IDENTIFICATION AND INFERENCE IN NONLINEAR MODELS USING TWO SAMPLES WITH NONCLASSICAL MEASUREMENT ERRORS*

By Xiaohong Chen† and Yingyao Hu

Yale University and Johns Hopkins University

This paper considers identification and inference 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, in 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, with no existence of instrumental variables nor 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. We also provide a sieve likelihood ratio model selection test to compare two possibly misspecified parametric latent models. A small Monte Carlo simulation is presented.

1. Introduction. The Measurement error problems are frequently encountered by researchers conducting empirical studies in 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

*The authors would like to thank R. Blundell, R. Carroll, P. Cross, S. Donald, B. Fitzenberger, E. Mammen, L. Nesheim, M. Stinchcombe, C. Taber, 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.

AMS 2000 subject classifications: Primary 62H12, 62D05; secondary 62G08

Keywords and phrases: data combination, nonlinear errors-in-variables model, nonclassical measurement error, nonparametric identification, misspecified parametric latent model, sieve likelihood estimation and inference

studies on identification and estimation of linear, nonlinear, and even nonparametric models with classical measurement errors (see, e.g., Fuller (1987), Cheng and Van Ness (1999), Wansbeek and Meijer (2000), Carroll, Ruppert, Stefanski and Crainiceanu (2006) for detailed reviews). However, numerous validation studies in economic 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 (see, e.g., Bound, Brown, and Mathiowetz (2001)). In fact, in many survey situations, a rational agent has an incentive to purposely report wrong values conditioning on his/her truth. 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, and both samples could have nonclassical measurement errors and the two samples could be arbitrarily correlated.

It is well known that, without additional parametric restrictions or sample information, a general nonlinear model cannot be identified in the presence of mismeasured covariates. There are currently three broad approaches to regaining identification of nonlinear EIV models. The first one is to impose parametric restrictions on measurement error distributions (see, e.g., Hsiao (1989), Fan (1991), Murphy and Van der Vaart (1996), Wang, Lin, Gutierrez and Carroll (1998), Liang, Hardle and Carroll (1999), Taupin (2001), 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., Amemiya and Fuller (1988), Carroll and Stefanski (1990), Hausman, Ichimura, Newey, and Powell (1991), Wang and Hsiao (1995), Buzas and Stefanski (1996), Li and Vuong (1998), Newey (2001), Li (2002), Wang (2004), Schennach (2004), Carroll, Ruppert, Crainiceanu, Tosteson and Karagas (2004), Lewbel (2007), Hu (2006) and Hu and Schennach (2006), to name only a few). The third approach to identifying nonlinear EIV models with nonclassical errors is to combine two samples (see, e.g., Hausman, Ichimura, Newey, and Powell (1991), Carroll and Wand (1991), Lee and Sepanski (1995), Chen, Hong, and Tamer (2005), Chen, Hong and Tarozzi (2007), Hu and Ridder (2006), and Ichimura and Martinez-Sanchis (2006), to name only a few).

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. See Carroll, Ruppert, and Stefanski (1995) and Ridder and Moffitt (2006) for a detailed survey of this approach.

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 subsamples of both samples to show that the measurement error distributions are identified. To be specific, we may identify the relationship between the measurement error distribution in the auxiliary sample and the ratio of the marginal distributions of latent true values in the subsamples. In fact, the ratio of the marginal distributions plays the role of an eigenvalue of an observed linear operator, while the measurement error distribution in the auxiliary sample is the corresponding eigenfunction. Therefore, the measurement error distribution may be identified through a diagonal decomposition of an observed linear operator under the normalization condition that the measurement error distribution in the auxiliary sample has zero mode (or zero median or mean). The latent nonlinear model of interest, $f_{Y|X^*,W}$, may then be nonparametrically identified. In this paper, we first illustrate our identification strategy using a nonlinear EIV model with nonclassical errors in discrete covariates of two samples. We then focus 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. But, 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 addition, we provide a sieve likelihood ratio model selection test to compare two possibly misspecified parametric nonlinear EIV models with nonclassical errors.

In this paper, for any two possibly vector-valued random variables A and B, we let $f_{A|B}$ denote the conditional density of A given B, f_A denote 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. The rest of the paper is organized as follows. 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 and the sieve likelihood ratio model selection test under possibly misspecified parametric latent models. Section 4 provides a Monte Carlo study and Section 5 briefly concludes. The Appendix contains the proofs of the main theorems.

2. Nonparametric Identification.

2.1. *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. For example, suppose that we are interested in the effect of the latent true college education level X^* on the employment status Y with the marital status W as a covariate, i.e., $f_{Y|X^*,W}$.

Instead of X^* we observe self-reported college education level X.

