Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors*

Raymond Carroll† Texas A&M University

Xiaohong Chen‡ Yale University

Yingyao Hu§ Johns Hopkins University

First version: February 2006; This version: October 2007

Abstract

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-ridden 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 latent 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 latent nonlinear model is parameterized, 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. A small Monte Carlo simulation and a real data application are presented.

KEY WORDS: Data combination; Measurement error; Misspecified parametric latent model; Nonclassical measurement error; Nonlinear errors-in-variables model; Nonparametric identification; Sieve quasi likelihood.

*The authors would like to thank 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 valuable suggestions. Chen acknowledges partial support from the National Science Foundation.

†Department of Statistics, Texas A&M University. Tel: 979-845-3141. E-mail: carroll@stat.tamu.edu

‡Department of Economics, Yale University. Tel: 203-432-5852. E-mail: xiaohong.chen@yale.edu

§Department of Economics, Johns Hopkins University. Tel: 410-516-7610. E-mail: yhu@jhu.edu.

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) and Carroll, et al. (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, e.g., Bound, et al. (2001). This motivates many recent studies on Errors-In-Variables (EIV) problems allowing for nonclassical measurement errors. In this paper, we provide one solution to the nonparametric identification of a general nonlinear EIV model by combining two samples, where both samples contain mismeasured covariates and neither contains an accurate measurement of the latent true variable. Our identification strategy does not require the existence of instrumental variables or repeated measurements, both samples could have nonclassical measurement errors and the two samples could be arbitrarily correlated.

There are currently three broad approaches to identification of general nonlinear EIV models. The first one is to impose parametric restrictions on measurement error distributions, see, e.g., Fan (1991), Wang, et al. (1998), Liang, et al. (1999) and Hong and Tamer (2003), and others. The second approach is to assume the existence of Instrumental Variables (IVs), such as a repeated measurement of the mismeasured covariates, that do not enter the latent model of interest but do contain information to recover features of latent true variables, see, e.g., Carroll and Stefanski (1990), Hausman, et al. (1991), Wang and Hsiao (1995), Buzas and Stefanski (1996), Li and Vuong (1998), Li (2002), Wang (2004), Schennach (2004), Carroll, et al. (2004), Hu (2006) and Hu and Schennach (2007). The third approach to identifying nonlinear EIV models with nonclassical errors is to combine

two samples, see, e.g., Carroll and Wand (1991), Lee and Sepanski (1995), Chen, et al. (2005), Chen, et al. (2007), and Hu and Ridder (2006).

The approach of combining samples has the advantages of allowing for arbitrary measurement errors in the primary sample, without the need of finding IVs or imposing parametric assumptions on measurement error distributions. However, all the currently published papers using this approach require that the auxiliary sample contain an accurate measurement of the true value; such a sample might be difficult to find in some applications.

In this paper, we provide nonparametric identification of a general nonlinear EIV model with measurement errors in covariates by combining a primary sample and an auxiliary sample, in which each sample contains only one measurement of the error-ridden explanatory variable, and the errors in both samples may be nonclassical. Our approach differs from the IV approach in that 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. Our approach is closer to the existing two-sample approach, since we also require an auxiliary sample and allow for nonclassical measurement errors in both samples. However, our identification strategy differs crucially from the existing two-sample approach in that neither of our samples contains an accurate measurement of the latent true variable.

We assume that both samples consist of a dependent variable (Y), some error-free covariates (W), and an error-ridden covariate (X), in which the measurement error has unknown distribution and could be arbitrarily correlated with the latent true values (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. These contrasting subsamples of the primary and the auxiliary samples may be different geographic areas, age groups, or other observed demographic characteristics. We use the difference between the distributions of the latent true values in the contrasting sub-

samples of both samples to show that the measurement error distributions are identified. In supplementary material available from the first author, we first illustrate our identification strategy using a nonlinear EIV model with nonclassical errors in discrete covariates of two samples. The main paper however focuses on nonparametric identification of a general latent nonlinear model with arbitrary measurement errors in continuous covariates.

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. Within this framework, we propose a sieve Quasi-Maximum Likelihood Estimation (Q-MLE) for the latent nonlinear model of interest using two samples with nonclassical measurement errors. Under possible misspecification of the latent parametric model, we establish root-n consistency and asymptotic normality of the sieve Q-MLE of the finite dimensional parameter of interest, as well as its semiparametric efficiency under correct specification.

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. The primary sample is a random sample from (X, W, Y), in which X is a mismeasured X^*; and the 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. Section 2 establishes the nonparametric identification of the latent probability 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 an empirical illustration. The Appendix contains technical arguments.

2 Nonparametric Identification

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 continuous regressor subject to a

possibly nonclassical measurement error, and W is an accurately measured discrete covariate. For example, the discrete covariate W may stand for subpopulations with different demographic characteristics, such as marital status, race, gender, profession, and geographic location. Suppose the supports of X, W, Y, and X^* are $\mathcal{X} \subseteq \mathbb{R}$, $\mathcal{W} = \{w_1, w_2, ..., w_J\}$, $\mathcal{Y} \subseteq \mathbb{R}$, and $\mathcal{X}^* \subseteq \mathbb{R}$, respectively. We assume

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 2.1(ii) implies that the measurement error in X is independent of all other variables in the model conditional on the true value X^*. The measurement error in X may still be correlated with the true value X^* in an arbitrary way, and hence is nonclassical.

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 $\mathcal{X}_a \times \mathcal{X}^* \times \mathcal{W} \times \mathcal{Y}$.*

The next condition requires that the latent structural probability model is the same in both samples, which is a reasonable stability assumption.

Assumption 2.3. *$f_{Y_a|X_a^*,W_a}(y|x^*, w) = f_{Y|X^*,W}(y|x^*, w)$ on $\mathcal{Y} \times \mathcal{X}^* \times \mathcal{W}$.*

Let $\mathcal{L}^p(\mathcal{X})$, $1 \leq p < \infty$ denote the space of functions with $\int_{\mathcal{X}} |h(x)|^p dx < \infty$, and let $\mathcal{L}^\infty(\mathcal{X})$ be the space of functions with $\sup_{x \in \mathcal{X}} |h(x)| < \infty$. For any $1 \leq p \leq \infty$, define the integral operator $L_{X|X^*} : \mathcal{L}^p(\mathcal{X}^*) \to \mathcal{L}^p(\mathcal{X})$ as:

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

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

$$\begin{aligned}
L_{X,Y|W_j} &: \mathcal{L}^p(\mathcal{Y}) \to \mathcal{L}^p(\mathcal{X}), \quad \left(L_{X,Y|W_j}h\right)(x) = \int f_{X,Y|W}(x, u|w_j) h(u) du, \\
L_{Y|X^*,W_j} &: \mathcal{L}^p(\mathcal{Y}) \to \mathcal{L}^p(\mathcal{X}^*), \quad \left(L_{Y|X^*,W_j}h\right)(x^*) = \int f_{Y|X^*,W_j}(u|x^*) h(u) du, \\
L_{X^*|W_j} &: \mathcal{L}^p(\mathcal{X}^*) \to \mathcal{L}^p(\mathcal{X}^*), \quad \left(L_{X^*|W_j}h\right)(x^*) = f_{X^*|W_j}(x^*) h(x^*).
\end{aligned}$$

We define the operators $L_{X_a|X_a^*} : \mathcal{L}^p(\mathcal{X}^*) \to \mathcal{L}^p(\mathcal{X}_a)$, $L_{X_a,Y_a|W_j} : \mathcal{L}^p(\mathcal{Y}) \to \mathcal{L}^p(\mathcal{X}_a)$, $L_{Y_a|X_a^*,W_j} : \mathcal{L}^p(\mathcal{Y}) \to \mathcal{L}^p(\mathcal{X}^*)$, and $L_{X_a^*|W_j} : \mathcal{L}^p(\mathcal{X}^*) \to \mathcal{L}^p(\mathcal{X}^*)$ for the auxiliary sample in the same way as their counterparts for the primary sample. Notice that the operators $L_{X^*|W_j}$ and $L_{X_a^*|W_j}$ are diagonal operators.

Assumptions 2.1, 2.2 and 2.3 imply that $L_{X,Y|W_j} = L_{X|X^*} L_{X^*|W_j} L_{Y|X^*,W_j}$ and $L_{X_a,Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{Y|X^*,W_j}$, where the operators $L_{X,Y|W_j}$ and $L_{X_a,Y_a|W_j}$ are observed given the data. We assume

Assumption 2.4. *(i) $L_{X_a|X_a^*}$ is injective, i.e., the set $\{h \in \mathcal{L}^p(\mathcal{X}^*) : L_{X_a|X_a^*} h = 0\} = \{0\}$; (ii) $L_{X,Y|W_j}$ and $L_{X_a,Y_a|W_j}$ are injective.*

Assumption 2.4(i) is commonly imposed in general deconvolution problems; see, e.g., Bissantz, Hohage, Munk and Ruymgaart (2007). Assumption 2.4(i) is the same as the *completeness* of the conditional density $f_{X_a^*|X_a}$, which is satisfied, for example, when $f_{X_a^*|X_a}$ belongs to an exponential family (see, e.g., Newey and Powell, 2003). Moreover, if we are willing to assume $\sup_{x^*,w} f_{X_a^*,W_a}(x^*, w) \leq c < \infty$, then a sufficient condition for Assumption 2.4(i) is the *bounded completeness* of the conditional density $f_{X_a^*|X_a}$; see, e.g., Lehmann and Romano (2005, page 118) and 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^*.

