_a - x_a^*), \ldots, p_{J_n}(x_a - x_a^*)\}^\top$, and let $B = (\gamma_{jk})$. Then $f(x_a, x_a^*) = (P^\top B Q)^2 = Q^\top B^\top P P^\top B Q$. We require for mean targeting that, for every x_a^*, $0 = \int (x_a - x_a^*) f(x_a - x_a^*)\,\mathrm{d}x_a$, which means $0 = B^\top \int (x_a - x_a^*) P P^\top\,\mathrm{d}x_a B$. However, the latter is $B^\top S B$, where S has all zeros except that its $(k, k+1)$ and $(k+1, k)$ components equal $(k+1)^{1/2}$ for $k = 1, \ldots, K_n - 1$. The restriction $0 = B^\top S B$ is readily achieved, e.g. for $J_n = 5$, $K_n = 4$, set $B_* = \mathrm{diag}(1, 0, 1, 0, 1)B$, then $B_*^\top S B_* = 0$ by algebra.

In applications, the sieve MLE method needs to choose the order of the sieve terms. Our experience is that the estimation of the finite-dimensional parameters θ are not very sensitive to the order of sieves. Of course if one cares about estimation of nonparametric density functions, then one could apply either the AIC or the generalised cross-validation.

4. Simulation and comparisons

4.1. *The simulation study*

This is the first paper to show nonparametric identifiability in the context of two samples, and the first to derive an estimator of the parametric part with no assumptions made about the distribution

of the measurement error or the latent variable. In this section, we are going to compare five estimators, as follows.

- The naive parametric model that ignores measurement error entirely.
- Our sieve MLE with no assumptions made about the distribution of the measurement error or the latent variable.
- A correctly specified fully parametric model for all components, with a parametric MLE.
- A fully parametric model with the measurement error model misspecified.
- A fully parametric model with the measurement error model misspecified and the latent variable model also misspecified.

The simulation will give some numerical experience into the cost of being nonparametric, and also the gain in robustness for being nonparametric.

The true response model is $f_{Y|X^*,W}(y|x^*, w; \theta_0) = \phi\{y - m(x^*, w; \theta_0)\}$, where $\phi(\cdot)$ is the standard normal density, $\theta = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}}$, $\theta_0 = (1, 1, 1)^{\mathrm{T}}$, and

$$m(x^*, w; \theta) = \theta_1 x^* + \theta_2 x^* w + \theta_3 (x^{*2} w + x^* w^2)/2,$$

in which $w \in \{-1, 0, 1\}$. We have two independent random samples, $\{X_i, W_i, Y_i\}_{i=1}^{n}$ and $\{X_{aj}, W_{aj}, Y_{aj}\}_{j=1}^{n_a}$, with a larger sample size ($n = 1500, n_a = 1000$) and with a smaller sample size ($n = 300, n_a = 200$). In the primary sample, we let $\Pr(W = 1) = \Pr(W = 0) = 1/3$, the unknown true conditional density $f_{X^*|W}$ be the standard normal density $\phi(x^*)$, and the mismeasured value X be $X = 0.1X^* + 0.6\mathrm{e}^{-0.1X^*}\varepsilon$ with $\varepsilon \sim \mathrm{N}(0, 1)$, i.e. multiplicative measurement error. In the auxiliary sample, we generate W_a in the same way as that for W in the primary sample, and the unknown true conditional density $f_{X_a^*|W_a}$ according to

$$f_{X_a^*|W_a}(x_a^*|w_a) = \begin{cases} \phi(x_a^*) & \text{for } w_a = -1, \\ \dfrac{1}{0.5}\phi\left(\dfrac{1}{0.5}x_a^*\right) & \text{for } w_a = 0, \\ \dfrac{1}{0.95}\phi\left(\dfrac{1}{0.95}x_a^* - 0.25\right) & \text{for } w_a = 1. \end{cases}$$

We let the mismeasured value X_a be $X_a = X_a^* + 0.5\mathrm{e}^{-0.1X_a^*}\nu$ with $\nu \sim \mathrm{N}(0, 1)$, which implies that x_a^* is the mode of the conditional density $f_{X_a|X_a^*}(\cdot|x_a^*)$. The simulation was repeated 1000 times.

We used the simple sieve expression $[p_1^{k_{1,n}}(x_1, x_2)^{\mathrm{T}}\beta_1]^2 = [\sum_{j=0}^{J_n}\sum_{k=0}^{K_n}\gamma_{jk}p_j(x_1 - x_2)q_k(x_2)]^2$ to approximate $f_{X|X^*}(x_1|x_2)$ and $f_{X_a|X_a^*}(x_1|x_2)$, with $k_{1,n} = (J_n + 1)(K_n + 1)$, $J_n = 5$, $K_n = 3$. We also use $[p_2^{k_{2,n}}(x^*)^{\mathrm{T}}\beta_2(w)]^2 = [\sum_{k=1}^{k_{2,n}}\gamma_k(w)q_k(x^*)]^2$ to approximate $f_{X^*|W_j=w}$ and $f_{X_a^*|W_j=w}$ with $W_j = -1, 0, 1$ and $k_{2,n} = 4$. The sieve bases $\{p_j(\cdot)\}$ and $\{q_k(\cdot)\}$ are Hermite polynomials bases, and we also impose the integration to one and the mean targeting constraints as described in Section 3.5.

For the parametric model with incorrectly specified measurement error distribution, we computed the parametric MLE when it was assumed that the measurement errors in the primary and auxiliary samples were homoscedastic with standard deviations 0.6079 and 0.6202, respectively. For the parametric model with measurement error and latent variable models misspecified, we did the following. Define $(\gamma, \gamma_{1a}, \gamma_{2a}, \gamma_{3a}, \gamma_{4a}) = (1.0, 1.0, 2.0, 1.05, 0.25)$, and define $\varphi(x) = \exp\{x - \exp(x)\}$. Then set $f_{X^*|W}(X^*|W; \gamma) = \gamma\varphi\{\gamma X^*\}$, and for $W_a = (-1, 0, 1)$, set $f_{X_a^*|W_a}(X_a^*|W_a; \gamma_a) = \gamma_{1a}\varphi(\gamma_{1a}X_a^*)$, $\gamma_{2a}\varphi(\gamma_{2a}X_a^*)$ and $\gamma_{3a}\varphi(\gamma_{3a}X_a^* - \gamma_{4a})$, respectively.

The simulation results are shown in Tables 1 and 2, illustrating larger and smaller sample sizes, and show what one might expect. First, accounting for measurement error matters: the 2-sample sieve MLE has a much smaller bias and MSE than the estimator ignoring measurement error.

Table 1. Simulation results ($n = 1500, n_a = 1000$, reps $= 1000$).

True value of θ	$\theta_1 = 1$	$\theta_2 = 1$	$\theta_3 = 1$
Ignoring measurement error			
Mean estimate	0.179	0.305	0.579
Standard error	0.084	0.118	0.195
Root MSE	0.825	0.705	0.464
2-sample sieve MLE			
Mean estimate	1.021	1.013	1.007
Standard error	0.209	0.182	0.200
Root MSE	0.210	0.182	0.200
Correctly specified parametric model			
Mean estimate	0.921	1.055	0.968
Standard error	0.053	0.068	0.087
Root MSE	0.095	0.088	0.093
Parametric model with misspecified measurement error distribution			
Mean estimate	1.429	1.382	1.427
Standard error	0.138	0.127	0.256
Root MSE	0.451	0.402	0.498
Parametric model with measurement error and latent variable models misspecified			
Mean estimate	1.127	1.472	1.522
Standard error	0.124	0.161	0.296
Root MSE	0.178	0.498	0.600

Table 2. Simulation results ($n = 300, n_a = 200$, reps $= 1000$).