In this illustration subsection, we use italic letters to highlight all the assumptions imposed for the nonparametric identification of $f_{Y|X^*,W}$, while detailed discussions of the assumptions are postponed to subsection 2.2. First, we assume that *the measurement error in X is independent of all other variables in the model conditional on the true value X^*, i.e., $f_{X|X^*,W,Y} = f_{X|X^*}$*. In our simple example, this assumption implies that all the people with the same latent true education level have the same pattern of misreporting, although the true education level could depend on other individual characteristics. Under this assumption, the probability distribution of the observables equals

$$f_{X,W,Y}(x,w,y) = \sum_{x^*=0,1} f_{X|X^*}(x|x^*) f_{X^*,W,Y}(x^*,w,y) \quad \text{for all } x,w,y. \tag{2.1}$$

We define the matrix representations of $f_{X|X^*}$ as follows:

$$L_{X|X^*} = \begin{pmatrix} f_{X|X^*}(0|0) & f_{X|X^*}(0|1) \\ f_{X|X^*}(1|0) & f_{X|X^*}(1|1) \end{pmatrix}.$$

Notice that the matrix $L_{X|X^*}$ contains the same information as the conditional density $f_{X|X^*}$. Equation (2.1) becomes for all w, y

$$\begin{pmatrix} f_{X,W,Y}(0,w,y) \\ f_{X,W,Y}(1,w,y) \end{pmatrix} = L_{X|X^*} \times \begin{pmatrix} f_{X^*,W,Y}(0,w,y) \\ f_{X^*,W,Y}(1,w,y) \end{pmatrix}. \tag{2.2}$$

Equation (2.2) implies that the density $f_{X^*,W,Y}$ would be identified provided that $L_{X|X^*}$ would be identifiable and invertible.

Equation (2.1) is equivalent to, for the subsamples of the married ($W = 1$) and of the unmarried ($W = 0$)

$$f_{X,Y|W=j}(x,y) = \sum_{x^*=0,1} f_{X|X^*}(x|x^*)\, f_{Y|X^*,W=j}(y|x^*) f_{X^*|W=j}(x^*), \tag{2.3}$$

in which $f_{X,Y|W=j}(x,y) \equiv f_{X,Y|W}(x,y|j)$ and $j = 0, 1$. By counting the numbers of knowns and unknowns in equation (2.3), one can see that the unknown densities $f_{X|X^*}$, $f_{Y|X^*,W=j}$ and $f_{X^*|W=j}$ cannot be identified using the primary sample alone.

In the auxiliary sample, we assume that *the measurement error in* X_a *satisfies the same conditional independence assumption as that in* X*, i.e.,* $f_{X_a|X_a^*,W_a,Y_a} = f_{X_a|X_a^*}$. Furthermore, we link the two samples by a stable assumption that *the effect of interest, i.e., the distribution of the employment status conditional on the true education level and the marital status, is the same in the two samples, i.e.,* $f_{Y_a|X_a^*,W_a}(y|x^*,w) = f_{Y|X^*,W}(y|x^*,w)$ *for all* y, x^*, w. Therefore, we have for the subsamples of the married ($W_a = 1$) and of the unmarried ($W_a = 0$):

(2.4)

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

Since all the variables are 0-1 dichotomous and probabilities sum to one, Equations (2.3) and (2.4) 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 possibly. However, equations (2.3) and (2.4) are nonlinear in the unknown values, we need additional restrictions to ensure the existence of unique solution.

Denote $W_j = \{j\}$ for $j = 0, 1$. Define the matrix representations of relevant densities for the subsamples of the married (W_1) and of the unmarried (W_0) in the primary sample as follows: for $j = 0, 1$,

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

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.

Using the matrix notations, equation (2.3) becomes for $j = 0, 1$,

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

that is

$$L_{X,Y|W_j} = L_{X|X^*} L_{X^*,Y|W_j} = L_{X|X^*} L_{X^*|W_j} L_{Y|X^*,W_j}. \tag{2.5}$$

Similarly, equation (2.4) becomes

$$L_{X_a,Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*,Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{Y|X^*,W_j}. \tag{2.6}$$

We assume that *the observable matrices* $L_{X,Y|W_j}$ *and* $L_{X_a,Y_a|W_j}$ *are invertible, that the diagonal matrices* $L_{X^*|W_j}$ *and* $L_{X_a^*|W_j}$ *are invertible, and that* $L_{X_a|X_a^*}$ *is invertible.* Then equations (2.6) and (2.5) imply that $L_{Y|X^*,W_j}$ and $L_{X|X^*}$ are invertible. We can then eliminate $L_{Y|X^*,W_j}$, to have for $j = 0, 1$

$$L_{X_a,Y_a|W_j} L^{-1}_{X,Y|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L^{-1}_{X^*|W_j} L^{-1}_{X|X^*}.$$