Assumption 2.4 implies that $L_{Y|X^*,W_j}$ and $L_{X|X^*}$ are invertible. In the Appendix we establish the diagonalization of an observed operator $L^{ij}_{X_a,X_a}$: $L^{ij}_{X_a,X_a} = L_{X_a|X_a^*} L^{ij}_{X_a^*} L^{-1}_{X_a|X_a^*}$ for all i, j, where the operator $L^{ij}_{X_a^*} \equiv \left(L_{X_a^*|W_j} L^{-1}_{X^*|W_j} L_{X^*|W_i} L^{-1}_{X_a^*|W_i} \right) : \mathcal{L}^p(\mathcal{X}^*) \to \mathcal{L}^p(\mathcal{X}^*)$ is a diagonal operator defined as: $\left(L^{ij}_{X_a^*} h \right)(x^*) = k^{ij}_{X_a^*}(x^*)\, h(x^*)$ with

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

In order to show the identification of $f_{X_a|X_a^*}$ and $k^{ij}_{X_a^*}(x^*)$, we assume

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

Notice that the subsets $W_1, W_2, ..., W_J \subset \mathcal{W}$ do not need to be collectively exhaustive. We may only consider those subsets in $\mathcal{W}$ in which these assumptions are satisfied. Since the indices i, j are exchangeable, the condition $\sup_{x^* \in \mathcal{X}^*} k^{ij}_{X^*_a}(x^*) < \infty$ may be replaced by $\inf_{x^* \in \mathcal{X}^*} k^{ij}_{X^*_a}(x^*) > 0$. Assumption 2.5 implies 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. Although there may exist duplicate eigenvalues in each decomposition corresponding to a pair of i and j, this assumption guarantees that each eigenfunction $f_{X_a|X^*_a}(\cdot|x^*)$ is uniquely determined by combining all the information from a series of decompositions of $L^{ij}_{X_a,X_a}$ for $i, j \in \{1, 2, ..., J\}$.

We now provide an example of the marginal distribution of X^* to illustrate that Assumption 2.5 is easily satisfied. Suppose that the distribution of X^* in the primary sample is the standard normal, i.e., $f_{X^*|W_j}(x^*) = \phi(x^*)$ for $j = 1, 2, 3$, where ϕ is the probability density function of the standard normal, and the distribution of X^*_a in the auxiliary sample is for $\sigma, \sigma' \in (0, 1)$ and $\mu \neq 0$

$$f_{X^*_a|W_j}(x^*) = \begin{cases} \phi(x^*) & \text{for } j = 1 \\ \sigma^{-1}\phi(\sigma^{-1}x^*) & \text{for } j = 2 \\ \frac{1}{\sigma'}\phi\left(\frac{x^*-\mu}{\sigma'}\right) & \text{for } j = 3 \end{cases}. \tag{2.1}$$

It is obvious that Assumption 2.5 is satisfied with

$$k^{ij}_{X^*_a}(x^*) = \begin{cases} \frac{\sigma^{-1}\phi\left(\sigma^{-1}x^*\right)}{\phi(x^*)} & \text{for } i = 1,\ j = 2 \\ \frac{\frac{1}{\sigma'}\phi\left(\frac{x^*-\mu}{\sigma'}\right)}{\phi(x^*)} & \text{for } i = 1,\ j = 3 \end{cases}. \tag{2.2}$$

Remark 2.1. (1) Assumption 2.5 does not hold if $f_{X^*|W=w_j}(x^*) = f_{X^*_a|W=w_j}(x^*)$ for all w_j and all $x^* \in \mathcal{X}^*$. This assumption requires that the two samples be from different populations. Given Assumption 2.3 and the invertibility of the operator $L_{Y|X^*,W_j}$, one could check Assumption 2.5 from the observed densities $f_{Y|W=w_j}$ and $f_{Y_a|W_a=w_j}$. In particular, if $f_{Y|W=w_j}(y) = f_{Y_a|W_a=w_j}(y)$ for all w_j and all $y \in \mathcal{Y}$, then Assumption 2.5 is not satisfied. **(2)** Assumption 2.5 does not hold if $f_{X^*|W=w_j}(x^*) = f_{X^*|W=w_i}(x^*)$ and $f_{X^*_a|W_a=w_j}(x^*) =$

$f_{X_a^*|W_a=w_i}(x^*)$ for all $w_j \neq w_i$ and all $x^* \in \mathcal{X}^*$. This means that the marginal distribution of X^* or X_a^* should be different in the subsamples corresponding to different w_j in at least one of the two samples. For example, if X^* or X_a^* are earnings and w_j corresponds to gender, then Assumption 2.5 requires that the earning distribution of males be different from that of females in one of the samples (either the primary or the auxiliary). Given the invertibility of the operators $L_{X|X^*}$ and $L_{X_a|X_a^*}$, one could check Assumption 2.5 from the observed densities $f_{X|W=w_j}$ and $f_{X_a|W_a=w_j}$. In particular, if $f_{X|W=w_j}(x) = f_{X|W=w_i}(x)$ for all $w_j \neq w_i$, and all $x \in \mathcal{X}$, then Assumption 2.5 requires the existence of an auxiliary sample such that $f_{X_a|W_a=w_j}(X_a) \neq f_{X_a|W_a=w_i}(X_a)$ with positive probability for some $w_j \neq w_i$.

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 $f_{X_a|X_a^*}(\cdot|x^*)$. Notice that the eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$ is identified up to the value of x^*. In other words, we have identified a probability density of X_a conditional on $X_a^* = x^*$ with the value of x^* unknown. An intuitive normalization assumption is that the value of x^* is the mean of this identified probability density, i.e., $x^* = \int x f_{X_a|X_a^*}(x|x^*)\,dx$; this assumption implies that the measurement error in the auxiliary sample has zero mean conditional on the latent true values. An alternative normalization assumption is that the value of x^* is the mode of this identified probability density, i.e., $x^* = \arg\max_x f_{X_a|X_a^*}(x|x^*)$; this assumption implies that the error distribution conditional on the latent true values has zero mode. The intuition behind this assumption is that people are more willing to report some values close to the latent true values than they are to report those far from the truth. Another normalization assumption may be that the value of x^* is the median of the identified probability density, i.e., $x^* = \inf\left\{z : \int_{-\infty}^{z} f_{X_a|X_a^*}(x|x^*)\,dx \geq \frac{1}{2}\right\}$; this assumption implies that the error distribution conditional on the latent true values has zero median, and that people have the same probability of over-reporting as that of underreporting. Obviously, the zero median condition can be generalized to an assumption that the error distribution conditional on the latent true values has a zero quantile.

Assumption 2.6. *One of the followings holds for all* $x^* \in \mathcal{X}^*$*: (i) (mean)* $\int x f_{X_a|X_a^*}(x|x^*)\,dx = x^*$*; 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* $\inf\left\{z : \int_{-\infty}^{z} f_{X_a|X_a^*}(x|x^*)\,dx \geq \gamma\right\} = x^*$.

Assumption 2.6 requires that the support of X_a not be smaller than that of X_a^*, and that, although the measurement error in the auxiliary sample $(X_a - X_a^*)$ could be nonclassical, it needs to satisfy some location regularity such as zero conditional mean, or zero conditional mode or zero conditional median.

We obtain the following identification result.

Theorem 2.1. Suppose Assumptions 2.1–2.6 hold. 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 $f_{X_a^*|W_j}$.

Remark 2.2. (1) When there exist extra common covariates in the two samples, we may consider more generally defined W and W_a, or relax assumptions on the error distributions in the auxiliary sample. On the one hand, this identification theorem still holds when we replace W and W_a with a scalar measurable function of W and W_a, respectively. On the other hand, we may relax Assumptions 2.1 and 2.2(ii) to allow the error distributions to be conditional on the true values and the extra common covariates. **(2)** The identification theorem does not require that the two samples be independent of each other.

3 Sieve Quasi Likelihood Estimation

Our identification result is very general and does not require the two samples to be independent. However, 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.

As shown in Section 2, 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):

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

Define

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

The latent parametric model is *correctly specified* if $g(y|x^*, w; \theta_0) = f_{Y|X^*,W}(y|x^*, w)$ for almost all y, x^*, w (and θ_0 is called true parameter value); otherwise it is *misspecified* (and θ_0 is called pseudo-true parameter value); see, e.g., White (1982).

Let $\alpha_0 \equiv (\theta_0^T, f_{01}, f_{01a}, f_{02}, f_{02a})^T \equiv (\theta_0^T, f_{X|X^*}, f_{X_a|X_a^*}, f_{X^*|W}, f_{X_a^*|W_a})^T$ 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 next provide a sieve (quasi-) MLE estimator $\widehat{\alpha}$ for α_0, and establish the root-n consistency and asymptotic normality of $\widehat{\theta}$ for θ_0, regardless of whether the latent model $g(y|x^*, w; \theta)$ is correctly specified or not.

3.1 Sieve Likelihood Under Possible Misspecification

Before we present a sieve (quasi-) MLE estimator $\widehat{\alpha}$ for α_0, we need to impose some mild smoothness restrictions on the unknown densities. The sieve method allows for unknown functions belonging to many different function spaces such as Sobolev space, Besov space, and others; see, e.g., Shen and Wong (1994) and Van de Geer (2000). But for the sake of concreteness and simplicity, we consider the widely used Hölder space of functions. Let $\xi = (\xi_1, \xi_2)^T \in \mathbb{R}^2$, $a = (a_1, a_2)^T$, and $\nabla^a h(\xi) \equiv \frac{\partial^{a_1+a_2} h(\xi_1,\xi_2)}{\partial \xi_1^{a_1} \partial \xi_2^{a_2}}$ denote the $(a_1+a_2)^{\text{th}}$ 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 $\gamma - \underline{\gamma} \in (0, 1]$. The Hölder space $\Lambda^\gamma(\mathcal{V})$ becomes a Banach

space under the Hölder norm:

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

We define a Hölder ball as $\Lambda_c^\gamma(\mathcal{V}) \equiv \{h \in \Lambda^\gamma(\mathcal{V}) : \|h\|_{\Lambda^\gamma} \leq c < \infty\}$. Denote

$$\mathcal{F}_1 = \left\{ f_1(\cdot|\cdot) \in \Lambda_c^{\gamma_1}\left(\mathcal{X}\times\mathcal{X}^*\right) : f_1(\cdot|x^*) > 0,\ \int_{\mathcal{X}} f_1(x|x^*)dx = 1 \text{ for all } x^* \in \mathcal{X}^* \right\},$$

$$\mathcal{F}_{1a} = \left\{ \begin{array}{c} f_{1a}(\cdot|\cdot) \in \Lambda_c^{\gamma_{1a}}\left(\mathcal{X}_a\times\mathcal{X}^*\right) : \text{ assumptions 2.4(i), 2.6 hold,} \\ f_{1a}(\cdot|x^*) > 0,\ \int_{\mathcal{X}_a} f_{1a}(x|x^*)dx = 1 \text{ for all } x^* \in \mathcal{X}^* \end{array} \right\},$$

$$\mathcal{F}_2 = \left\{ \begin{array}{c} f_2\left(\cdot|w\right) \in \Lambda_c^{\gamma_2}\left(\mathcal{X}^*\right) : \text{ Assumption 2.5 holds,} \\ f_2\left(\cdot|w\right) > 0,\ \int_{\mathcal{X}^*} f_2\left(x^*|w\right)dx^* = 1 \text{ for all } w \in \mathcal{W} \end{array} \right\},$$