True value of θ	$\theta_1 = 1$	$\theta_2 = 1$	$\theta_3 = 1$
Ignoring measurement error			
Mean estimate	0.181	0.312	0.576
Standard error	0.191	0.269	0.420
Root MSE	0.841	0.739	0.596
2-sample sieve MLE			
Mean estimate	1.017	1.052	1.084
Standard error	0.351	0.278	0.777
Root MSE	0.351	0.283	0.781
Correctly specified parametric model			
Mean estimate	0.966	1.135	1.115
Standard error	0.130	0.156	0.203
Root MSE	0.134	0.206	0.233
Parametric model with misspecified measurement error distribution			
Mean estimate	1.547	1.532	1.659
Standard error	0.270	0.271	0.514
Root MSE	0.610	0.597	0.836
Parametric model with measurement error and latent variable models misspecified			
Mean estimate	1.179	1.601	1.822
Standard error	0.301	0.323	0.596
Root MSE	0.350	0.682	1.015

Second, there are substantial costs for being nonparametric: compared to a correctly specified parametric model, 2-sample sieve MLE is simply more variable, a not very surprising result. Third, there are costs for model misspecification of either the measurement error distribution or the distribution for (X^*, X_a^*).

4.2. *Testing Assumption 2.3*

Huang et al. (2006) develop a method that can allow the testing of assumptions about latent variable distributions in measurement error models. Here we show that a modification of their basic idea is capable of detecting violations of Assumption 2.3.

We performed 500 simulations of the following experiment. For the main data set we had $n = 1000$, $W = (W_1, W_2)$, where $W_1 = \text{Bernoulli}(0.5)$ and $W_2 = \text{Bernoulli}(0.3)$ are independent of one another, $X^* = W_1 + W_2 + \text{N}(0, 0.25)$, $X = X^* + \text{N}(0, 0.25)$ and finally $Y = 1.5X^* + 0.3W_1 + 0.7W_2 + \text{N}(0, 0.01)$. For the auxiliary data set, we had $n_a = 1000$, $W_a = (W_{1a}, W_{2a})$, where $W_{1a} = \text{Bernoulli}(0.3)$ and $W_{2a} = \text{Bernoulli}(0.7)$ are independent of one another, $X^* = 1.5W_1 + 0.5W_2 + \text{N}(0, 0.25)$, $X = X^* + \text{N}(0, 0.25)$ and finally $Y = 0.5X^* + 1.3W_1 + 1.7W_2 + \text{N}(0, 0.01)$. Note that Assumption 2.3 is violated because the regression models are very different.

The fitting method was as follows. We assumed that Assumption 2.3 holds and the homoscedastic model that has $E(Y|X^*, W) = \beta_0 + \beta_1 X^* + \beta_2 W_1 + \beta_3 W_2$. In addition, we assumed that the distribution of X^* given W had different means and variances depending on the four levels of W. The same thing but with different parameters was assumed for the distribution of X_a^* given W_a. We also assumed that the measurement error in X and X_a was additive and homoscedastic but with possibly different variances. Maximum likelihood, which is also method of moments in this context, was used to fit the simulated data sets.

Following Huang et al. (2006), for each of the 500 simulated data sets, we computed a perturbed data set. Specifically, we added to both X and X_a normal random variables with mean zero and variance 0.25, and then we refit the perturbed data. The idea of Huang et al. (2006) is that if we assumed that the model is actually true, then adding additional measurement error will increase variability that will not generate any bias. Conversely, if the assumed model is false, then perturbing the data by adding additional measurement error will cause a bias.

The results were as follows. For β_1, the mean difference between the original and perturbed estimates was -0.16 with a standard deviation 0.02, and thus an effect size of -8.38. For β_2, the mean difference between the original and perturbed estimates was 0.17 with a standard deviation 0.03, and thus an effect size of 5.04. For β_3, the mean difference between the original and perturbed estimates was 0.14 with a standard deviation 0.03, and thus an effect size of 4.75. Clearly, this calculation shows that the Huang et al. (2006) method can detect serious departures from Assumption 2.3.

5. Real data example

5.1. *Background and analysis*

As an illustrative example, we consider two nutritional epidemiology data sets, the eating at America's table study (EATS, Subar et al. (2001)) and the observing protein and energy nutrition study (OPEN, Kipnis et al. (2003)). In both studies, the response Y is the log(1.0 + the amount of beta-carotene from foods as measured by a food frequency questionnaire). In addition, X is the log(1.0 + the amount of beta-carotene from foods as measured by a 24-h recall). We also observed two categorical variables W, namely gender and whether the person was > 50 years of age. Here X^* is the individual's true long-term transformed beta-carotene intake as measured by a hypothetical infinite number of 24-h recalls. The sample sizes for EATS and OPEN were 965 and 481, respectively.

With EATS as the primary study and OPEN as the auxiliary study, the assumption of nondifferential measurement error in the 24 h recalls (Assumptions 2.1(ii) and 2.2(ii))) is standard in this context. Both studies took place in the United States, and thus the stability Assumptions 2.3(i) and

2.6 also seem reasonable. The main difference between EATS and OPEN is that the former was a national study, while the latter took place in the relatively affluent Montgomery County Maryland. Thus, one would expect the distributions of X^* given W and X_a^* given W_a to be different, and of course one would expect that the distribution of true transformed beta-carotene intake will depend on gender and age. Thus, assumption 2.5 seems reasonable in this context. Indeed, for those aged under 50, Wilcoxon rank tests comparing the two transformed 24-h recalls between the two data sets are statistically significant both for men and for women. Within OPEN, the Wilcoxon rank test is also statistically significant when comparing genders or when comparing age categories, while no such differences are observed for EATS. However, in EATS the 24-h recalls for women had more statistically significant variability than those for men, using a Wilcoxon test on the absolute differences from the means.

The data are $\{Y_{ij}, X_{ij}, W_{ij}\}$ for $j = 1, 2$, where $j = 1$ is the primary sample, EATS, and $j = 2$ is the auxiliary sample, OPEN. Here $W_{ij} = (W_{ij1}, W_{ij2})$ is the gender (male = 0) and age ($> 50 = 1$) of the individual. The latent model of interest is

$$Y_{ij} = \theta_4 + \theta_1 X_{ij}^* + \theta_2 W_{ij1} + \theta_3 W_{ij2} + \epsilon_{ij}, \quad X_{ij} = X_{ij}^* + U_{ij}, \tag{2}$$

where ϵ_{ij} is assumed to be independent of the true regressors $(X_{ij}^*, W_{ij1}, W_{ij2})$.

We consider four estimators for $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)^{\mathrm{T}}$.