Since this equation holds for $j = 0, 1$, we may then eliminate $L_{X|X^*}$, to have

$$
\begin{aligned}
& L_{X_a,X_a} \\
\equiv\ & \left(L_{X_a,Y_a|W_1} L^{-1}_{X,Y|W_1}\right)\left(L_{X_a,Y_a|W_0} L^{-1}_{X,Y|W_0}\right)^{-1} \\
=\ & L_{X_a|X_a^*}\left(L_{X_a^*|W_1} L^{-1}_{X^*|W_1} L_{X^*|W_0} L^{-1}_{X_a^*|W_0}\right) L^{-1}_{X_a|X_a^*} \\
=\ & \begin{pmatrix} f_{X_a|X_a^*}(0|0) & f_{X_a|X_a^*}(0|1) \\ f_{X_a|X_a^*}(1|0) & f_{X_a|X_a^*}(1|1) \end{pmatrix} \begin{pmatrix} k_{X_a^*}(0) & 0 \\ 0 & k_{X_a^*}(1) \end{pmatrix} \times \\
& \times \begin{pmatrix} f_{X_a|X_a^*}(0|0) & f_{X_a|X_a^*}(0|1) \\ f_{X_a|X_a^*}(1|0) & f_{X_a|X_a^*}(1|1) \end{pmatrix}^{-1},
\end{aligned}
\tag{2.7}
$$

with

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

Notice that the matrix L_{X_a,X_a} on the left-hand side of the equation (2.7) can be viewed as observed given the data. Equation (2.7) provides an eigenvalue-eigenvector decomposition of L_{X_a,X_a}. If such a decomposition is unique, then we may identify (or solve) $L_{X_a|X_a^*}$, i.e., $f_{X_a|X_a^*}$, from the observed matrix L_{X_a,X_a}.

To ensure a unique eigenvalue-eigenvector decomposition of L_{X_a,X_a}, we assume that *the eigenvalues are distinctive; i.e.,* $k_{X_a^*}(0) \neq k_{X_a^*}(1)$. This assumption requires that the distributions of the latent education level of the married or the unmarried in the primary sample are different from those in the auxiliary sample, and that the distribution of the latent education level of the married is different from that of the unmarried in one of the two samples. 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, for an observed L_{X_a,X_a}, we may have an eigenvalue-eigenvector decomposition as follows:

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

The value of each entry on the right-hand side of equation (2.8) can be directly computed from the observed matrix L_{X_a,X_a}. The only ambiguity left in equation (2.8) is the value of the indices x_1^* and x_2^*, or the indexing of the eigenvalues and eigenvectors. In other words, the identification of $f_{X_a|X_a^*}$ boils down to finding a 1-to-1 mapping between the two sets of indices of the eigenvalues and eigenvectors: $\{x_1^*, x_2^*\} \Longleftrightarrow \{0,1\}$. For this, we make a normalization assumption that *people with (or without) true college education in the auxiliary sample are more likely to report that they have (or do not have) college education; i.e.,* $f_{X_a|X_a^*}(x^*|x^*) > 0.5$ *for* $x^* = 0,1$. (This assumption also implies the invertibility of $L_{X_a|X_a^*}$.) Since the values of $f_{X_a|X_a^*}(0|x_1^*)$ and $f_{X_a|X_a^*}(1|x_1^*)$ are known in equation (2.8), this assumption pins down the index x_1^* as follows:

$$
x_1^* = \begin{cases} 0 & \text{if } f_{X_a|X_a^*}(0|x_1^*) > 0.5 \\ 1 & \text{if } f_{X_a|X_a^*}(1|x_1^*) > 0.5 \end{cases}.
$$

The value of x_2^* may be found in the same way. In summary, 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}.

After identifying $L_{X_a|X_a^*}$, we can then identify $L_{X_a^*,Y_a|W_j}$ or $f_{X_a^*,Y_a|W_j}$ from equation (2.6) 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. Moreover, we can then identify $L_{X,X^*|W_j}$ or $f_{X,X^*|W_j}$ from equation (2.5) 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.

This simple example with dichotomous variables demonstrates that we can nonparametrically identify $f_{Y|X^*,W} = f_{Y_a|X_a^*,W_a}$, $f_{X|X^*}$ and $f_{X_a|X_a^*}$ using the two samples in which both samples contain nonclassical measurement errors. We next show that such a nonparametric identification strategy is in fact generally applicable.

2.2. *The continuous latent regressor case.* We are interested in identifying a latent (structural) 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. *$f_{X|X^*,W,Y}(x|x^*, w, y) = f_{X|X^*}(x|x^*)$ for all $x \in \mathcal{X}$, $x^* \in \mathcal{X}^*$, $w \in \mathcal{W}$, and $y \in \mathcal{Y}$.*

Assumption 2.1 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) X_a^*, W_a, and Y_a have the same supports as X^*, W, and Y, respectively; (ii) $f_{X_a|X_a^*,W_a,Y_a}(x|x^*, w, y) = f_{X_a|X_a^*}(x|x^*)$ for all $x \in \mathcal{X}$, $x^* \in \mathcal{X}^*$, $w \in \mathcal{W}$, and $y \in \mathcal{Y}$.*

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

ASSUMPTION 2.3. *$f_{Y_a|X_a^*,W_a}(y|x^*, w) = f_{Y|X^*,W}(y|x^*, w)$ for all $x^* \in \mathcal{X}^*$, $w \in \mathcal{W}$, and $y \in \mathcal{Y}$.*

We note that, under assumption 2.3, the joint distributions of the true value X^* and covariate W in the primary sample may still be different from those of X_a^* and W_a in the auxiliary sample.