We impose the following smoothness restrictions on the densities:

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)* $f_{X^*|W}\left(\cdot|w\right), f_{X_a^*|W_a}\left(\cdot|w\right) \in \mathcal{F}_2$ *with* $\gamma_2 > 1/2$ *for all* $w \in \mathcal{W}$*.*

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

$$\left\{Z_t^T \equiv (S_tX_t, S_tW_t, S_tY_t, S_t, (1-S_t)X_t, (1-S_t)W_t, (1-S_t)Y_t)\right\}_{t=1}^{n+n_a}$$

such that $\{X_t, W_t, Y_t, S_t = 1\}_{t=1}^n$ is the primary sample and $\{X_t, W_t, Y_t, S_t = 0\}_{t=n+1}^{n+n_a}$ is the auxiliary sample. Denote $p \equiv \Pr(S_t = 1) \in (0,1)$. 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 $\alpha \equiv (\theta^T, f_1, f_{1a}, f_2, f_{2a})^T \in \mathcal{A}$ is given by:

$$\begin{aligned} & \textstyle\sum_{t=1}^{n+n_a} \{S_t \log\left[p \times f(X_t, W_t, Y_t|S_t = 1; \alpha)\right] + (1-S_t)\log\left[(1-p)\times f(X_t, W_t, Y_t|S_t = 0; \alpha)\right]\} \\ = \ & n\log(p) + n_a \log\{(1-p)\} + \textstyle\sum_{t=1}^{n+n_a} \ell(Z_t;\alpha), \end{aligned}$$

in which

$$\begin{aligned} \ell(Z_t;\alpha) &\equiv S_t\ell_p(Z_t;\theta,f_1,f_2) + (1-S_t)\ell_a(Z_t;f_{1a},f_{2a}), \\ \ell_p(Z_t;\theta,f_1,f_2) &= \log\int f_1(X_t|x^*)g(Y_t|x^*,W_t;\theta)f_2(x^*|W_t)dx^* + \log f_W(W_t), \\ \ell_a(Z_t;f_{1a},f_{2a}) &= \log\int f_{1a}(X_t|x_a^*)g(Y_t|x_a^*,W_t;\theta)f_{2a}(x_a^*|W_t)dx_a^* + \log\{f_{W_a}(W_t)\}. \end{aligned}$$

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

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

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

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

We could apply infinite-dimensional approximating spaces as sieves $\mathcal{F}_j^n$ for $\mathcal{F}_j, j = 1, 1a, 2$. However, in applications 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. Then we denote the sieve space for $\mathcal{F}_1$, $\mathcal{F}_{1a}$, and $\mathcal{F}_2$ as follows:

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

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

3.1.1 Consistency Under a Strong Norm

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 $\omega_j(\xi) = \left(1 + \|\xi\|_E^2\right)^{-\varsigma_j/2}$, $\varsigma_j > 0$ for $j = 1, 1a, 2$. We assume each of $\mathcal{X}$, $\mathcal{X}_a$, $\mathcal{X}^*$ is $\mathbb{R}$, and

Assumption 3.2. *(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. In addition, $\lim_{n\to\infty} \frac{n}{n+n_a} = p \in (0,1)$; (ii) $g(y|x^*,w;\theta)$ is continuous in $\theta \in \Theta$, and Θ is a compact subset of $\mathbb{R}^{d_\theta}$; and (iii) $\theta_0 \in \Theta$ is the unique maximizer of $\int[\log g(y|x^*,w;\theta)] f_{Y|X^*,W}(y|x^*,w)dy$ over $\theta \in \Theta$.*

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

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)* $k_n \equiv \max\{k_{1,n}, k_{1a,n}, k_{2,n}\} \to \infty$ *and* $k_n/n \to 0$.

Assumption 3.2(i) is a typical condition used in cross-sectional analyses with two samples. Assumption 3.2(ii–iii) are typical conditions for parametric (quasi-) MLE of θ_0 if X^* could be observed without error. The following consistency lemma is a direct application of theorem 3.1 (or remark 3.3) of Chen (2006); hence, we omit its proof.

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

3.1.2 Convergence Rate Under a Weaker Metric

Given Lemma 3.1, we can now restrict our attention to a shrinking $||\cdot||_s-$neighborhood around α_0. Let $\mathcal{A}_{0s} \equiv \{\alpha \in \mathcal{A} : ||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \leq c_0 < c\}$ and $\mathcal{A}_{0sn} \equiv \{\alpha \in \mathcal{A}_n : ||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \leq c_0 < c\}$. Then, for the purpose of establishing a convergence rate under a pseudo metric that is weaker than $||\cdot||_s$, we can treat $\mathcal{A}_{0s}$ as the new parameter space and $\mathcal{A}_{0sn}$ as its sieve space, and assume that both $\mathcal{A}_{0s}$ and $\mathcal{A}_{0sn}$ are convex parameter spaces. For any α_1, $\alpha_2 \in \mathcal{A}_{0s}$, we consider a continuous path $\{\alpha(\tau) : \tau \in [0,1]\}$ in $\mathcal{A}_{0s}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. For simplicity we assume that for any α, $\alpha + v \in \mathcal{A}_{0s}$, $\{\alpha + \tau v : \tau \in [0,1]\}$ is a continuous path in $\mathcal{A}_{0s}$, and that $\ell(Z_t;\alpha+\tau v)$ is twice continuously differentiable at $\tau = 0$ for almost all Z_t and any direction $v \in \mathcal{A}_{0s}$. We define the pathwise first derivative as

$$\frac{d\ell(Z_t;\alpha)}{d\alpha}[v] \equiv \frac{d\ell(Z_t;\alpha+\tau v)}{d\tau}|_{\tau=0} \text{ a.s. } Z_t,$$

and the pathwise second derivative as

$$\frac{d^2\ell(Z_t;\alpha)}{d\alpha d\alpha^T}[v,v] \equiv \frac{d^2\ell(Z_t;\alpha+\tau v)}{d\tau^2}|_{\tau=0} \quad \text{a.s. } Z_t.$$

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

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

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 3.5. *(i)* $\varsigma_j > \gamma_j$ *for* $j = 1, 1a, 2$*; (ii)* $k_n^{-\gamma} = o([n+n_a]^{-1/4})$ *with* $\gamma \equiv \min\{\gamma_1/2, \gamma_{1a}/2, \gamma_2\} > 1/2$.

Assumption 3.6. *(i)* $\mathcal{A}_{0s}$ *is convex at* α_0 *and* $\theta_0 \in int\,(\Theta)$*; (ii)* $\ell(Z_t;\alpha)$ *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 3.7. $\sup_{\widetilde{\alpha}\in\mathcal{A}_{0s}} \sup_{\alpha\in\mathcal{A}_{0sn}} \left|\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}\left[\frac{\alpha-\alpha_0}{\|\alpha-\alpha_0\|_s}\right]\right| \leq U(Z_t)$ *for a random variable* $U(Z_t)$ *with* $E\{[U(Z_t)]^2\} < \infty$.

Assumption 3.8. *(i)* $\sup_{v\in\mathcal{A}_{0s}:||v||_s=1} -E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[v,v]\right) \leq C < \infty$*; (ii) uniformly over* $\widetilde{\alpha} \in \mathcal{A}_{0s}$ *and* $\alpha \in \mathcal{A}_{0sn}$*, we have*

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

Assumption 3.5 guarantees that the sieve approximation error under the strong norm $||\cdot||_s$ goes to zero faster than $[n+n_a]^{-1/4}$. Assumption 3.6 makes sure that the twice pathwise derivatives are well defined with respect to $\alpha \in \mathcal{A}_{0s}$; hence, the pseudo metric $\|\alpha - \alpha_0\|_2$ is well defined on $\mathcal{A}_{0s}$. Assumption 3.7 imposes an envelope condition. Assumption 3.8(i) implies that $\|\alpha - \alpha_0\|_2 \leq \sqrt{C}\,\|\alpha - \alpha_0\|_s$ for all $\alpha \in \mathcal{A}_{0s}$. Assumption 3.8(ii) implies that there are positive finite constants C_1 and C_2, such that for all $\alpha \in \mathcal{A}_{0sn}$, $C_1\,\|\alpha - \alpha_0\|_2^2 \leq E[\ell(Z_t;\alpha_0) - \ell(Z_t;\alpha)] \leq C_2\,\|\alpha - \alpha_0\|_2^2$. The following convergence rate theorem is a direct application of Theorem 3.2 of Shen and Wong (2004) to the local parameter space $\mathcal{A}_{0s}$ and the local sieve space $\mathcal{A}_{0sn}$; hence, we omit its proof.