- The naive OLS estimator with measurement errors ignored.
- A parametric maximum likelihood estimator under the additional Assumptions: $\epsilon_{ij} = \mathrm{N}(0, \sigma_\epsilon^2)$, $U_{ij} = \mathrm{N}(0, \sigma_u^2)$, $X_{i1}^* = a_0 + a_1 W_{ij1} + a_2 W_{ij2} + \nu_{i1}$, and $X_{i2}^* = b_0 + b_1 W_{ij1} + b_2 W_{ij2} + \nu_{i2}$, with $\nu_{ij} = \mathrm{N}(0, \sigma_{\nu,j}^2)$. Note that for this parametric MLE, the measurement error status is assumed to not depend on j.
- The sieve MLE under the additional restriction that the latent model of interest is Equation (2) with $\epsilon_{ij} = \mathrm{Normal}(0, \sigma_\epsilon^2)$.
- The sieve MLE with no assumptions about the distribution of ϵ_{ij}.

To compute the third and the fourth estimators, we use the same set up as in the simulation study. In addition, to approximate $f_\epsilon(\epsilon)$ we used Hermite polynomials with $k_{3,n} = 3$ to compute the fourth sieve MLE.

We also implemented 500 bootstraps by resampling (Y, X, W) within each population. The results are given in Table 3. We see that the measurement errors cause significant attenuation

Table 3. Estimates and Bootstrap analysis of the OPEN and EATS data sets.

	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
	Naive OLS				2-S SMLE w/normal reg. err.			
Estimate	0.242	0.084	0.037	−0.046	0.500	0.076	0.032	0.423
Boot mean	0.242	0.083	0.035	−0.044	0.618	0.080	0.035	0.317
Boot median	0.242	0.083	0.033	−0.043	0.555	0.078	0.037	0.348
Boot s.e.	0.019	0.040	0.040	0.034	0.236	0.042	0.047	0.286
Boot 95% CI	0.204	0.007	−0.039	−0.121	0.254	−0.005	−0.061	−0.287
	0.284	0.161	0.122	0.017	1.176	0.162	0.133	0.846
	Parametric MLE				2-sample sieve MLE			
Estimate	0.461	0.131	−0.019	−0.073	0.780	0.067	−0.024	−0.058
Boot mean	0.485	0.135	−0.027	−0.074	0.714	0.120	0.080	−0.067
Boot median	0.466	0.132	−0.021	−0.073	0.761	0.119	0.078	−0.066
Boot s.e.	0.194	0.061	0.064	0.045	0.312	0.129	0.121	0.082
Boot 95% CI	0.292	0.041	−0.211	−0.181	0.101	−0.115	−0.185	−0.223
	1.179	0.288	0.078	0.002	1.263	0.374	0.328	0.097

in the estimation of θ_1. The corrected estimators have much greater variability than the naive estimator, with variability increasing as assumptions are relaxed.

In Figure 1, we plot the sieve estimated density functions for the measurement error models in EATS and OPEN, as well as the density functions for the latent covariates. The measurement error density estimates are rough, as expected from the deconvolution literature, but they appear somewhat vaguely centered at the true value of the latent variable and clearly depend upon it, the latter being the point of most interest. The latent variable density estimates are easier to visualise because W and W_a have only four levels: there is more skewness in the EATS data than in the OPEN data.

5.2. *Testing Assumption 2.3*

We used the same idea as in Section 4.2 to test whether the distribution of the response given the true covariates is the same in the two samples.

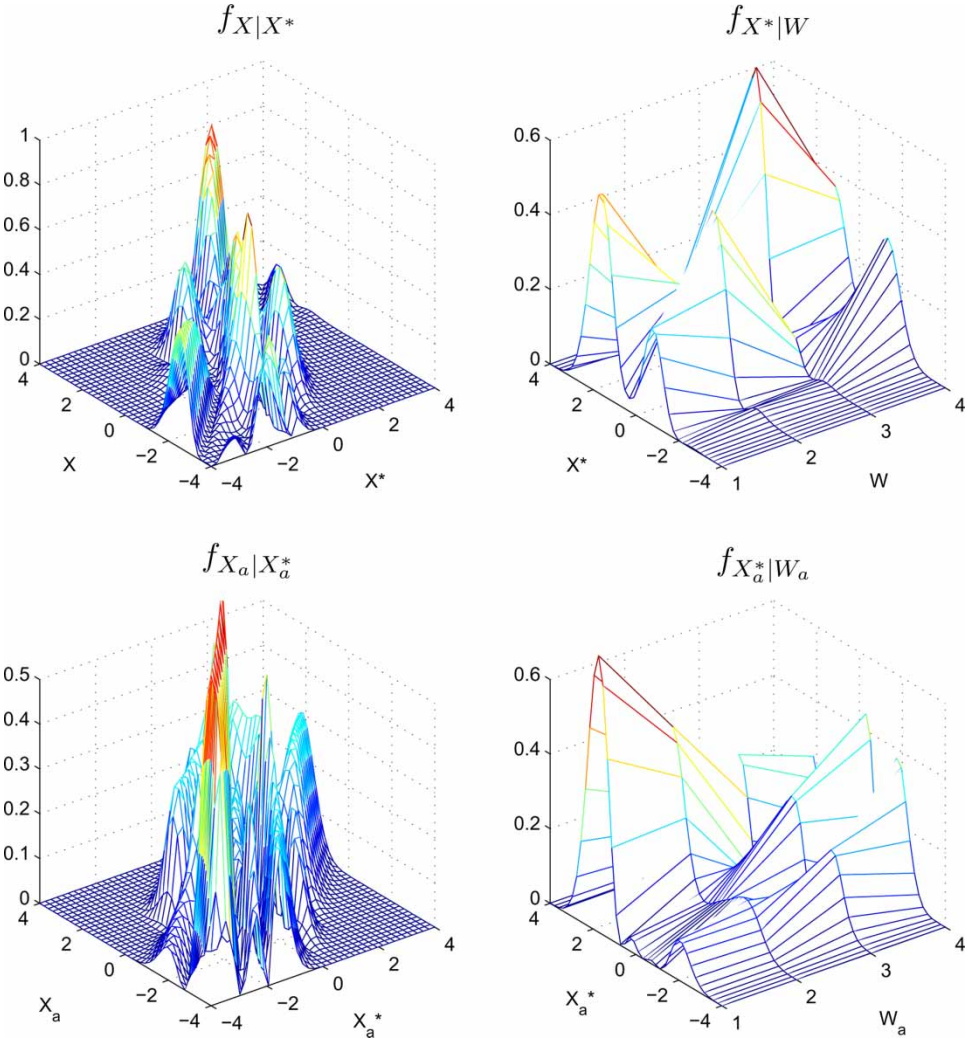

Figure 1. Analysis of the nutrition data set. Left side: sieve estimated measurement error density. Right side: sieve estimated latent variable density.

From our analysis, the measurement error variances for OPEN and EATS were estimated as approximately 0.3 and 0.6. We generated 500 data sets where we replaced X in EATS by $X + \mathrm{N}(0.0, 0.5)$ and we replaced X_a in OPEN by $X_a + \mathrm{N}(0.0, 0.3)$. We fit the same method of moments estimation as in Section 4.2 but applied to these perturbed data sets. If the model assumption fails, then we would expect to see statistically significant bias.

The results were as follows. For β_1, the mean difference between the original and perturbed estimates was 0.035 with a standard deviation 0.035. For β_2, the mean difference between the original and perturbed estimates was 0.01 with a standard deviation 0.014. For β_3, the mean difference between the original and perturbed estimates was 0.01 with a standard deviation 0.015. It appears then that while Assumption 2.3 may be violated in this example, the size of that violation is not likely to be large.

We also refit the data using our sieve-based approach, which makes no assumptions that the measurement errors are homoscedastic, with similar results. That is, in all cases, the mean difference between the original and perturbed estimates were much smaller than the standard deviation of those differences, indicating once again that the evidence that our assumptions are badly violated is weak.