Let $\mathcal{L}^p(\mathcal{X})$, $1 \le p < \infty$ 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 $\sup_{x\in\mathcal{X}} |h(x)| < \infty$. Then it is clear that for any fixed $w \in \mathcal{W}$, $y \in \mathcal{Y}$, $f_{X,W,Y}(\cdot, w, y) \in \mathcal{L}^p(\mathcal{X})$, and $f_{X^*,W,Y}(\cdot, w, y) \in \mathcal{L}^p(\mathcal{X}^*)$ for all $1 \le p \le \infty$. Let $\mathcal{H}_X \subseteq \mathcal{L}^2(\mathcal{X})$ and $\mathcal{H}_{X^*} \subseteq \mathcal{L}^2(\mathcal{X}^*)$. Define the integral operator $L_{X|X^*} : \mathcal{H}_{X^*} \to \mathcal{H}_X$ as:

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

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

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

We also define the operators $L_{X_a|X_a^*}$, $L_{X_a,Y_a|W_j}$, $L_{Y_a|X_a^*,W_j}$, and $L_{X_a^*|W_j}$ for the auxiliary sample in the same way as their counterparts for the primary sample. Notice that operators $L_{X^*|W_j}$ and $L_{X_a^*|W_j}$ are diagonal operators.

Under the operators formulation, assumption 2.1 implies

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

in the primary sample; assumptions 2.2 and 2.3 imply

$$L_{X_a,Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{Y|X^*,W_j}$$

in the auxiliary sample. We assume

Assumption 2.4. *$L_{X_a|X_a^*} : \mathcal{H}_{X_a^*} \to \mathcal{H}_{X_a}$ is injective, i.e., the set $\{h \in \mathcal{H}_{X_a^*} : L_{X_a|X_a^*}h = 0\} = \{0\}$.*

Assumption 2.4 is equivalent to assume that the linear operator $L_{X_a|X_a^*}$ is *invertible*. Recall that the conditional expectation operator of X_a^* given X_a, $E_{X_a^*|X_a} : \mathcal{L}^2(\mathcal{X}^*) \to \mathcal{L}^2(\mathcal{X}_a)$, is defined as

$$\begin{aligned} \{E_{X_a^*|X_a} h'\}(x) &= \int_{\mathcal{X}^*} f_{X_a^*|X_a}(x^*|x)\, h'(x^*) dx^* \\ &= E[h'(X_a^*)\,|X_a = x] \text{ for any } h' \in \mathcal{L}^2(\mathcal{X}^*),\, x \in \mathcal{X}_a. \end{aligned}$$

We have $\{L_{X_a|X_a^*} h\}(x) = \int_{\mathcal{X}^*} f_{X_a|X_a^*}(x|x^*)\, h(x^*)\, dx^* = E\left[\frac{f_{X_a}(x) h(X_a^*)}{f_{X_a^*}(X_a^*)} |X_a = x\right]$ for any $h \in \mathcal{H}_{X_a^*}$, $x \in \mathcal{X}_a$. Assumption 2.4 is equivalent to $E\left[h(X_a^*)\, \frac{f_{X_a}(X_a)}{f_{X_a^*}(X_a^*)} |X_a\right] = 0$ implies $h = 0$. If $0 < f_{X_a^*}(x^*) < \infty$ over $int(\mathcal{X}^*)$ and $0 < f_{X_a}(x) < \infty$ over $int(\mathcal{X}_a)$ (which are very minor restrictions), then assumption 2.4 is the same as the identification condition imposed in Newey and Powell (2003), and Carrasco, Florens, and Renault (2006), among others. As these authors point out, this condition is implied by 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. 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 is the *bounded completeness* of the conditional density $f_{X_a^*|X_a}$; see, e.g., Blundell, Chen, and Kristensen (2007) and Chernozhukov, Imbens, and Newey (2007). Distributions that are complete are automatically bounded complete. There are much larger families of distributions that are bounded complete (and some of them may not be complete). See, e.g., Lehmann (1986, page 173), Hoeffding (1977) and Mattner (1993) for many examples. When X_a and X_a^* are discrete, assumption 2.4 requires that the support of X_a is not smaller than that of X_a^*.

ASSUMPTION 2.5. *(i)* $f_{X^*|W_j} > 0$ *and* $f_{X_a^*|W_j} > 0$*; (ii)* $L_{X,Y|W_j}$ *is injective; (iii)* $L_{X,Y|W_j}$ *is injective.*

Assumptions 2.4 and 2.5 imply that $L_{Y|X^*,W_j}$ and $L_{X|X^*}$ are invertible. In the Appendix we establish the diagonalization of an observed operator

$L^{ij}_{X_a,X_a}$:

$$L^{ij}_{X_a,X_a} = L_{X_a|X_a^*} L^{ij}_{X_a^*} L^{-1}_{X_a|X_a^*} \quad \text{for all } i,j,$$

where the operator $L^{ij}_{X_a^*} \equiv \left(L_{X_a^*|W_j} L^{-1}_{X^*|W_j} L_{X^*|W_i} L^{-1}_{X_a^*|W_i} \right)$ is a diagonal operator defined as: $\left(L^{ij}_{X_a^*} h \right)(x^*) = k^{ij}_{X_a^*}(x^*)\, h(x^*)$ with

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

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

ASSUMPTION 2.6. $\sup_{x^* \in \mathcal{X}^*} k^{ij}_{X_a^*}(x^*) < \infty$ *for all* $i, j \in \{1, 2, ..., J\}$.