Theorem 3.1. Under assumptions 3.1–3.8, if $k_n = O\left([n+n_a]^{\frac{1}{2\gamma+1}}\right)$, then

$$\|\widehat{\alpha}_n - \alpha_0\|_2 = O_P\left(\max\left[k_n^{-\gamma}, \{k_n/(n+n_a)\}^{1/2}\right]\right) = O_P\left([n+n_a]^{\frac{-\gamma}{2\gamma+1}}\right).$$

3.2 Asymptotic Normality Under Possible Misspecification

We can derive the asymptotic distribution of the sieve quasi MLE $\widehat{\theta}_n$ regardless of whether the latent parametric model $g(y|x^*, w; \theta_0)$ is correctly specified or not. First, we define an inner product corresponding to the pseudo metric $\|\cdot\|_2$:

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

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the metric $\|\cdot\|_2$. Then $\left(\overline{\mathbf{V}}, \|\cdot\|_2\right)$ 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. Then the pathwise first derivative can be written as

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

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

$$\begin{aligned}\frac{d\ell(Z;\alpha_0)}{dh}[h-h_0] &= \frac{d\ell(Z;\theta_0, h_0(1-\tau)+\tau h)}{d\tau}|_{\tau=0} \\ &= \frac{d\ell(Z_t;\alpha_0)}{df_1}[f_1 - f_{01}] + \frac{d\ell(Z_t;\alpha_0)}{df_{1a}}[f_{1a} - f_{01a}] \\ &\quad + \frac{d\ell(Z_t;\alpha_0)}{df_2}[f_2 - f_{02}] + \frac{d\ell(Z_t;\alpha_0)}{df_{2a}}[f_{2a} - f_{02a}].\end{aligned}$$

Note that

$$\begin{aligned}&E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[\alpha-\alpha_0, \alpha-\alpha_0]\right) \\ =\ &(\theta-\theta_0)^T E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\theta d\theta^T} - 2\frac{d^2\ell(Z;\alpha_0)}{d\theta dh^T}[\mu] + \frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu,\mu]\right)(\theta-\theta_0),\end{aligned}$$

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

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

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

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

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

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

$$\frac{d\ell(Z;\alpha_0)}{dh}[\mu^*]=\left(\frac{d\ell(Z;\alpha_0)}{dh}\left[\mu^{*1}\right],...,\frac{d\ell(Z;\alpha_0)}{dh}\left[\mu^{*d_\theta}\right]\right),$$

$$\frac{d^2\ell(Z;\alpha_0)}{\partial\theta dh^T}[\mu^*]=\left(\frac{d^2\ell(Z;\alpha_0)}{\partial\theta dh}[\mu^{*1}],...,\frac{d^2\ell(Z;\alpha_0)}{\partial\theta dh}[\mu^{*d_\theta}]\right),$$

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

Also denote

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

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

$$\begin{aligned}&\sup_{\alpha-\alpha_0\neq 0}\frac{|\lambda^T(\theta-\theta_0)|^2}{||\alpha-\alpha_0||_2^2}\\ =\;&\sup_{\theta\neq\theta_0,\mu\neq 0}\frac{(\theta-\theta_0)^T\lambda\lambda^T(\theta-\theta_0)}{(\theta-\theta_0)^TE\left\{-\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\theta d\theta^T}-2\frac{d^2\ell(Z;\alpha_0)}{d\theta dh^T}[\mu]+\frac{d^2\ell(Z;\alpha_0)}{dhdh^T}[\mu,\mu]\right)\right\}(\theta-\theta_0)}\\ =\;&\lambda^T(V_*)^{-1}\lambda,\end{aligned}$$

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

Suppose that V_* is nonsingular. For any fixed $\lambda\neq 0$, denote $\upsilon^*\equiv(v_\theta^*,v_h^*)$ with $v_\theta^*\equiv(V_*)^{-1}\lambda$ and $v_h^*\equiv-\mu^*\times v_\theta^*$. Then the Riesz representation theorem implies: $\lambda^T(\theta-\theta_0)=\langle\upsilon^*,\alpha-\alpha_0\rangle_2$ for all $\alpha\in\mathcal{A}$. In the appendix, we show that

$$\lambda^T\left(\widehat{\theta}_n-\theta_0\right)=\langle\upsilon^*,\widehat{\alpha}_n-\alpha_0\rangle_2=\frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]+o_p\left(\frac{1}{\sqrt{n+n_a}}\right).$$

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

Assumption 3.10. *There is a $\upsilon_n^* \in \mathcal{A}_n - \{\alpha_0\}$, such that $||\upsilon_n^* - \upsilon^*||_2 = o(1)$ and $\|\upsilon_n^* - \upsilon^*\|_2 \times \|\widehat{\alpha}_n - \alpha_0\|_2 = o_P(\frac{1}{\sqrt{n+n_a}})$.*

Assumption 3.11. *There is a random variable $U(Z_t)$ with $E\{[U(Z_t)]^2\} < \infty$ and a non-negative measurable function η with $\lim_{\delta \to 0} \eta(\delta) = 0$, such that, for all $\alpha \in \mathcal{N}_{0n}$,*

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

Assumption 3.12. *Uniformly over $\overline{\alpha} \in \mathcal{N}_0$ and $\alpha \in \mathcal{N}_{0n}$,*

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

Assumption 3.13. *$E\left\{ \left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^* - \upsilon^*] \right)^2 \right\}$ goes to zero as $\|\upsilon_n^* - \upsilon^*\|_2$ goes to zero.*

It is easily seen that Assumption 3.13 is automatically satisfied when the latent parametric model is correctly specified. The other assumptions are necessary for the proofs. Denote

$$\mathcal{S}_{\theta_0} \equiv \frac{d\ell(Z_t; \alpha_0)}{d\theta^T} - \frac{d\ell(Z_t; \alpha_0)}{dh}[\mu^*] \quad \text{and} \quad I_* \equiv E\left[\mathcal{S}_{\theta_0}^T \mathcal{S}_{\theta_0}\right].$$

The following asymptotic normality result applies to possibly misspecified models.

Theorem 3.2. Under Assumptions 3.1–3.13, we have $\sqrt{n+n_a}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathrm{N}\left(0, V_*^{-1} I_* V_*^{-1}\right)$.

3.3 Semiparametric Efficiency under Correct Specification

In this subsection we assume that $g(y|x^*, w; \theta_0)$ correctly specifies the true unknown conditional density $f_{Y|X^*,W}(y|x^*, w)$. We can then establish the semiparametric efficiency of the

two-sample sieve MLE $\widehat{\theta}_n$ for the parameter of interest θ_0. Recall the definitions of Fisher inner product and the Fisher norm:

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

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 $\|\cdot\|$. For each parametric component θ^k of θ, $k = 1, 2, ..., d_\theta$, an alternative way to obtain $\mu^* = \left(\mu^{*1}, \mu^{*2}, ..., \mu^{*d_\theta}\right)$ is to compute $\mu^{*k} \equiv \left(\mu_1^{*k}, \mu_{1a}^{*k}, \mu_2^{*k}, \mu_{2a}^{*k}\right)^T \in \overline{\mathcal{U}}$ as the solution to

$$\begin{aligned} & \inf_{\mu^k \in \overline{\mathcal{U}}} E\left\{ \left(\frac{d\ell(Z_t;\alpha_0)}{d\theta^k} - \frac{d\ell(Z_t;\alpha_0)}{dh}\left[\mu^k\right] \right)^2 \right\} \\ = & \inf_{(\mu_1, \mu_{1a}, \mu_2, \mu_{2a})^T \in \overline{\mathcal{U}}} E\left\{ \begin{pmatrix} \frac{d\ell(Z_t;\alpha_0)}{d\theta^k} - \frac{d\ell(Z_t;\alpha_0)}{df_1}[\mu_1] - \frac{d\ell(Z_t;\alpha_0)}{df_{1a}}[\mu_{1a}] \\ -\frac{d\ell(Z_t;\alpha_0)}{df_2}[\mu_2] - \frac{d\ell(Z_t;\alpha_0)}{df_{2a}}[\mu_{2a}] \end{pmatrix}^2 \right\}. \end{aligned}$$

Then $\mathcal{S}_{\theta_0} \equiv \frac{d\ell(Z_t;\alpha_0)}{d\theta^T} - \frac{d\ell(Z_t;\alpha_0)}{dh}[\mu^*]$ becomes the semiparametric efficient score for θ_0, and $I_* \equiv E\left[\mathcal{S}_{\theta_0}^T \mathcal{S}_{\theta_0}\right] = V_*$ becomes the semiparametric information bound for θ_0.

Given the expression of the density function, the pathwise first derivative at α_0 is

$$\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\alpha - \alpha_0] = S_t \frac{d\ell_p(Z_t;\theta_0, f_{01}, f_{02})}{d\alpha}[\alpha - \alpha_0] + (1 - S_t)\frac{d\ell_a(Z_t; f_{01a}, f_{02a})}{d\alpha}[\alpha - \alpha_0].$$

Thus $I_* \equiv E\left[\mathcal{S}_{\theta_0}^T \mathcal{S}_{\theta_0}\right] = pI_{*p} + (1-p)I_{*a}$, with

$$I_{*p} = E\begin{bmatrix} \left(\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02})}{d\theta^T} - \sum_{j=1}^2 \frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02})}{df_j}\left[\mu_j^*\right] \right)^T \\ \left(\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02})}{d\theta^T} - \sum_{j=1}^2 \frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02})}{df_j}\left[\mu_j^*\right] \right) \end{bmatrix},$$

$$I_{*a} = E\begin{bmatrix} \left(\sum_{j=1}^2 \frac{d\ell_a(Z_t;f_{01a},f_{02a})}{df_{ja}}\left[\mu_{ja}^*\right] \right)^T \\ \left(\sum_{j=1}^2 \frac{d\ell_a(Z_t;f_{01a},f_{02a})}{df_{ja}}\left[\mu_{ja}^*\right] \right) \end{bmatrix}.$$

To order $o_p\{(n+n_a)^{-1/2}\}$, the influence function representation is

$$\lambda^T\left(\widehat{\theta}_n - \theta_0\right) = \frac{1}{n+n_a}\left\{ \sum_{i=1}^n \frac{d\ell_p(Z_{pi};\theta_0,f_{01},f_{02})}{d\alpha}[\upsilon^*] + \sum_{j=1}^{n_a} \frac{d\ell_a(Z_{aj};f_{01a},f_{02a})}{d\alpha}[\upsilon^*] \right\},$$

and the asymptotic distribution of $\sqrt{n+n_a}\left(\widehat{\theta}_n - \theta_0\right)$ is $\mathrm{N}\left(0, I_*^{-1}\right)$. Combining our Theorem 3.2 and Theorem 4 of Shen (1997), we immediately obtain

Theorem 3.3. Suppose that $g(y|x^*, w; \theta_0) = f_{Y|X^*,W}(y|x^*, w)$ for almost all y, x^*, w, that I_* is positive definite, and that Assumptions 3.1– 3.12 hold. Then the two-sample sieve MLE $\widehat{\theta}_n$ is semiparametrically efficient, and $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathrm{N}\left(0, [I_{*p} + \frac{1-p}{p} I_{*a}]^{-1}\right) = \mathrm{N}\left(0, pI_*^{-1}\right)$.