6. Summary

In the absence of knowledge about the measurement error distribution or an instrumental variable such as a replicate, the use of two samples to correct for the effects of measurement error is well established in the literature. One basic assumption in this approach is that the underlying regression function is the same in the two samples. However, all published papers have assumed that the latent variable X^* is measured exactly in one of the two samples. Our paper does not require such validation data, and is thus the first paper to allow estimation in the absence of knowledge about the measurement error distribution, of an instrumental variable and of validation data.

We note two points. First, we have used the terms 'primary' and 'auxiliary samples', but of course these can be interchanged. Second, if W is continuous, our results hold, but different methods of proof are required.

We have provided very general conditions ensuring identifiability: essentially, we require that the distribution of X^* depends on exactly measured covariates, and that this distribution varies in some way across the two data sets.

Finally, in the presence of a parametric regression model, we have provided a sieve quasi-MLE approach to estimation, with the measurement error distribution and the distribution of the latent variable remaining nonparametric. We derived asymptotic theory when the presumed regression model is incorrectly or correctly specified. Simulations and two examples show that our method has good performance despite the generality of the approach.

A long version of this paper is available at http://www.stat.tamu.edu/ftp/pub/rjcarroll/2009.papers.directory/CCH_Long.pdf. It contains more detailed identification results, a second empirical example, and proofs of the asymptotic normality of our sieve estimator.

Acknowledgements

The authors would like to thank the editor, an associate editor, two anonymous referees, P. Cross, S. Donald, E. Mammen, M. Stinchcombe, and conference participants at the 2006 North American Summer Meeting of the Econometric Society and the 2006 Southern Economic Association annual meeting for their valuable suggestions. We thank Arthur Schatzkin, Amy Subar and Victor Kipnis for making the data in our example available to us. Chen acknowledges support from the National Science Foundation (SES-0631613). Carroll's research was supported by grants from the National Cancer Institute (CA57030, CA104620), and partially supported by Award Number KUS-CI-016-04 made by King Abdullah University of Science and Technology (KAUST).

References

Ai, C., and Chen, X. (2007), 'Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables', *Journal of Econometrics*, 141, 5–43.

Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007), 'Convergence Rates of General Regularisation Methods for Statistical Inverse Problems and Applications', *SIAM Journal on Numerical Analysis*, 45, 2610–2636.

Bound, J., Brown, C., and Mathiowetz, N. (2001), 'Measurement Error in Survey Data', in *Handbook of Econometrics* (Vol. 5), eds. J.J. Heckman and E. Leamer, Elsevier Science.

Carroll, R.J., Ruppert, D., Crainiceanu, C., Tosteson, T., and Karagas, R. (2004), 'Nonlinear and Nonparametric Regression and Instrumental Variables', *The 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.), Boca Raton: CRC Press.

Carroll, R.J., and Wand, M.P. (1991), 'Semiparametric Estimation in Logistic Measurement Error Models', *Journal of the Royal Statistical Society: Series B*, 53, 573–585.

Chen, X. (2007), 'Large Sample Sieve Estimation of Semi-nonparametric Models', in *Handbook of Econometrics* (Vol. 6B), eds. J.J. Heckman and E. Leamer, Elsevier Science.

Chen, X., Hong, H., and Nekipelov, D. (2007), 'Nonlinear Models of Measurement Errors', unpublished manuscript.

Chen, X., Hong, H., and Tamer, E. (2005), 'Measurement Error Models with Auxiliary Data', *The Review of Economic Studies*, 72, 343–366.

Chen, X., Linton, O., and Van Keilegom, I. (2003), 'Estimation of Semiparametric Models when the Criterion Function is not Smooth', *Econometrica*, 71, 1591–1608.

Cheng, C.L., and Van Ness, J.W. (1999), *Statistical Regression with Measurement Error*, London: Arnold.

Dunford, N., and Schwartz, J.T. (1971), *Linear Operators, Part* 3: *Spectral Operators*, NewYork: JohnWiley & Sons.

Fan, J. (1991), 'On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems', *Annals of Statistics*, 19, 1257–1272.

Hausman, J., Ichimura, H., Newey, W., and Powell, J. (1991), 'Identification and Estimation of Polynomial Errors-in-variables Models', *Journal of Econometrics*, 50, 273–295.

Hong, H., and Tamer, E. (2003), 'A Simple Estimator for Nonlinear Error in Variable Models', *Journal of Econometrics*, 117, 1–19.

Hu, Y., and Schennach, S.M. (2008), 'Instrumental Variable Treatment of Nonclassical Measurement Error Models', *Econometrica*, 76, 195–216.

Huang, X., Stefanski, L.A., and Davidian, M. (2006), 'Latent-model Robustness in Structural Measurement Error Models', *Biometrika*, 93, 53–64.

Kipnis, V., Subar, A.F., Midthune, D., Freedman, L.S., Ballard-Barbash, R., Troiano, R., Bingham, S., Schoeller, D.A., Schatzkin, A., and Carroll, R.J. (2003), 'The Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study', *American Journal of Epidemiology*, 158, 14–21.

Lee, L.-F., and Sepanski, J.H. (1995), 'Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data', *The Journal of the American Statistical Association*, 90, 429–440.

Li, T. (2002), 'Robust and Consistent Estimation of Nonlinear Errors-in-variables Models', *Journal of Econometrics*, 110, 1–26.

Li, T., and Vuong, Q. (1998), 'Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators', *Journal of Multivariate Analysis*, 65, 139–165.

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

Mattner, L. (1993), 'Some Incomplete but Bounded Complete Location Families', *Annals of Statistics*, 21, 2158–2162.

Shen, X. (1997), 'On Methods of Sieves and Penalization', *Annals of Statistics*, 25, 2555–2591.

Shen, X., and Wong, W. (1994), 'Convergence Rate of Sieve Estimates', *Annals of Statistics*, 22, 580–615.

Subar, A.F., Thompson, F.E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001), 'Comparative Validation of the Block, Willett and National Cancer Institute Food Frequency Questionnaires: The Eating at America's Table Study', *American Journal of Epidemiology*, 154, 1089–1099.

Wang, L. (2004), 'Estimation of Nonlinear Models with Berkson Measurement Errors', *Annals of Statistics*, 32, 2559–2579.

Zinde-Walsh, V. (2007), 'Errors-in-variables Models: A Generalized Functions Approach', Working paper, McGill University and CIREQ.