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.

ASSUMPTION 2.7. *For any* $x_1^* \neq x_2^*$, *there exist* $i, j \in \{1, 2, ..., J\}$, *such that* $k^{ij}_{X_a^*}(x_1^*) \neq k^{ij}_{X_a^*}(x_2^*)$.

Assumption 2.7 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 assumptions 2.6 and 2.7 are easily satisfied. Suppose that the distribution of X^* in the primary sample is the standard normal, i.e., $f_{X^*|w_j}(x^*) = \psi(x^*)$ for $j = 1, 2, 3$, where ψ is the probability density function of the standard normal, and the distribution of X_a^* in the auxiliary sample is for $0 < \sigma < 1$ and $\mu \neq 0$

$$f_{X_a^*|w_j}(x^*) = \begin{cases} \psi(x^*) & \text{for } j = 1 \\ \sigma^{-1}\psi(\sigma^{-1}x^*) & \text{for } j = 2 \\ \psi(x^* - \mu) & \text{for } j = 3 \end{cases}. \tag{2.9}$$

It is obvious that assumption 2.6 is satisfied with

$$k_{X_a^*}^{ij}(x^*) = \begin{cases} \sigma^{-1}\exp\left(-\frac{1-\sigma^{-2}}{2}(x^*)^2\right) & \text{for } i = 1,\ j = 2 \\ \frac{\psi(x^*-\mu)}{\psi(x^*)} & \text{for } i = 1,\ j = 3 \end{cases}. \tag{2.10}$$

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^*.

REMARK 2.1. ***(1)*** *Assumption 2.7 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 2.7 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 2.7 is not satisfied.* ***(2)*** *Assumption 2.7 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 2.7 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 2.7 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 2.7 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 overreporting 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 2.8. *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 2.8 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. Recall that, in the dichotomous case, assumption 2.8 with zero conditional mode also implies the invertibility of $L_{X_a|X_a^*}$ (i.e., assumption 2.4). 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 2.8 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.

We obtain the following identification result.

THEOREM 2.1. *Suppose assumptions 2.1–2.8 hold. Then, the densities* $f_{X,W,Y}$ *and* f_{X_a,W_a,Y_a} *uniquely determine* $f_{Y|X^*,W}$, $f_{X|X^*}$, *and* $f_{X_a|X_a^*}$.

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. The identification theorem is still valid. 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 and Inference. 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 2.1–2.8. 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).

In this section we provide a root-n consistent and asymptotically normally distributed sieve MLE of θ_0, regardless of whether the latent parametric model $g(y|x^*, w; \theta)$ is correctly specified or not. When $g(y|x^*, w; \theta)$ is misspecified, the estimator is better called the "sieve quasi MLE" than the "sieve MLE." (In this paper we have used both terminologies since we allow the latent model $g(y|x^*, w; \theta)$ to either correctly or incorrectly specify the true latent conditional density $f_{Y|X^*,W}$.) Under the correct specification of the latent model, we show that the sieve MLE of θ_0 is automatically semiparametrically efficient, and provide a simple consistent estimator of its asymptotic variance. In addition, we provide a sieve likelihood ratio model selection test of two non-nested parametric specifications of $f_{Y|X^*,W}$ when both could be misspecified.

3.1. *Sieve likelihood estimation under possible misspecification.* 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 value, 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 introduce a dummy random variable S, with $S = 1$ indicating the primary sample and $S = 0$ indicating the auxiliary sample. Then we have a big combined sample

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

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 $p \equiv \Pr(S_t = 1) \in (0, 1)$. Then the observed joint likelihood for α_0 is

$$\prod_{t=1}^{n+n_a} \left[p \times f(X_t, W_t, Y_t|S_t = 1; \alpha_0)\right]^{S_t} \left[(1 - p) \times f(X_t, W_t, Y_t|S_t = 0; \alpha_0)\right]^{1-S_t},$$

in which

$$f(X, W, Y|S = 1; \alpha_0) = f_W(W) \int f_{01}(X|x^*) g(Y|x^*, W; \theta_0) f_{02}(x^*|W) dx^*,$$

$$f(X, W, Y|S = 0; \alpha_0) = f_{W_a}(W) \int f_{01a}(X|x^*) g(Y|x^*, W; \theta_0) f_{02a}(x^*|W) dx^*.$$

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 (1993, 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)-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:

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

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

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

$$\mathcal{F}_{1a} = \left\{ \begin{array}{c} f_{1a}(\cdot|\cdot) \in \Lambda^{\gamma_{1a}}_c (\mathcal{X}_a \times \mathcal{X}^*) : \textit{ assumptions 2.4, 2.8 hold,} \\ f_{1a}(\cdot|x^*) > 0, \ \int_{\mathcal{X}_a} f_{1a}(x|x^*)dx = 1 \textit{ for all } x^* \in \mathcal{X}^* \end{array} \right\},$$

$$\mathcal{F}_2 = \left\{ \begin{array}{c} f_2(\cdot|w) \in \Lambda^{\gamma_2}_c (\mathcal{X}^*) : \textit{ assumptions 2.6, 2.7 hold,} \\ f_2(\cdot|w) > 0, \ \int_{\mathcal{X}^*} f_2(x^*|w)\, dx^* = 1 \textit{ for all } w \in \mathcal{W} \end{array} \right\},$$

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

Denote $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2$ and $\alpha = \left(\theta^T, f_1, f_{1a}, f_2, f_{2a}\right)^T$. Then the log-joint likelihood for $\alpha \in \mathcal{A}$ is given by:

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

in which

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

$$\ell_p(Z_t; \theta, f_1, f_2) = \ln \int f_1(X_t|x^*) g(Y_t|x^*, W_t; \theta) f_2(x^*|W_t) dx^* + \ln f_W(W_t),$$

$$\ell_a(Z_t; f_{1a}, f_{2a}) = \ln \int f_{1a}(X_t|x^*_a) g(Y_t|x^*_a, W_t; \theta) f_{2a}(x^*_a|W_t) dx^*_a + \ln f_{W_a}(W_t).$$

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 $i-th$ 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

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

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:

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

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.* 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. Assumption 3.3(ii) requires that the log density be Hölder continuous under the metric $\|\cdot\|_s$ over the sieve space. The following consistency lemma is a direct application of lemma A.1 of Newey and Powell (2003) or theorem 3.1 (or remark 3.1(4), remark 3.3) of Chen (2006); hence, we omit its proof.

LEMMA 3.1. *Let $\widehat{\alpha}_n$ be the two-sample sieve MLE. 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. *$\sup_{\widetilde{\alpha}\in\mathcal{A}_{0s}}\sup_{\alpha\in\mathcal{A}_{0sn}}\left|\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}\left[\frac{\alpha-\alpha_0}{\|\alpha-\alpha_0\|_s}\right]\right|\le U(Z_t)$ for a random variable $U(Z_t)$ with $E\{[U(Z_t)]^2\}<\infty$.*

ASSUMPTION 3.8. *(i) $\sup_{v\in\mathcal{A}_{0s}:||v||_s=1}-E\left(\frac{d^2\ell(Z_t;\alpha_0)}{d\alpha d\alpha^T}[v,v]\right)\le C<\infty$; (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 $\|\alpha-\alpha_0\|_2\le\sqrt{C}\,\|\alpha-\alpha_0\|_s$ 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 $\alpha\in\mathcal{A}_{0sn}$, $C_1\|\alpha-\alpha_0\|_2^2\le E[\ell(Z_t;\alpha_0)-\ell(Z_t;\alpha)]\le C_2\|\alpha-\alpha_0\|_2^2$; that is, $\|\alpha-\alpha_0\|_2^2$ is equivalent to the Kullback-Leibler discrepancy on the local sieve space $\mathcal{A}_{0sn}$. The following convergence rate theorem is a direct application of theorem 3.2 of Shen and Wong (2004) or theorem 3.2 of Chen (2006) 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*

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

3.1.3. *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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\upsilon^* \equiv (\upsilon_\theta^*, \upsilon_h^*)$ with $\upsilon_\theta^* \equiv (V_*)^{-1}\lambda$ and $\upsilon_h^* \equiv -\mu^* \times \upsilon_\theta^*$. Then the Riesz representation theorem implies: $\lambda^T(\theta-\theta_0) = \langle\upsilon^*, \alpha-\alpha_0\rangle_2$ for all $\alpha \in \mathcal{A}$. In the appendix, we show that

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

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 $\alpha \in \mathcal{N}_{0n}$,*

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

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.*

Assumption 3.9 is critical for obtaining the $\sqrt{n}$ convergence of sieve quasi MLE $\widehat{\theta}_n$ to θ_0 and its asymptotic normality. We notice that it is possible that θ_0 is uniquely identified but assumption 3.9 is not satisfied. If this happens, θ_0 can still be consistently estimated, but the best achievable convergence rate is slower than the $\sqrt{n}-$rate. 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 $E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^* - \upsilon^*]\right)^2\right\} = \|\upsilon_n^* - \upsilon^*\|_2^2$ under correct specification. 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* $\sqrt{n+n_a}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, V_*^{-1} I_* V_*^{-1}\right)$.

3.1.4. *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. First we recall the Fisher metric $\|\cdot\|$ on $\mathcal{A}$: for any α_1, $\alpha_2 \in \mathcal{A}$,

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

and the Fisher norm–induced inner product:

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

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

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

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 under correct specification, $I_* \equiv E\left[\mathcal{S}_{\theta_0}^T \mathcal{S}_{\theta_0}\right] = V_*$, which is the semiparametric information bound for θ_0.