Following Ai and Chen (2003), the asymptotic efficient variance, I_*^{-1}, of the sieve MLE $\widehat{\theta}_n$ (under correct specification) can be consistently estimated by $\widehat{I}_*^{-1}$, with

$$\widehat{I}_* = \frac{1}{n+n_a} \sum_{t=1}^{n+n_a} \left(\frac{d\ell(Z_t; \widehat{\alpha})}{d\theta^T} - \frac{d\ell(Z_t; \widehat{\alpha})}{dh} [\widehat{\mu}^*] \right)^T \left(\frac{d\ell(Z_t; \widehat{\alpha})}{d\theta^T} - \frac{d\ell(Z_t; \widehat{\alpha})}{dh} [\widehat{\mu}^*] \right),$$

in which $\widehat{\mu}^* = \left(\widehat{\mu}^{*1}, \widehat{\mu}^{*2}, ..., \widehat{\mu}^{*d_\theta}\right)$ and $\widehat{\mu}^{*k} \equiv \left(\widehat{\mu}_1^{*k}, \widehat{\mu}_{1a}^{*k}, \widehat{\mu}_2^{*k}, \widehat{\mu}_{2a}^{*k}\right)^T$ solves the following sieve minimization problem: for $k = 1, 2, ..., d_\theta$,

$$\min_{\mu^k \in \mathcal{F}_n} \sum_{t=1}^{n+n_a} \left(\begin{array}{c} \frac{d\ell(Z_t;\widehat{\alpha})}{d\theta^k} - \frac{d\ell(Z_t;\widehat{\alpha})}{df_1}\left[\mu_1^k\right] - \frac{d\ell(Z_t;\widehat{\alpha})}{df_{1a}}\left[\mu_{1a}^k\right] \\ -\frac{d\ell(Z_t;\widehat{\alpha})}{df_2}\left[\mu_2^k\right] - \frac{d\ell(Z_t;\widehat{\alpha})}{df_{2a}}\left[\mu_{2a}^k\right] \end{array} \right)^2,$$

in which $\mathcal{F}_n \equiv \mathcal{F}_1^n \times \mathcal{F}_{1a}^n \times \mathcal{F}_2^n \times \mathcal{F}_2^n$, and where

$$\frac{d\ell(Z_t; \widehat{\alpha})}{dh} \left[\widehat{\mu}^{*k}\right] \equiv \frac{d\ell(Z_t; \widehat{\alpha})}{df_1} \left[\widehat{\mu}_1^{*k}\right] + \frac{d\ell(Z_t; \widehat{\alpha})}{df_{1a}} \left[\widehat{\mu}_{1a}^{*k}\right] + \frac{d\ell(Z_t; \widehat{\alpha})}{df_2} \left[\widehat{\mu}_2^{*k}\right] + \frac{d\ell(Z_t; \widehat{\alpha})}{df_{2a}} \left[\widehat{\mu}_{2a}^{*k}\right],$$

and

$$\frac{d\ell(Z_t; \widehat{\alpha})}{dh} [\widehat{\mu}^*] = \left(\frac{d\ell(Z_t; \widehat{\alpha})}{dh} \left[\widehat{\mu}^{*1}\right], ..., \frac{d\ell(Z_t; \widehat{\alpha})}{dh} \left[\widehat{\mu}^{*d_\theta}\right] \right).$$

4 Simulation

We present a small simulation study to illustrate the finite sample performance of the two-sample sieve MLE. The true latent 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, $\theta = (\theta_1, \theta_2, \theta_3)^T$, $\theta_0 = (1, 1, 1)^T$ and

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

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 $n = 1500$ and $n_a = 1000$. 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 $X = 0.1X^* + e^{-0.1X^*}\varepsilon$ with $\varepsilon \sim \mathrm{N}(0, 0.36)$.

Table 1: Simulation results ($n = 1500, n_a = 1000, reps = 400$)

true value of θ:	$\theta_1 = 1$	$\theta_2 = 1$	$\theta_3 = 1$
ignoring measurement error:			
– mean estimate	0.175	0.307	0.595
– standard error	0.084	0.123	0.188
– root mse	0.829	0.703	0.446
infeasible MLE:			
– mean estimate	0.999	1.001	1.000
– standard error	0.028	0.034	0.035
– root mse	0.028	0.034	0.035
2-sample sieve MLE:			
– mean estimate	1.038	1.065	1.049
– standard error	0.087	0.135	0.244
– root mse	0.095	0.150	0.248

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

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

We let the mismeasured value X_a be $X_a = X_a^* + 0.5e^{-X_a^*}\nu$ 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 results shown in Table 1 include three estimators for $\theta_0 = (1,1,1)^T$. The first estimator is the standard probit MLE using the primary sample $\{X_i, W_i, Y_i\}_{i=1}^n$ alone as if it were accurate; this estimator is inconsistent and its bias should dominate the square root of mean square error (root mse). The second estimator is the standard probit MLE using accurate data $\{Y_i, X_i^*, W_i\}_{i=1}^n$. This estimator is consistent and the most efficient; however, we call it "infeasible MLE" since X_i^* is not observed in practice. The third estimator is the two-sample sieve MLE developed in this paper, in which the true functional form $f_{Y|X^*,W}(y|x^*, w; \theta)$ is known up to the unknown θ, and the conditional densities $f_{X|X^*}$, $f_{X_a|X_a^*}$, $f_{X^*|W}$ and $f_{X_a^*|W}$ are unknown. We use the simple sieve expression $p_1^{k_{1,n}}(x_1, x_2)^T\beta_1 = \sum_{j=0}^{J_n}\sum_{k=0}^{K_n}\gamma_{jk}p_j\left(x_1 - x_2\right)q_k\left(x_2\right)$ 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 $p_2^{k_{2,n}}(x^*)^T\beta_2(w) = \sum_{k=1}^{k_{2,n}}\gamma_k(w)q_k\left(x^*\right)$ 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.

The simulation was repeated 400 times. The simulation results in Table 1 show that the 2-sample sieve MLE has a much smaller bias than the estimator ignoring measurement error. Surprisingly, the sieve Q-MLE has only a slightly larger standard error that the naive estimator, hence it has much smaller total root mse. In summary, our 2-sample sieve MLE performs well in this Monte Carlo simulation.

5 Illustrative Example

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-hour 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-hour recalls. The sample sizes for EATS and OPEN were 965 and 481, respectively. In EATS, there were 315 (131) men under (over) the age of 50 and 364 (155) women under (over) the age of 50. In OPEN, there were 102 (157) men under (over) the age of 50 and 98 (124) women under (over) the age of 50.

With EATS as the primary study and OPEN as the auxiliary study, the assumption of nondifferential measurement error as expressed in Assumptions 2.1 and 2.2 are standard in this context. While our 2-sample sieve Q-MLE does not make this assumption, it makes sense to believe that the measurement error distributions are the same in the two studies. Both studies took place in the United State, and thus the stability Assumption 2.3 also

Table 2: 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	0.562	0.100	0.054	-0.116
– Boot Mean	0.242	0.083	0.035	-0.044	0.617	0.091	0.041	-0.105
– Boot Median	0.242	0.083	0.033	-0.043	0.546	0.088	0.041	-0.092
– Boot s.e.	0.019	0.040	0.040	0.034	0.238	0.044	0.040	0.091
– Boot 95% CI	0.204	0.007	-0.039	-0.121	0.335	0.004	-0.031	-0.299
	0.284	0.161	0.122	0.017	1.362	0.186	0.125	0.025
	parametric MLE:				2-sample sieve MLE:			
– Estimate	0.461	0.131	-0.019	-0.073	0.749	0.151	0.064	0.188
– Boot Mean	0.485	0.135	-0.027	-0.074	0.727	0.124	0.072	0.064
– Boot Median	0.466	0.132	-0.021	-0.073	0.764	0.126	0.065	0.032
– Boot s.e.	0.194	0.061	0.064	0.045	0.318	0.067	0.060	0.171
– Boot 95% CI	0.292	0.041	-0.211	-0.181	0.091	-0.007	-0.020	-0.136
	1.179	0.288	0.078	0.002	1.304	0.243	0.225	0.549

seems reasonable. The main difference between EATS and OPEN is that the former was a national study, while the latter took place exclusively 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, assumptions 2.4 - 2.6 seem reasonable in this context. Indeed, for those aged under 50, Wilcoxon rank tests comparing the two transformed 24-hour 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-hour recalls for women had statistically significantly more variability than those for men, as measured by a Wilcoxon test performed 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 ethic status (Caucasian = 1) of the individual. The latent model of interest is

$$Y_{ij} = \theta_4 + \theta_1 X_{ij}^* + \theta_2 W_{ij1} + \theta_3 W_{ij2} + \epsilon_{ij}, \quad X_{ij} = X_{ij}^* + U_{ij}, \tag{4.1}$$

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

We consider four estimators for $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)^T$. The first estimator is a naive OLS estimator with measurement errors in X_{ij} ignored. The second estimator is a parametric maximum likelihood estimator under the additional Assumptions: $\epsilon_{ij} = \text{Normal}(0, \sigma^2_\epsilon)$, $U_{ij} = \text{Normal}(0, \sigma^2_u)$, $X^*_{i1} = a_0 + a_1 W_{ij1} + a_2 W_{ij2} + \nu_{i1}$, and $X^*_{i2} = b_0 + b_1 W_{ij1} + b_2 W_{ij2} + \nu_{i2}$, with $\nu_{ij} = \text{Normal}(0, \sigma^2_{\nu,j})$. Note that for this parametric MLE, the measurement error status is assumed to not depend on j. The third and fourth estimators are the semiparametric sieve MLEs: the third estimator assumes the latent model of interest is (4.1) with $\epsilon_{ij} = \text{Normal}(0, \sigma^2_\epsilon)$; the fourth estimator does not assume the functional form of f_ϵ but estimates it via sieve MLE. We assume Assumption 2.6(i) holds in the EATS sample. To compute the third and the fourth estimators, we use the Hermite polynomial sieve bases as those in the simulation study: $p_1^{k_{1,n}}(x_1, x_2)^T \beta_1$ to approximate $f_{X|X^*}(x_1|x_2)$ and $f_{X_a|X^*_a}(x_1|x_2)$ with $k_{1,n} = (J_n + 1)(K_n + 1)$, $J_n = 5$, $K_n = 3$, and $p_2^{k_{2,n}}(x^*)^T \beta_2(w)$ to approximate $f_{X^*|W=w}$ and $f_{X^*_a|W=w}$ with $k_{2,n} = 5$. In addition, we use Hermite polynomial sieve $p_3^{k_{3,n}}(\epsilon)^T \beta_1$ with $k_{2,n} = 3$ to approximate $f_\epsilon(\epsilon)$ 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 2. We see that the measurement errors cause significant attenuation in the estimation of θ_1. The corrected estimators have much greater variability than the naive estimator, with variability increasing as assumptions are relaxed.