Appendix A. Mathematical Proofs

A1. *Identification*

Let $\mathcal{L}^2(\mathcal{X})$ denote the space of functions with $\int_{\mathcal{X}} |h(x)|^2 \, \mathrm{d}x < \infty$. Define the integral operator $L_{X|X^*} : \mathcal{L}^2(\mathcal{X}^*) \to \mathcal{L}^2(\mathcal{X})$ as

$$\{L_{X|X^*}h\}(x) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*)h(x^*) \, \mathrm{d}x^* \quad \text{for any } h \in \mathcal{L}^2(\mathcal{X}^*), \quad x \in \mathcal{X}.$$

Denote $W_j = \{w_j\}$ for $j = 1, \ldots, J$ and define the following operators for the primary sample

$$L_{X,Y|W_j} : \mathcal{L}^2(\mathcal{Y}) \to \mathcal{L}^2(\mathcal{X}), \quad (L_{X,Y|W_j}h)(x) = \int f_{X,Y|W}(x, u|w_j)h(u)\,\mathrm{d}u,$$

$$L_{Y|X^*,W_j} : \mathcal{L}^2(\mathcal{Y}) \to \mathcal{L}^2(\mathcal{X}^*), \quad (L_{Y|X^*,W_j}h)(x^*) = \int f_{Y|X^*,W_j}(u|x^*)h(u)\,\mathrm{d}u,$$

$$L_{X^*|W_j} : \mathcal{L}^2(\mathcal{X}^*) \to \mathcal{L}^2(\mathcal{X}^*), \quad (L_{X^*|W_j}h)(x^*) = f_{X^*|W_j}(x^*)h(x^*).$$

Similarly, we may define $L_{Y,X|W_j} : \mathcal{L}^2(\mathcal{X}) \to \mathcal{L}^2(\mathcal{Y})$. We define the operators $L_{X_a|X_a^*} : \mathcal{L}^2(\mathcal{X}^*) \to \mathcal{L}^2(\mathcal{X}_a)$, $L_{X_a,Y_a|W_j} : \mathcal{L}^p(\mathcal{Y}) \to \mathcal{L}^2(\mathcal{X}_a)$, $L_{Y_a|X_a^*,W_j} : \mathcal{L}^2(\mathcal{Y}) \to \mathcal{L}^2(\mathcal{X}^*)$, $L_{X_a^*|W_j} : \mathcal{L}^2(\mathcal{X}^*) \to \mathcal{L}^2(\mathcal{X}^*)$, and $L_{Y_a,X_a|W_j} : \mathcal{L}^2(\mathcal{X}_a) \to \mathcal{L}^2(\mathcal{Y})$ for the auxiliary sample in the same way. Notice that the operators $L_{X^*|W_j}$ and $L_{X_a^*|W_j}$ are diagonal operators, and the operators $L_{X,Y|W_j}$ and $L_{X_a,Y_a|W_j}$ are observed from the data.

ASSUMPTION A.1 (i) $L_{X,Y|W_j}$ *has a right-inverse* (*denoted as* $A = (L_{X,Y|W_j})^{-1}$), *i.e.* $L_{X,Y|W_j}A = I$. (ii) $L_{X_a,Y_a|W_j}$ *has a right-inverse.*

Assumption A.1(i) is equivalent to saying that the adjoint operator of $L_{X,Y|W_j}$ has a left-inverse, i.e. $L_{Y,X|W_j}$ is injective, i.e. the set $\{h \in \mathcal{L}^2(\mathcal{X}) : L_{Y,X|W_j}h = 0\} = \{0\}$.

Proof of Theorem 2.1 Under Assumption 2.1,

$$f_{X,W,Y}(x, w, y) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*) f_{X^*,W,Y}(x^*, w, y)\,\mathrm{d}x^* \quad \text{for all } x, w, y. \tag{A1}$$

For each value w_j of W, Assumptions 2.1–2.3 imply that

$$f_{X,Y|W}(x, y|w_j) = \int f_{X|X^*}(x|x^*) f_{Y|X^*,W}(y|x^*, w_j) f_{X^*|W_j}(x^*)\,\mathrm{d}x^*, \tag{A2}$$

$$f_{X_a,Y_a|W_a}(x, y|w_j) = \int f_{X_a|X_a^*}(x|x^*) f_{Y|X^*,W}(y|x^*, w_j) f_{X_a^*|W_j}(x^*)\,\mathrm{d}x^* \tag{A3}$$

By Equation (A2) and the definition of the operators, we have, for any function $h \in \mathcal{L}^2(\mathcal{Y})$,

$$\begin{aligned}
(L_{X,Y|W_j}h)(x) &= \int f_{X,Y|W_j}(x, u|w_j)h(u)\,\mathrm{d}u \\
&= \int \left(\int f_{X|X^*}(x|x^*) f_{Y|X^*,W}(u|x^*, w_j) f_{X^*|W_j}(x^*)\,\mathrm{d}x^* \right) h(u)\,\mathrm{d}u \\
&= \int f_{X|X^*}(x|x^*) f_{X^*|W_j}(x^*) \left(\int f_{Y|X^*,W}(u|x^*, w_j)h(u)\,\mathrm{d}u \right) \mathrm{d}x^* \\
&= \int f_{X|X^*}(x|x^*) f_{X^*|W_j}(x^*)(L_{Y|X^*,W_j}h)(x^*)\,\mathrm{d}x^* \\
&= \int f_{X|X^*}(x|x^*)(L_{X^*|W_j}L_{Y|X^*,W_j}h)(x^*)\,\mathrm{d}x^* \\
&= (L_{X|X^*}L_{X^*|W_j}L_{Y|X^*,W_j}h)(x).
\end{aligned}$$

This means we have the operator equivalence

$$L_{X,Y|W_j} = L_{X|X^*}L_{X^*|W_j}L_{Y|X^*,W_j} \tag{A4}$$

in the primary sample. Similarly, we have in the auxiliary sample,

$$L_{X_a,Y_a|W_j} = L_{X_a|X_a^*}L_{X_a^*|W_j}L_{Y|X^*,W_j}. \tag{A5}$$

note that the left-hand sides of Equations (A4) and (A5) are observed. Assumptions 2.4 and A.1 imply that all the operators involved in Equations (A4) and (A5) are invertible. Hence

$$L_{X_a,Y_a|W_j}L_{X,Y|W_j}^{-1} = L_{X_a|X_a^*}L_{X_a^*|W_j}L_{X^*|W_j}^{-1}L_{X|X^*}^{-1}. \tag{A6}$$

This equation holds for all W_i and W_j. We may then eliminate $L_{X|X^*}$ to have

$$L_{X_a,X_a}^{ij} \equiv (L_{X_a,Y_a|W_j}L_{X,Y|W_j}^{-1})(L_{X_a,Y_a|W_i}L_{X,Y|W_i}^{-1})^{-1} = L_{X_a|X_a^*}L_{X_a^*}^{ij}L_{X_a|X_a^*}^{-1}. \tag{A7}$$

The operator L_{X_a,X_a}^{ij} on the left-hand side is observed for all i and j. An important observation is that the operator $L_{X_a^*}^{ij} \equiv (L_{X_a^*|W_j}L_{X^*|W_j}^{-1}L_{X^*|W_i}L_{X_a^*|W_i}^{-1}) : \mathcal{L}^2(\mathcal{X}^*) \to \mathcal{L}^2(\mathcal{X}^*)$ is a diagonal operator defined as $(L_{X_a^*}^{ij}h)(x^*) \equiv k_{X_a^*}^{ij}(x^*)h(x^*)$ with

$k^{ij}_{X^*_a}(x^*)$ defined in Equation (1). Equation (A7) implies a diagonalisation of an observed operator $L^{ij}_{X_a,X_a}$. An eigenvalue of $L^{ij}_{X_a,X_a}$ equals$k^{ij}_{X^*_a}(x^*)$ for a value of x^*, which corresponds to an eigenfunction $f_{X_a|X^*_a}(\cdot|x^*)$.