Given the expression of the density function, the pathwise first derivative at α_0 can be written as

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

See Appendix for the expressions of $\frac{d\ell_p(Z_t;\theta_0,f_{01},f_{02})}{d\alpha}[\alpha-\alpha_0]$ and $\frac{d\ell_a(Z_t;f_{01a},f_{02a})}{d\alpha}[\alpha-\alpha_0]$. Thus

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

with

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

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

Therefore, the influence function representation of our two-sample sieve MLE is:

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

and the asymptotic distribution of $\sqrt{n+n_a}\left(\widehat{\theta}_n - \theta_0\right)$ is $N\left(0, I_*^{-1}\right)$. 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 $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, [I_{*p} + \frac{1-p}{p} I_{*a}]^{-1}\right) = N\left(0, pI_*^{-1}\right)$.*

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

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

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 $\mathcal{F}_n \equiv \mathcal{F}_1^n \times \mathcal{F}_{1a}^n \times \mathcal{F}_2^n \times \mathcal{F}_2^n$. Denote

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

and

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

3.2. *Sieve likelihood ratio model selection test.* In many empirical applications, researchers often estimate different parametrically specified structure models in order to select one that fits the data the "best". We shall consider two non-nested, possibly misspecified, parametric latent structure models: $\{g_1(y|x^*, w; \theta_1) : \theta_2 \in \Theta_1\}$ and $\{g_2(y|x^*, w; \theta_2) : \theta_2 \in \Theta_2\}$. If X^* were observed without error in the primary sample, researchers could apply Vuong's (1989) likelihood ratio test to select a "best" parametric model that is closest to the true underlying conditional density $f_{Y|X^*,W}(y|x^*, w)$ according to the KLIC. In this subsection, we shall extend Vuong's result to the case in which X^* is not observed in either sample.

Consider two parametric families of models $\{g_j(y|x^*, w; \theta_j) : \theta_j \in \Theta_j\}$, Θ_j a compact subset of $\mathbb{R}^{d_{\theta_j}}$, $j = 1, 2$ for the latent true conditional density

$f_{Y|X^*,W}$. Define

$$\theta_{0j} \equiv \arg\max_{\theta_j \in \Theta_j} \int [\log g_j(y|x^*, w; \theta_j)] f_{Y|X^*,W}(y|x^*, w) dy.$$

According to Vuong (1989), the two models are *nested* if $g_1(y|x^*, w; \theta_{01}) = g_2(y|x^*, w; \theta_{02})$ for almost all $y \in \mathcal{Y}, x^* \in \mathcal{X}^*, w \in \mathcal{W}$; the two models are *non-nested* if $g_1(Y|X^*, W; \theta_{01}) \neq g_2(Y|X^*, W; \theta_{02})$ with positive probability.

For $j = 1, 2$, denote $\alpha_{0j} = (\theta_{0j}^T, f_{01}, f_{01a}, f_{02}, f_{02a})^T \in \mathcal{A}_j$ with $\mathcal{A}_j = \Theta_j \times \mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2$, and let $\ell_j(Z_t; \alpha_{0j})$ denote the log-likelihood according to model j evaluated at data Z_t. Following Vuong (1989), we select model 1 if H_0 holds, in which

$$H_0: \; E\left\{\ell_2(Z_t; \alpha_{02}) - \ell_1(Z_t; \alpha_{01})\right\} \leq 0,$$

and we select model 2 if H_1 holds, in which

$$H_1: \; E\left\{\ell_2(Z_t; \alpha_{02}) - \ell_1(Z_t; \alpha_{01})\right\} > 0.$$

For $j = 1, 2$, denote $\mathcal{A}_{j,n} = \Theta_j \times \mathcal{F}_1^n \times \mathcal{F}_{1a}^n \times \mathcal{F}_2^n \times \mathcal{F}_2^n$ and define the sieve quasi MLE for $\alpha_{0j} \in \mathcal{A}_j$ as

$$\begin{aligned}
\widehat{\alpha}_j &= \arg\max_{\alpha_j \in \mathcal{A}_{j,n}} \sum_{t=1}^{n+n_a} \ell_j(Z_t; \alpha_j) \\
&= \arg\max_{\alpha_j \in \mathcal{A}_{j,n}} \left[\sum_{t=1}^{n} \ell_{j,p}(Z_{pt}; \theta_j, f_1, f_2) + \sum_{t=1}^{n_a} \ell_{j,a}(Z_{at}; f_{1a}, f_{2a}) \right].
\end{aligned}$$

In the following, we denote $\sigma^2 \equiv Var\left(\ell_2(Z_t; \alpha_{02}) - \ell_1(Z_t; \alpha_{01})\right)$ and

$$\hat{\sigma}^2 = \frac{1}{n+n_a} \sum_{t=1}^{n+n_a} \left[\begin{array}{c} \{\ell_2(Z_t; \widehat{\alpha}_2) - \ell_1(Z_t; \widehat{\alpha}_1)\} \\ -\frac{1}{n+n_a} \sum_{s=1}^{n+n_a} \{\ell_2(Z_s; \widehat{\alpha}_2) - \ell_1(Z_s; \widehat{\alpha}_1)\} \end{array} \right]^2.$$