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 be in 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 have provided very general conditions ensuring identifiability: essentially, we require that the distribution of X^* depend 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 an example show that our method has good performance despite the generality of the approach.

7 Appendix: Mathematical Proofs

Proof : (Theorem 2.1) Under Assumption 2.1,

$$f_{X,W,Y}(x,w,y) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*) f_{X^*,W,Y}(x^*,w,y)dx^* \quad \text{for all } x,w,y. \tag{A.1}$$

For each value w_j of W, assumptions 2.1-2.3 imply that

$$f_{X,Y|W}(x,y|w_j) = \int f_{X|X^*}(x|x^*) f_{Y|X^*,W}(y|x^*,w_j) f_{X^*|W_j}(x^*)dx^*, \tag{A.2}$$

$$f_{X_a,Y_a|W_a}(x,y|w_j) = \int f_{X_a|X_a^*}(x|x^*) f_{Y|X^*,W}(y|x^*,w_j) f_{X_a^*|W_j}(x^*)dx^* \tag{A.3}$$

By equation (A.2), for any function $h \in \mathcal{L}^p(\mathcal{Y})$,

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

This means we have the operator equivalence

$$L_{X,Y|W_j} = L_{X|X^*} L_{X^*|W_j} L_{Y|X^*,W_j} \tag{A.4}$$

in the primary sample. Similarly, equation (A.3) and the definition of the operators imply

$$L_{X_a,Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{Y|X^*,W_j} \tag{A.5}$$

in the auxiliary sample. While the left-hand sides of equations (A.4) and (A.5) are observed, the right-hand sides contain unknown operators corresponding to the error distributions ($L_{X|X^*}$ and $L_{X_a|X_a^*}$), the marginal distributions of the latent true values ($L_{X^*|W_j}$ and $L_{X_a^*|W_j}$), and the conditional distribution of the dependent variable ($L_{Y|X^*,W_j}$).

Assumption 2.4 imply that all the operators involved in equations (A.4) and (A.5) are invertible. Under assumption 2.4, for any given W_j we can eliminate $L_{Y|X^*,W_j}$ in equations (A.4) and (A.5) to obtain

$$L_{X_a,Y_a|W_j} L_{X,Y|W_j}^{-1} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{X^*|W_j}^{-1} L_{X|X^*}^{-1} : \mathcal{L}^p(\mathcal{X}) \to \mathcal{L}^p(\mathcal{X}_a). \tag{A.6}$$

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

$$\begin{aligned} L_{X_a,X_a}^{ij} &\equiv \left(L_{X_a,Y_a|W_j} L_{X,Y|W_j}^{-1}\right)\left(L_{X_a,Y_a|W_i} L_{X,Y|W_i}^{-1}\right)^{-1} \\ &= L_{X_a|X_a^*}\left(L_{X_a^*|W_j} L_{X^*|W_j}^{-1} L_{X^*|W_i} L_{X_a^*|W_i}^{-1}\right) L_{X_a|X_a^*}^{-1} \\ &= L_{X_a|X_a^*} L_{X_a^*}^{ij} L_{X_a|X_a^*}^{-1} : \mathcal{L}^p(\mathcal{X}_a) \to \mathcal{L}^p(\mathcal{X}_a) \end{aligned} \tag{A.7}$$

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 $L_{X_a^*}^{ij} \equiv \left(L_{X_a^*|W_j} L_{X^*|W_j}^{-1} L_{X^*|W_i} L_{X_a^*|W_i}^{-1}\right) : \mathcal{L}^p(\mathcal{X}^*) \to \mathcal{L}^p(\mathcal{X}^*)$ is a diagonal operator defined as $\left(L_{X_a^*}^{ij} h\right)(x^*) \equiv k_{X_a^*}^{ij}(x^*)\, h(x^*)$ with

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

Equation (A.7) implies a diagonalization of an observed operator L_{X_a,X_a}^{ij}. An eigenvalue of L_{X_a,X_a}^{ij} equals $k_{X_a^*}^{ij}(x^*)$ 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; see, e.g., Dunford and Schwartz (1971, theorem XV.4.3.5, p. 1939). Equation (A.7) 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 diagonalization of the operator $L^{ij}_{X_a,X_a}$, equation (A.7) does not guarantee distinctive eigenvalues. If there exist duplicate eigenvalues, there exist two linearly independent eigenfunctions corresponding to the same eigenvalue. A linear combination of the two eigenfunctions is also an eigenfunction corresponding to the same eigenvalue. Therefore, the eigenfunctions may not be identified in each decomposition corresponding to a pair of i and j. 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^*. To eliminate this ambiguity, we need to normalize each eigenfunction. Notice that $f_{X_a|X_a^*}(\cdot|x^*)$ is a conditional probability density for each x^*; hence, $\int f_{X_a|X_a^*}(x|x^*)\,dx = 1$ for all x^*. This property of conditional density provides a perfect normalization condition.

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 $f_{X_a|X_a^*}(\cdot|x^*)$. Notice that the eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$ is identified up to the value of x^*. In other words, we have identified a probability density of X_a conditional on $X_a^* = x^*$ with the value of x^* unknown. Note that assumption 2.6 identifies the exact value of x^* for each eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$. For example, an intuitive assumption is that the value of x^* is the mean of this identified probability density, i.e., $x^* = \int x f_{X_a|X_a^*}(x|x^*)\,dx$; this assumption 2.6(i) is equivalent to that the measurement

error in the auxiliary sample $(X_a - X_a^*)$ has zero mean conditional on the latent true values.

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 (A.3), 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

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

By equation (A.4) and the injectivity of the identified operator $L_{Y|X^*,W_j}$, we have

$$L_{X|X^*} L_{X^*|W_j} = L_{X,Y|W_j} L^{-1}_{Y|X^*,W_j}. \tag{A.8}$$

The left-hand side of equation (A.8) 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 (A.8) 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

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

Proof : (Theorem 3.2) For any $\alpha \in \mathcal{N}_{0n}$, define

$$r[Z_t;\alpha,\alpha_0] \equiv \ell(Z_t;\alpha) - \ell(Z_t;\alpha_0) - \frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\alpha - \alpha_0].$$

Denote the centered empirical process indexed by any measurable function h as

$$\mu_n\left(h(Z_t)\right) \equiv \frac{1}{n+n_a}\textstyle\sum_{t=1}^{n+n_a}\{h(Z_t) - E[h(Z_t)]\}.$$

Let $\varepsilon_n > 0$ be at the order of $o([n+n_a]^{-1/2})$. By definition of the two-sample sieve quasi MLE $\widehat{\alpha}_n$, we have

$$\begin{aligned}
0 \leq & \frac{1}{n+n_a}\textstyle\sum_{t=1}^{n+n_a}[\ell(Z_t;\widehat{\alpha}) - \ell(Z_t;\widehat{\alpha} \pm \varepsilon_n v_n^*)] \\
= & \ \mu_n\left(\ell(Z_t;\widehat{\alpha}) - \ell(Z_t;\widehat{\alpha} \pm \varepsilon_n v_n^*)\right) + E\left(\ell(Z_t;\widehat{\alpha}) - \ell(Z_t;\widehat{\alpha} \pm \varepsilon_n v_n^*)\right) \\
= & \ \mp\varepsilon_n \times \frac{1}{n+n_a}\textstyle\sum_{t=1}^{n+n_a}\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[v_n^*] + \mu_n\left(r[Z_t;\widehat{\alpha},\alpha_0] - r[Z_t;\widehat{\alpha} \pm \varepsilon_n v_n^*,\alpha_0]\right) \\
& + E\left(r[Z_t;\widehat{\alpha},\alpha_0] - r[Z_t;\widehat{\alpha} \pm \varepsilon_n v_n^*,\alpha_0]\right).
\end{aligned}$$

In the following we will show that:

$$\frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^*-\upsilon^*] = o_P(\frac{1}{\sqrt{n+n_a}}); \tag{A.9}$$

$$E\left(r[Z_t;\widehat{\alpha},\alpha_0]-r[Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*,\alpha_0]\right) = \pm\varepsilon_n\left\langle\widehat{\alpha}-\alpha_0,\upsilon^*\right\rangle_2+\varepsilon_n o_P(\frac{1}{\sqrt{n+n_a}}); \tag{A.10}$$

$$\mu_n\left(r[Z_t;\widehat{\alpha},\alpha_0]-r[Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*,\alpha_0]\right) = \varepsilon_n\times o_P(\frac{1}{\sqrt{n+n_a}}). \tag{A.11}$$

Notice that assumptions 3.1, 3.2(ii)(iii), and 3.6 imply $E\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]\right)=0$. Under (A.9) - (A.11) we have:

$$\begin{aligned}0 &\leq \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}[\ell(Z_t;\widehat{\alpha})-\ell(Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*)]\\ &= \mp\varepsilon_n\times\mu_n\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]\right)\pm\varepsilon_n\times\left\langle\widehat{\alpha}-\alpha_0,\upsilon^*\right\rangle_2+\varepsilon_n\times o_P(\frac{1}{\sqrt{n+n_a}}).\end{aligned}$$

Hence

$$\sqrt{n+n_a}\left\langle\widehat{\alpha}-\alpha_0,\upsilon^*\right\rangle_2=\sqrt{n+n_a}\mu_n\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]\right)+o_P(1)\Rightarrow N\left(0,\sigma_*^2\right),$$

$$\sigma_*^2\equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon^*]\right)^2\right\}=(\upsilon_\theta^*)^TE\left[\mathcal{S}_{\theta_0}^T\mathcal{S}_{\theta_0}\right](\upsilon_\theta^*)=\lambda^T(V_*)^{-1}I_*(V_*)^{-1}\lambda.$$