We now show the identification of $f_{X_a|X^*_a}$ and $k^{ij}_{X^*_a}(x^*)$. First, we require the operator $L^{ij}_{X_a,X_a}$ to be bounded so that the diagonal decomposition may be unique (Dunford and Schwartz 1971, Theorem XV.4.3.5, p. 1939). Equation (A7) implies that the operator $L^{ij}_{X_a,X_a}$ has the same spectrum as the diagonal operator $L^{ij}_{X^*_a}$. Since an operator is bounded by the largest element of its spectrum, Assumption 2.5 guarantees that the operator $L^{ij}_{X_a,X_a}$ is bounded. Second, although it implies a diagonalisation of the operator $L^{ij}_{X_a,X_a}$, Equation (A7) does not guarantee distinctive eigenvalues. However, such ambiguity can be eliminated by noting that the observed operators $L^{ij}_{X_a,X_a}$ for all i, j share the same eigenfunctions $f_{X_a|X^*_a}(\cdot|x^*)$. Assumption 2.5 guarantees that, for any two different eigenfunctions $f_{X_a|X^*_a}(\cdot|x^*_1)$ and $f_{X_a|X^*_a}(\cdot|x^*_2)$, one can always find two subsets W_j and W_i such that the two different eigenfunctions correspond to two different eigenvalues $k^{ij}_{X^*_a}(x^*_1)$ and $k^{ij}_{X^*_a}(x^*_2)$ and, therefore, are identified. ■

The third ambiguity is that, for a given value of x^*, an eigenfunction $f_{X_a|X^*_a}(\cdot|x^*)$ times a constant is still an eigenfunction corresponding to x^*. This ambiguity is eliminated by noting that $\int f_{X_a|X^*_a}(x|x^*)\,\mathrm{d}x = 1$ for all x^*.

Fourth, in order to fully identify each eigenfunction, i.e. $f_{X_a|X^*_a}$, we need to identify the exact value of x^* in each eigenfunction $f_{X_a|X^*_a}(\cdot|x^*)$. However, note that assumption 2.6 identifies the exact value of x^* for each eigenfunction $f_{X_a|X^*_a}(\cdot|x^*)$.

After fully identifying the density function $f_{X_a|X^*_a}$, we now show that the density of interest $f_{Y|X^*,W}$ and $f_{X|X^*}$ are also identified. By Equation (A3), we have $f_{X_a,Y_a|W_a} = L_{X_a|X^*_a} f_{Y_a,X^*_a|W_a}$. By the injectivity of operator $L_{X_a|X^*_a}$, the joint density $f_{Y_a,X^*_a|W_a}$ may be identified as follows: $f_{Y_a,X^*_a|W_a} = L^{-1}_{X_a|X^*_a} f_{X_a,Y_a|W_a}$. Assumption 2.3 implies that $f_{Y_a|X^*_a,W_a} = f_{Y|X^*,W}$ so that we may identify $f_{Y|X^*,W}$ through

$$f_{Y|X^*,W}(y|x^*, w) = \frac{f_{Y_a,X^*_a|W_a}(y, x^*|w)}{\int f_{Y_a,X^*_a|W_a}(y, x^*|w)\,\mathrm{d}y} \quad \text{for all } x^* \text{ and } w.$$

By equation (A4) and the injectivity of the identified operator $L_{Y|X^*,W_j}$, we have

$$L_{X|X^*} L_{X^*|W_j} = L_{X,Y|W_j} L^{-1}_{Y|X^*,W_j}. \tag{A8}$$

The left-hand side of Equation (A8) equals an operator with the kernel function $f_{X,X^*|W=w_j} \equiv f_{X|X^*} f_{X^*|W=w_j}$. Since the right-hand side of Equation (A8) has been identified, the kernel $f_{X,X^*|W=w_j}$ on the left-hand side is also identified. We may then identify $f_{X|X^*}$ through

$$f_{X|X^*}(x|x^*) = \frac{f_{X,X^*|W=w_j}(x, x^*)}{\int f_{X,X^*|W=w_j}(x, x^*)\,\mathrm{d}x} \quad \text{for all } x^* \in \mathcal{X}^*.$$

A2. *Conditions and asymptotic normality of sieve Q-MLE*

A2.1. *Rates of convergence*

Given consistency Lemma 3.1, we can restrict our attention to a shrinking $\|\cdot\|_s$-neighborhood around α_0. Let $\mathcal{A}_{0s} \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_s = o(1), \|\alpha\|_s \le c_0 < c\}$ and $\mathcal{A}_{0sn} \equiv \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_0\|_s = o(1), \|\alpha\|_s \le c_0 < c\}$. We assume that both $\mathcal{A}_{0s}$ and $\mathcal{A}_{0sn}$ are convex parameter spaces, and that $\ell(Z_i; \alpha + \tau v)$ is twice continuously differentiable at $\tau = 0$ for almost all Z_i and any direction $v \in \mathcal{A}_{0s}$.

We define the pathwise first and second derivatives of the sieve loglikelihood in the direction v as

$$\frac{\mathrm{d}\ell(Z_i; \alpha)}{\mathrm{d}\alpha}[v] \equiv \left.\frac{\mathrm{d}\ell(Z_i; \alpha + \tau v)}{d\tau}\right|_{\tau=0}; \quad \frac{\mathrm{d}^2\ell(Z_i; \alpha)}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[v, v] \equiv \left.\frac{\mathrm{d}^2\ell(Z_i; \alpha + \tau v)}{d\tau^2}\right|_{\tau=0}.$$

Following Ai and Chen (2007), for any $\alpha_1, \alpha_2 \in \mathcal{A}_{0s}$, we define a pseudo metric $\|\cdot\|_2$ as

$$\|\alpha_1 - \alpha_2\|_2 \equiv \sqrt{-E\left(\frac{\mathrm{d}^2\ell(Z_i; \alpha_0)}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[\alpha_1 - \alpha_2, \alpha_1 - \alpha_2]\right)}.$$

We show that $\widehat{\alpha}_n$ converges to α_0 at a rate faster than $n^{-1/4}$ under the pseudo metric $\|\cdot\|_2$ and the following assumptions:

ASSUMPTION A.2 (i) $\varsigma_j > \gamma_j$ *for* $j = 1, 1a, 2$; (ii) $\max\{k_{1,n}^{-\gamma_1/2}, k_{1a,n}^{-\gamma_{1a}/2}, k_{2,n}^{-\gamma_2}\} = o([n + n_a]^{-1/4})$.

ASSUMPTION A.3 (i) *$\mathcal{A}_{0s}$ is convex at α_0 and $\theta_0 \in int(\Theta)$*; (ii) *$\ell(Z_i;\alpha)$ is twice continuously pathwise differentiable with respect to $\alpha \in \mathcal{A}_{0s}$, and* $\log g(y|x^*, w;\theta)$ *is twice continuously differentiable at θ_0.*

ASSUMPTION A.4

$$\sup_{\widetilde{\alpha}\in\mathcal{A}_{0s}} \sup_{\alpha\in\mathcal{A}_{0sn}} \left| \frac{\mathrm{d}\ell(Z_i;\widetilde{\alpha})}{\mathrm{d}\alpha} \left[\frac{\alpha-\alpha_0}{\|\alpha-\alpha_0\|_s} \right] \right| \le U(Z_i)$$

for a random variable $U(Z_i)$ with $E\{[U(Z_i)]^2\} < \infty$.