Thus, assumptions 3.2(i), 3.7, and 3.9 together imply that $\sigma_*^2<\infty$ and

$$\sqrt{n+n_a}\lambda^T(\widehat{\theta}_n-\theta_0)=\sqrt{n+n_a}\left\langle\widehat{\alpha}-\alpha_0,\upsilon^*\right\rangle_2+o_P(1)\Rightarrow N\left(0,\sigma_*^2\right).$$

To complete the proof, it remains to establish (A.9) - (A.11). Notice that (A.9) is implied by the Chebyshev inequality, i.i.d. data, and assumptions 3.10 and 3.13. For (A.10) and (A.11) we notice that$r[Z_t;\widehat{\alpha},\alpha_0]-r[Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*,\alpha_0]=\mp\varepsilon_n\times\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]$, in which $\widetilde{\alpha}\in\mathcal{N}_{0n}$ is in between $\widehat{\alpha}$ and $\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*$, and $\overline{\alpha}\in\mathcal{N}_0$ is in between $\widetilde{\alpha}\in\mathcal{N}_{0n}$ and α_0. Therefore, for (A.10), by the definition of inner product $\langle\cdot,\cdot\rangle_2$, we have:

$$\begin{aligned}&E\left(r[Z_t;\widehat{\alpha},\alpha_0]-r[Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*,\alpha_0]\right)\\ &= \mp\varepsilon_n\times E\left(\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]\right)\\ &= \pm\varepsilon_n\times\left\langle\widetilde{\alpha}-\alpha_0,\upsilon_n^*\right\rangle_2\mp\varepsilon_n\times E\left(\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]-\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[\widetilde{\alpha}-\alpha_0,\upsilon_n^*]\right)\\ &= \pm\varepsilon_n\times\left\langle\widehat{\alpha}-\alpha_0,\upsilon_n^*\right\rangle_2\pm\varepsilon_n\times\left\langle\widetilde{\alpha}-\widehat{\alpha},\upsilon_n^*\right\rangle_2+o_P(\frac{\varepsilon_n}{\sqrt{n+n_a}})\\ &= \pm\varepsilon_n\times\left\langle\widehat{\alpha}-\alpha_0,\upsilon^*\right\rangle_2+O_P(\varepsilon_n^2)+o_P(\frac{\varepsilon_n}{\sqrt{n+n_a}})\end{aligned}$$

in which the last two equalities hold due to the definition of $\widetilde{\alpha}$, assumptions 3.10 and 3.12, and $\langle \widehat{\alpha} - \alpha_0, \upsilon_n^* - \upsilon^* \rangle_2 = o_P(\frac{1}{\sqrt{n+n_a}})$ and $||\upsilon_n^*||_2^2 \to ||\upsilon^*||_2^2 < \infty$. Hence, (A.10) is satisfied. For (A.11), we notice

$$\mu_n \left(r[Z_t; \widehat{\alpha}, \alpha_0] - r[Z_t; \widehat{\alpha} \pm \varepsilon_n \upsilon_n^*, \alpha_0] \right) = \mp \varepsilon_n \times \mu_n \left(\frac{d\ell(Z_t; \widetilde{\alpha})}{d\alpha}[\upsilon_n^*] - \frac{d\ell(Z_t; \alpha_0)}{d\alpha}[\upsilon_n^*] \right)$$

in which $\widetilde{\alpha} \in \mathcal{N}_{0n}$ is in between $\widehat{\alpha}$ and $\widehat{\alpha} \pm \varepsilon_n \upsilon_n^*$. Since the class $\left\{ \frac{d\ell(Z_t; \widetilde{\alpha})}{d\alpha}[\upsilon_n^*] : \widetilde{\alpha} \in \mathcal{A}_{0s} \right\}$ is Donsker under assumptions 3.1, 3.2, 3.6, and 3.7, and since

$$E\left\{ \left(\frac{d\ell(Z_t; \widetilde{\alpha})}{d\alpha}[\upsilon_n^*] - \frac{d\ell(Z_t; \alpha_0)}{d\alpha}[\upsilon_n^*] \right)^2 \right\} = E\left\{ \left(\frac{d^2\ell(Z_t; \overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha} - \alpha_0, \upsilon_n^*] \right)^2 \right\}$$

goes to zero as $||\widetilde{\alpha} - \alpha_0||_s$ goes to zero under assumption 3.11, we have (A.11) holds.

References

[1] Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica* 71, 1795–1843.

[2] Ai, C., and X. Chen (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," *Journal of Econometrics* 141, 5-43.

[3] Bissantz, N., T. Hohage, A. Munk and F. Ruymgaart (2007): "Convergence Rates of General Regularization Methods for Statistical Inverse Problems and Applications," *SIAM Journal on Numerical Analysis*, forthcoming.

[4] Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics*, vol. 5, ed. by J. J. Heckman and E. Leamer, Elsevier Science.

[5] Buzas, J., and L. Stefanski (1996): "Instrumental Variable Estimation in Generalized Linear Measurement Error Models," *Journal of the American Statistical Association* 91, 999–1006.

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

[7] Carroll, R. J., D. Ruppert, L. A. Stefanski and C. Crainiceanu, 2006, *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, CRI.

[8] Carroll, R. J., and L. A. Stefanski (1990): "Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors," *Journal of the American Statistical Association* 85, 652–663.

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

[10] Chen, X. (2006): "Large Sample Sieve Estimation of Semi-nonparametric Models," in *Handbook of Econometrics*, vol. 6, ed. by J. J. Heckman and E. Leamer, Elsevier Science.

[11] Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies* 72, 343–366.

[12] Chen, X., H. Hong, and A. Tarozzi (2007): "Semiparametric Efficiency in GMM Models with Nonclassical Measurement Error," forthcoming in *Annals of Statistics*.

[13] Cheng, C. L., Van Ness, J. W., 1999, *Statistical Regression with Measurement Error*, Arnold, London.

[14] Dunford, N., and J. T. Schwartz (1971): *Linear Operators, Part 3: Spectral Operators*. New York: John Wiley & Sons.

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

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

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

[18] Hu, Y. (2006): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables," working paper, University of Texas at Austin.

[19] Hu, Y., and G. Ridder (2006): "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," working paper, University of Southern California.

[20] Hu, Y., and S. M. Schennach (2007): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica,* forthcoming.

[21] 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.

[22] Lee, L.-F., and J. H. Sepanski (1995): "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data," *Journal of the American Statistical Association* 90 (429).

[23] Lehmann, E.L., and J. Romano (2005): *Testing Statistical Hypotheses*, 3rd ed. Springer: New York.

[24] Li, T., and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis* 65, 139–165.

[25] Li, T. (2002): "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models," *Journal of Econometrics* 110, 1–26.

[26] Liang, H., W. Hardle, and R. Carroll, 1999, "Estimation in a Semiparametric Partially Linear Errors-in-Variables Model," The Annals of Statistics, Vol. 27, No. 5, 1519-1535.

[27] Mattner, L. (1993): "Some Incomplete but Bounded Complete Location Families," *Annals of Statistics*, 21, 2158-2162.

[28] Newey, W., and J. Powell (2003): "Instrumental Variables Estimation of Nonparametric Models," *Econometrica* 71, 1557–1569.

[29] Schennach, S. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica* 72(1), 33–76.

[30] Shen, X. (1997): "On Methods of Sieves and Penalization," *Annals of Statistics* 25, 2555–2591.

[31] Shen, X., and W. Wong (1994) "Convergence Rate of Sieve Estimates," *The Annals of Statistics* 22, 580–615.

[32] 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.

[33] Van de Geer, S. (2000), *Empirical Processes in M-estimation*, Cambridge University Press.

[34] Wang, L., 2004, "Estimation of nonlinear models with Berkson measurement errors," The Annals of Statistics 32, no. 6, 2559–2579.

[35] Wang, L., and C. Hsiao (1995): "Simulation-Based Semiparametric Estimation of Nonlinear Errors-in-Variables Models," working paper, University of Southern California.

[36] Wang, N., X. Lin, R. Gutierrez, and R. Carroll, 1998, "Bias analysis and SIMEX approach in generalized linear mixed measurement error models," J. Amer. Statist. Assoc. 93, no. 441, 249–261.

[37] White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica* 50, 143–161.

Additional Material for the Paper by Carroll, Chen and Hu

B.1 The Dichotomous Case: an Illustration

We first illustrate our identification strategy by describing a special case in which all the variables X^*, X, W, Y are 0-1 dichotomous. Denote $W_j = \{j\}$ for $j = 0, 1$, then all the probability distributions $f_{X,Y|W_j}$, $f_{Y|X^*,W_j}$, $f_{X^*|W_j}$ and $f_{X|X^*}$ can be equivalently represented in terms of matrices $L_{X,Y|W_j}$, $L_{Y|X^*,W_j}$, $L_{X^*|W_j}$ and $L_{X|X^*}$:

$$\begin{aligned} L_{X,Y|W_j} &\equiv \begin{pmatrix} f_{X,Y|W_j}(0,0) & f_{X,Y|W_j}(0,1) \\ f_{X,Y|W_j}(1,0) & f_{X,Y|W_j}(1,1) \end{pmatrix}, \quad L_{X|X^*} \equiv \begin{pmatrix} f_{X|X^*}(0|0) & f_{X|X^*}(0|1) \\ f_{X|X^*}(1|0) & f_{X|X^*}(1|1) \end{pmatrix}, \\ L_{Y|X^*,W_j} &\equiv \begin{pmatrix} f_{Y|X^*,W_j}(0|0) & f_{Y|X^*,W_j}(0|1) \\ f_{Y|X^*,W_j}(1|0) & f_{Y|X^*,W_j}(1|1) \end{pmatrix}^T, \quad L_{X^*|W_j} \equiv \begin{pmatrix} f_{X^*|W_j}(0) & 0 \\ 0 & f_{X^*|W_j}(1) \end{pmatrix}, \end{aligned}$$

where the superscript ^{T} stands for the transpose of a matrix. Let $W_{aj} = \{j\}$ for $j = 0, 1$. We similarly define the matrix representations $L_{X_a,Y_a|W_{aj}}$, $L_{X_a|X_a^*}$, and $L_{X_a^*|W_{aj}}$ of the corresponding densities $f_{X_a,Y_a|W_{aj}}$, $f_{X_a|X_a^*}$, and $f_{X_a^*|W_{aj}}$ in the auxiliary sample. To simplify notation, in the following we use W_j instead of W_{aj} in the auxiliary sample, and denote

$$k_{X_a^*}(x^*) \equiv \frac{f_{X_a^*|W_1}(x^*) f_{X^*|W_0}(x^*)}{f_{X^*|W_1}(x^*) f_{X_a^*|W_0}(x^*)} \quad \text{for } x^* \in \{0, 1\}.$$

We first state an identification result for the dichotomous case.

Theorem A.1. Suppose that the random variables X^*, X, W, Y and X_a^*, X_a, W_a, Y_a all have supports $\{0, 1\}$, and the following conditions hold: (1) $f_{X|X^*,W,Y} = f_{X|X^*}$; (2) $f_{X_a|X_a^*,W_a,Y_a} = f_{X_a|X_a^*}$; (3) $f_{Y_a|X_a^*,W_a} = f_{Y|X^*,W}$; (4) for $j = 0, 1$, $L_{X,Y|W_j}$ and $L_{X_a,Y_a|W_j}$ *are invertible, and* $f_{X^*|W_j}(0)$, $f_{X_a^*|W_j}(0) \in (0, 1)$*; (5)* $k_{X_a^*}(0) \neq k_{X_a^*}(1)$; (6) $f_{X_a|X_a^*}(x^*|x^*) > 0.5$ *for* $x^* = 0, 1$. Then: $f_{X,W,Y}$ 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 $f_{X_a^*|W_j}$.

Theorem A.1 can be viewed as a special case of the general identification theorem 2.1; hence we shall discuss its conditions in the next subsection. Nevertheless, we sketch a proof of theorem A.1 here to illustrate our general identification strategy. Conditions (1), (2) and

(3) imply that for $j = 0, 1$, and for all $x, y \in \{0, 1\}$,

$$f_{X,Y|W=j}(x, y) = \sum_{x^*=0,1} f_{X|X^*}(x|x^*)\, f_{Y|X^*,W=j}(y|x^*) f_{X^*|W=j}(x^*), \tag{B.1}$$

$$f_{X_a,Y_a|W_a=j}(x, y) = \sum_{x^*=0,1} f_{X_a|X_a^*}(x|x^*)\, f_{Y|X^*,W=j}(y|x^*) f_{X_a^*|W_a=j}(x^*). \tag{B.2}$$

Since all the variables are 0-1 dichotomous and probabilities sum to one, Equations (B.1) and (B.2) involve 12 distinct known probability values of $f_{X,Y|W=j}$ and $f_{X_a,Y_a|W_a=j}$, and 12 distinct unknown values of $f_{X|X^*}$, $f_{Y|X^*,W=j}$, $f_{X^*|W=j}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a=j}$, which makes exact identification (unique solution) of the 12 distinct unknown values possible. However, equations (B.1) and (B.2) are nonlinear in the unknown values, thus we need additional restrictions (such as conditions (4), (5) and (6)) to ensure the existence of unique solution.

Using the matrix notations, equations (B.1) and (B.2) can be respectively expressed as

$$L_{X,Y|W_j} = L_{X|X^*} L_{X^*,Y|W_j} = L_{X|X^*} L_{X^*|W_j} L_{Y|X^*,W_j} \quad \text{for } j = 0, 1, \tag{B.3}$$

and

$$L_{X_a,Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*,Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{Y|X^*,W_j} \quad \text{for } j = 0, 1. \tag{B.4}$$

Condition (6) implies that $L_{X_a|X_a^*}$ is invertible, this, condition (4) and equations (B.3) - (B.4) imply that $L_{Y|X^*,W_j}$, $L_{X|X^*}$, $L_{X^*|W_j}$ and $L_{X_a^*|W_j}$ are invertible. Thus,

$$L_{X_a,Y_a|W_j} L^{-1}_{X,Y|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L^{-1}_{X^*|W_j} L^{-1}_{X|X^*} \quad \text{for } j = 0, 1.$$

We may further eliminate $L_{X|X^*}$, and obtain

$$\begin{aligned} L_{X_a,X_a} &\equiv \left(L_{X_a,Y_a|W_1} L^{-1}_{X,Y|W_1}\right) \left(L_{X_a,Y_a|W_0} L^{-1}_{X,Y|W_0}\right)^{-1} \\ &= L_{X_a|X_a^*} \begin{pmatrix} k_{X_a^*}(0) & 0 \\ 0 & k_{X_a^*}(1) \end{pmatrix} L^{-1}_{X_a|X_a^*}. \end{aligned} \tag{B.5}$$

Equation (B.5) provides an eigenvalue-eigenvector decomposition of the observed (or known) matrix L_{X_a,X_a}. Condition (5) implies that the eigenvalues are distinct. Notice that each eigenvector is a column in $L_{X_a|X_a^*}$, which is a conditional density; hence each eigenvector

is automatically normalized. Therefore, from the observed L_{X_a,X_a}, we can compute its eigenvalue-eigenvector decomposition as follows:

$$L_{X_a,X_a} = \begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix} \begin{pmatrix} k_{X_a^*}(x_1^*) & 0 \\ 0 & k_{X_a^*}(x_2^*) \end{pmatrix} \times \\ \times \begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix}^{-1}, \tag{B.6}$$

that is, the value of each entry on the right-hand side of equation (B.6) can be directly computed from the observed matrix L_{X_a,X_a}. The only ambiguity left is the value of the indices x_1^* and x_2^*, or the indexing of the eigenvalues and eigenvectors. Since for $j = 1, 2$, the values of $f_{X_a|X_a^*}(0|x_j^*)$ and $f_{X_a|X_a^*}(1|x_j^*)$ are known in equation (B.6), condition (6) pins down the index x_j^* to be: $x_j^* = 0$ if $f_{X_a|X_a^*}(0|x_j^*) > 0.5$ and $x_j^* = 1$ if $f_{X_a|X_a^*}(1|x_j^*) > 0.5$. Thus we have identified $L_{X_a|X_a^*}$ (i.e., $f_{X_a|X_a^*}$) from the decomposition of the observed matrix L_{X_a,X_a}. Next, we can identify $L_{X_a^*,Y_a|W_j}$ ($f_{X_a^*,Y_a|W_j}$) from equation (B.4) as $L_{X_a^*,Y_a|W_j} = L_{X_a|X_a^*}^{-1} L_{X_a,Y_a|W_j}$; hence the conditional density $f_{Y|X^*,W_j} = f_{Y_a|X_a^*,W_j}$ and the marginal density $f_{X_a^*|W_j}$ are identified. We can then identify $L_{X,X^*|W_j}$ ($f_{X,X^*|W_j}$) from equation (B.3) as $L_{X,X^*|W_j} = L_{X,Y|W_j} L_{Y|X^*,W_j}^{-1}$; hence the densities $f_{X|X^*}$ and $f_{X^*|W_j}$ are identified.

B.2 Discussion of Equation 2.2

For $i = 1$,$j = 2$, any two eigenvalues $k_{X_a^*}^{12}(x_1^*)$ and $k_{X_a^*}^{12}(x_2^*)$ of L_{X_a,X_a}^{12} may be the same if and only if $x_1^* = -x_2^*$. In other words, we cannot distinguish the eigenfunctions $f_{X_a|X_a^*}(\cdot|x_1^*)$ and $f_{X_a|X_a^*}(\cdot|x_2^*)$ in the decomposition of L_{X_a,X_a}^{12} if and only if $x_1^* = -x_2^*$. Since $k_{X_a^*}^{ij}(x^*)$ for $i = 1, j = 3$ is not symmetric around zero, the eigenvalues $k_{X_a^*}^{13}(x_1^*)$ and $k_{X_a^*}^{13}(x_2^*)$ of L_{X_a,X_a}^{13} are different for any $x_1^* = -x_2^*$. Notice that the operators L_{X_a,X_a}^{12} and L_{X_a,X_a}^{13} share the same eigenfunctions $f_{X_a|X_a^*}(\cdot|x_1^*)$ and $f_{X_a|X_a^*}(\cdot|x_2^*)$. Therefore, we may distinguish the eigenfunctions $f_{X_a|X_a^*}(\cdot|x_1^*)$ and $f_{X_a|X_a^*}(\cdot|x_2^*)$ with $x_1^* = -x_2^*$ in the decomposition of L_{X_a,X_a}^{13}. By combining the information obtained from the decompositions of L_{X_a,X_a}^{12} and L_{X_a,X_a}^{13}, we can distinguish the eigenfunctions corresponding to any two different values of x^*.

B.3 Further Discussion of Assumption 2.6

In the dichotomous case, Assumption 2.6 with zero conditional mode also implies the invertibility of $L_{X_a|X_a^*}$, i.e., Assumption 2.4(i). However, this is not true in the general discrete case. For the general discrete case, a comparable sufficient condition for the invertibility of $L_{X_a|X_a^*}$ is strictly diagonal dominance (i.e., the diagonal entries of $L_{X_a|X_a^*}$ are all larger than 0.5), but Assumption 2.6 with zero mode only requires that the diagonal entries of $L_{X_a|X_a^*}$ be the largest in each row, which cannot guarantee the invertibility of $L_{X_a|X_a^*}$ when the support of X_a^* contains more than 2 values.

B.4 Discussion of Assumptions 3.9-3.13

Assumption 3.9 is necessary for obtaining the $\sqrt{n}$ convergence of sieve quasi MLE $\widehat{\theta}_n$ to θ_0 and its asymptotic normality. Assumption 3.10 implies that the asymptotic bias of the Riesz representer is negligible. Assumptions 3.11 and 3.12 control the remainder term. Assumption 3.13 is automatically satisfied when the latent parametric model is correctly specified, since $E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}\left[\upsilon_n^* - \upsilon^*\right]\right)^2\right\} = \|\upsilon_n^* - \upsilon^*\|_2^2$ under correct specification.