ASSUMPTION A.5

(i) $$\sup_{v\in\mathcal{A}_{0s}:\|v\|_s=1} -E\left(\frac{\mathrm{d}^2\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[v,v] \right) \le C < \infty;$$

(ii) *uniformly over $\widetilde{\alpha} \in \mathcal{A}_{0s}$ and $\alpha \in \mathcal{A}_{0sn}$, we have*

$$-E\left(\frac{\mathrm{d}^2\ell(Z_i;\widetilde{\alpha})}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[\alpha-\alpha_0,\alpha-\alpha_0] \right) = \|\alpha-\alpha_0\|_2^2 \times \{1+o(1)\}.$$

The assumptions are straightforward and standard. The following convergence rate theorem is a direct application of (Shen and Wong 1994, Theorem 3.2) to the local parameter space $\mathcal{A}_{0s}$ and the local sieve space $\mathcal{A}_{0sn}$; hence, we omit its proof.

THEOREM A.1 *Let $\gamma \equiv \min\{\gamma_1/2, \gamma_{1a}/2, \gamma_2\} > 1/2$. Under assumptions 3.1–A.5, if $k_{1,n} = O([n+n_a]^{1/(\gamma_1+1)})$, $k_{1a,n} = O([n+n_a]^{1/(\gamma_{1a}+1)})$, and $k_{2,n} = O([n+n_a]^{1/(2\gamma_2+1)})$, then*

$$\|\widehat{\alpha}_n - \alpha_0\|_2 = O_P([n+n_a]^{-\gamma/(2\gamma+1)}) = o_P([n+n_a]^{-1/4}).$$

A2.2. *Conditions for asymptotic normality*

We also define an inner product corresponding to the pseudo metric $\|\cdot\|_2$:

$$\langle v_1, v_2\rangle_2 \equiv -E\left[\frac{\mathrm{d}^2\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[v_1,v_2] \right],$$

where

$$\frac{\mathrm{d}^2\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[v_1,v_2] \equiv \frac{\mathrm{d}^2\ell(Z_i;\alpha_0+\tau_1v_1+\tau_2v_2)}{\mathrm{d}\tau_1\,\mathrm{d}\tau_2}|_{\tau_1=\tau_2=0}.$$

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the metric $\|\cdot\|_2$. Then $(\overline{\mathbf{V}}, \|\cdot\|_2)$ is a Hilbert space and we can represent $\overline{\mathbf{V}} = \mathbb{R}^{d_\theta} \times \overline{\mathcal{U}}$ with $\overline{\mathcal{U}} \equiv \overline{\mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2} - \{(f_{01}, f_{01a}, f_{02}, f_{02a})\}$. Let $h = (f_1, f_{1a}, f_2, f_{2a})$ denote all the unknown densities. The pathwise first derivative can be written as

$$\begin{aligned}\frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha}[\alpha-\alpha_0] &= \frac{\mathrm{d}\ell(Z_i;\alpha_0)}{d\theta^{\mathrm{T}}}(\theta-\theta_0) + \frac{\mathrm{d}\ell(Z;\alpha_0)}{\mathrm{d}h}[h-h_0] \\ &= \left(\frac{\mathrm{d}\ell(Z_i;\alpha_0)}{d\theta^{\mathrm{T}}} - \frac{\mathrm{d}\ell(Z;\alpha_0)}{\mathrm{d}h}[\mu] \right)(\theta-\theta_0),\end{aligned}$$

with $h - h_0 \equiv -\mu \times (\theta-\theta_0)$, and in which

$$\begin{aligned}\frac{\mathrm{d}\ell(Z;\alpha_0)}{\mathrm{d}h}[h-h_0] &= \frac{\mathrm{d}\ell(Z;\theta_0, h_0(1-\tau)+\tau h)}{d\tau}|_{\tau=0} \\ &= \frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_1}[f_1-f_{01}] + \frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_{1a}}[f_{1a}-f_{01a}] \\ &\quad + \frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_2}[f_2-f_{02}] + \frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_{2a}}[f_{2a}-f_{02a}].\end{aligned}$$

Note that

$$E\left(\frac{\mathrm{d}^2\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[\alpha-\alpha_0,\alpha-\alpha_0] \right) = (\theta-\theta_0)^{\mathrm{T}} E\left(\frac{\mathrm{d}^2\ell(Z_i;\alpha_0)}{d\theta d\theta^{\mathrm{T}}} - 2\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{d\theta\,\mathrm{d}h^{\mathrm{T}}}[\mu] + \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu,\mu] \right)(\theta-\theta_0),$$

with $h - h_0 \equiv -\mu \times (\theta - \theta_0)$, and in which

$$\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{d\theta \mathrm{d}h^{\mathrm{T}}}[h-h_0] = \frac{d(\partial\ell(Z;\theta_0, h_0(1-\tau)+\tau h)/\partial\theta)}{d\tau}|_{\tau=0},$$
$$\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[h-h_0, h-h_0] = \frac{\mathrm{d}^2\ell(Z;\theta_0, h_0(1-\tau)+\tau h)}{d\tau^2}|_{\tau=0}.$$

For each component θ^k (of θ), $k = 1, \ldots, d_\theta$, suppose there exists a $\mu^{*k} \in \overline{\mathcal{U}}$ that solves:

$$\mu^{*k} : \inf_{\mu^k \in \overline{\mathcal{U}}} E\left\{-\left(\frac{\partial^2\ell(Z;\alpha_0)}{\partial\theta^k\partial\theta^k} - 2\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\partial\theta^k \mathrm{d}h^{\mathrm{T}}}[\mu^k] + \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu^k,\mu^k]\right)\right\}.$$

Denote $\mu^* = (\mu^{*1}, \mu^{*2}, \ldots, \mu^{*d_\theta})$ with each $\mu^{*k} \in \overline{\mathcal{U}}$, and

$$\frac{\mathrm{d}\ell(Z;\alpha_0)}{\mathrm{d}h}[\mu^*] = \left(\frac{\mathrm{d}\ell(Z;\alpha_0)}{\mathrm{d}h}[\mu^{*1}], \ldots, \frac{\mathrm{d}\ell(Z;\alpha_0)}{\mathrm{d}h}[\mu^{*d_\theta}]\right),$$
$$\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\partial\theta \mathrm{d}h^{\mathrm{T}}}[\mu^*] = \left(\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\partial\theta \mathrm{d}h}[\mu^{*1}], \ldots, \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\partial\theta \mathrm{d}h}[\mu^{*d_\theta}]\right),$$
$$\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu^*,\mu^*] = \begin{pmatrix} \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu^{*1},\mu^{*1}] & \cdots & \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu^{*1},\mu^{*d_\theta}] \\ \cdots & \cdots & \cdots \\ \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu^{*d_\theta},\mu^{*1}] & \cdots & \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu^{*d_\theta},\mu^{*d_\theta}] \end{pmatrix}.$$

Also denote

$$V_* \equiv -E\left(\frac{\partial^2\ell(Z;\alpha_0)}{\partial\theta\partial\theta^{\mathrm{T}}} - 2\frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\partial\theta \mathrm{d}h^{\mathrm{T}}}[\mu^*] + \frac{\mathrm{d}^2\ell(Z;\alpha_0)}{\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}}}[\mu^*,\mu^*]\right). \tag{A9}$$

Now we consider a linear functional of α, which is $\lambda^{\mathrm{T}}\theta$ for any $\lambda \in \mathbb{R}^{d_\theta}$ with $\lambda \neq 0$. Since

$$\sup_{\alpha-\alpha_0\neq 0}\frac{|\lambda^{\mathrm{T}}(\theta-\theta_0)|^2}{\|\alpha-\alpha_0\|_2^2} = \sup_{\theta\neq\theta_0,\mu\neq 0}\frac{(\theta-\theta_0)^{\mathrm{T}}\lambda\lambda^{\mathrm{T}}(\theta-\theta_0)}{\begin{array}{l}(\theta-\theta_0)^{\mathrm{T}}E\{-((\mathrm{d}^2\ell(Z_i;\alpha_0))/(d\theta d\theta^{\mathrm{T}}) - 2(\mathrm{d}^2\ell(Z;\alpha_0))/(d\theta \mathrm{d}h^{\mathrm{T}})[\mu] \\ \qquad +(\mathrm{d}^2\ell(Z;\alpha_0))/(\mathrm{d}h\,\mathrm{d}h^{\mathrm{T}})[\mu,\mu])\}(\theta-\theta_0)\end{array}}$$
$$= \lambda^{\mathrm{T}}(V_*)^{-1}\lambda,$$

the functional $\lambda^{\mathrm{T}}(\theta-\theta_0)$ is *bounded* if and only if the matrix V_* is nonsingular.

Suppose that V_* is nonsingular. For any fixed $\lambda \neq 0$, denote $\upsilon^* \equiv (\upsilon_\theta^*, \upsilon_h^*)$ with $\upsilon_\theta^* \equiv (V_*)^{-1}\lambda$ and $\upsilon_h^* \equiv -\mu^* \times \upsilon_\theta^*$. Then the Riesz representation theorem implies: $\lambda^{\mathrm{T}}(\theta-\theta_0) = \langle \upsilon^*, \alpha-\alpha_0\rangle_2$ for all $\alpha \in \mathcal{A}$. In the longer version of this paper, we establish the following:

$$\lambda^{\mathrm{T}}(\widehat{\theta}_n - \theta_0) = \langle \upsilon^*, \widehat{\alpha}_n - \alpha_0\rangle_2 = \frac{1}{n+n_a}\sum_{i=1}^{n+n_a}\frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha}[\upsilon^*] + o_p\{(n+n_a)^{-1/2}\}. \tag{A10}$$

Denote $\mathcal{N}_0 = \{\alpha \in \mathcal{A}_{0s} : \|\alpha-\alpha_0\|_2 = o([n+n_a]^{-1/4})\}$ and $\mathcal{N}_{0n} = \{\alpha \in \mathcal{A}_{0sn} : \|\alpha-\alpha_0\|_2 = o([n+n_a]^{-1/4})\}$. We impose the following additional conditions for asymptotic normality of sieve quasi MLE $\widehat{\theta}_n$:

ASSUMPTION A.6 *μ^* exists (i.e. $\mu^{*k} \in \overline{\mathcal{U}}$ for $k = 1, \ldots, d_\theta$), and V_* is positive-definite.*

ASSUMPTION A.7 *There is a $\upsilon_n^* \in \mathcal{A}_n - \{\alpha_0\}$, such that $\|\upsilon_n^* - \upsilon^*\|_2 = o(1)$ and $\|\upsilon_n^* - \upsilon^*\|_2 \times \|\widehat{\alpha}_n - \alpha_0\|_2 = o_P(1/\sqrt{n+n_a})$.*

ASSUMPTION A.8 *There is a random variable $U(Z_i)$ with $E\{[U(Z_i)]^2\} < \infty$ and a non-negative measurable function η with $\lim_{\delta\to 0}\eta(\delta) = 0$, such that, for all $\alpha \in \mathcal{N}_{0n}$,*

$$\sup_{\overline{\alpha}\in\mathcal{N}_0}\left|\frac{\mathrm{d}^2\ell(Z_i;\overline{\alpha})}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[\alpha-\alpha_0, \upsilon_n^*]\right| \leq U(Z_i)\times\eta(\|\alpha-\alpha_0\|_s).$$

ASSUMPTION A.9 *Uniformly over* $\overline{\alpha} \in \mathcal{N}_0$ *and* $\alpha \in \mathcal{N}_{0n}$,

$$E\left(\frac{\mathrm{d}^2\ell(Z_i;\overline{\alpha})}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[\alpha-\alpha_0,\upsilon_n^*]-\frac{\mathrm{d}^2\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha\mathrm{d}\alpha^{\mathrm{T}}}[\alpha-\alpha_0,\upsilon_n^*]\right)=o\left(\frac{1}{\sqrt{n+n_a}}\right).$$

ASSUMPTION A.10 $E\{((\mathrm{d}\ell(Z_i;\alpha_0))/(\mathrm{d}\alpha)[\upsilon_n^*-\upsilon^*])^2\}$ *goes to zero as* $\|\upsilon_n^*-\upsilon^*\|_2$ *goes to zero.*

Assumption A.10 is automatically satisfied when the latent parametric model is correctly specified. Recall the definitions of Fisher inner product and the Fisher norm:

$$\langle v_1,v_2\rangle\equiv E\left\{\left(\frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha}[v_1]\right)\left(\frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}\alpha}[v_2]\right)\right\},\quad \|v\|\equiv\sqrt{\langle v,v\rangle}.$$

Under correct specification, $g(y|x^*,w;\theta_0)=f_{Y|X^*,W}(y|x^*,w)$, it can be shown that $\|v\|=\|v\|_2$ and $\langle v_1,v_2\rangle=\langle v_1,v_2\rangle_2$. Thus, the space $\overline{\mathbf{V}}$ is also the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the Fisher metric $\|\cdot\|$.

Suppose that θ has d_θ components, and write its kth component as θ^k. Write $\mu^*=(\mu^{*1},\mu^{*2},\ldots,\mu^{*d_\theta})$, where we compute $\mu^{*k}\equiv(\mu_1^{*k},\mu_{1a}^{*k},\mu_2^{*k},\mu_{2a}^{*k})^{\mathrm{T}}\in\overline{\mathcal{U}}$ as the solution to

$$\begin{aligned}&\inf_{\mu^k\in\overline{\mathcal{U}}}E\left\{\left(\frac{\mathrm{d}\ell(Z_i;\alpha_0)}{d\theta^k}-\frac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}h}[\mu^k]\right)^2\right\}\\&\quad=\inf_{(\mu_1,\mu_{1a},\mu_2,\mu_{2a})^{\mathrm{T}}\in\overline{\mathcal{U}}}E\left\{\left(\begin{array}{c}\dfrac{\mathrm{d}\ell(Z_i;\alpha_0)}{d\theta^k}-\dfrac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_1}[\mu_1]-\dfrac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_{1a}}[\mu_{1a}]\\-\dfrac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_2}[\mu_2]-\dfrac{\mathrm{d}\ell(Z_i;\alpha_0)}{\mathrm{d}f_{2a}}[\mu_{2a}]\end{array}\right)^2\right\}.\end{aligned}$$

Implicitly, this defines $(\mathrm{d}\ell(Z_i;\alpha_0))/\mathrm{d}h[\mu^*]$. Then $\mathcal{S}_{\theta_0}\equiv(\mathrm{d}\ell(Z_i;\alpha_0))/d\theta^{\mathrm{T}}-(\mathrm{d}\ell(Z_i;\alpha_0))/\mathrm{d}h[\mu^*]$ becomes the semiparametric efficient score for θ_0, and

$$I_*\equiv E[\mathcal{S}_{\theta_0}^{\mathrm{T}}\mathcal{S}_{\theta_0}]=V_* \tag{A11}$$

becomes the semiparametric information bound for θ_0.

We refer readers to the longer version for the proof of the asymptotic normality Theorem 3.1.