THEOREM 3.4. *Suppose both models 1 and 2 satisfy assumptions 3.1–3.8,* *and* $\sigma^2 < \infty$. *Then*

$$\begin{aligned} & \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \begin{pmatrix} \{\ell_2(Z_t;\widehat{\alpha}_2) - \ell_1(Z_t;\widehat{\alpha}_1)\} \\ -E\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\} \end{pmatrix} \\ = & \frac{1}{\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \begin{pmatrix} \{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\} \\ -E\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\} \end{pmatrix} + o_P(1) \\ & \xrightarrow{d} N\left(0, \sigma^2\right). \end{aligned}$$

Suppose models 1 and 2 are non-nested, then

$$\frac{1}{\hat{\sigma}\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \begin{pmatrix} \{\ell_2(Z_t;\widehat{\alpha}_2) - \ell_1(Z_t;\widehat{\alpha}_1)\} \\ -E\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\} \end{pmatrix} \xrightarrow{d} N(0,1).$$

Thus under the least favorable null hypothesis of $E\{\ell_2(Z_t;\alpha_{02}) - \ell_1(Z_t;\alpha_{01})\} = 0$, we have $\frac{1}{\hat{\sigma}\sqrt{n+n_a}} \sum_{t=1}^{n+n_a} \{\ell_2(Z_t;\widehat{\alpha}_2) - \ell_1(Z_t;\widehat{\alpha}_1)\} \xrightarrow{d} N(0,1)$, which can be used to provide a sieve likelihood ratio model selection test of H_0 against H_1.

4. Simulation. In this section we present a simulation study to illustrate the finite sample performance of the two-sample sieve MLE. The true latent probability density model $f_{Y|X^*,W}$ is:

$$f_{Y|X^*,W}(y|x^*, w; \theta) = \phi\left(y - m\left(x^*, w; \theta\right)\right),$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution and

$$m\left(X^*, W; \theta\right) = \beta_1 X^* + \beta_2 X^* W + \beta_3 \left(X^{*2} W + X^* W^2\right)/2,$$

in which $\theta = (\beta_1, \beta_2, \beta_3)^T$ is unknown and $W \in \{-1, 0, 1\}$. 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 $\theta_0 = (1,1,1)^T$, $X^*|W \sim N(0,1)$, and $\Pr(W = 1) = \Pr(W = 0) = 1/3$. The mismeasured value X equals

$$X = 0.1X^* + e^{-0.1X^*}\varepsilon \quad \text{with} \quad \varepsilon \sim N(0, 0.36).$$

In the auxiliary sample we generate W_a in the same way that we generate W in the primary sample. We set the unknown true conditional density $f_{X_a^*|W_a}$ as follows:

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

The mismeasured value X_a equals

$$X_a = X_a^* + 0.5e^{-X_a^*}\nu, \quad \text{with} \quad \nu \sim N(0,1),$$

which implies that x_a^* is the mode of the conditional density $f_{X_a|X_a^*}(\cdot|x_a^*)$.

We use the simple sieve expression $p_1^{k_{1,n}}(x_1,x_2)^T\beta_1 = \sum_{j=0}^{J_n}\sum_{k=0}^{K_n}\gamma_{jk}p_j(x_1-x_2)\,q_k(x_2)$ to approximate the conditional densities $f_{X|X^*}(x_1|x_2)$ and $f_{X_a|X_a^*}(x_1|x_2)$, with $k_{1,n} = (J_n+1)(K_n+1)$; and $p_2^{k_{2,n}}(x^*)^T\beta_2(w) = \sum_{k=1}^{k_{2,n}}\gamma_k(w)q_k(x^*)$ to approximate the conditional densities $f_{X^*|W_j=w}$, $f_{X_a^*|W_j=w}$ with $W_j = -1, 0, 1$. The bases $\{p_j(\cdot)\}$ and $\{q_k(\cdot)\}$ are Hermite polynomials bases.

The simulation results shown in Table 1 include three estimators. 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 squared 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 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 the last column, we also report the square root of the sum of the three mean square errors of β_1, β_2, and β_3. The simulation repetition times is 400. The simulation results show that the 2-sample sieve MLE has a much smaller bias (and a slightly bigger standard error) than the estimator ignoring measurement error. Moreover, the 2-sample sieve MLE has a smaller total root MSE than the inconsistent estimator. In summary, our 2-sample sieve MLE performs well in this Monte Carlo simulation.

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

true value of β:	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_3 = 1$
ignoring meas. error:			
– mean estimate	0.1753	0.3075	0.5953
– standard error	0.08422	0.1227	0.1879
– root mse	0.8290	0.7033	0.4461
infeasible MLE:			
– mean estimate	0.9998	1.001	1.000
– standard error	0.02792	0.03382	0.03549
– root mse	0.02792	0.03382	0.03549
2-sample sieve MLE:			
– mean estimate	1.024	1.038	0.9866
– standard error	0.08670	0.1229	0.2290
– root mse	0.08999	0.1286	0.2293

note: $J_n = 5$, $K_n = 3$ in $\widehat{f}_{X|X^*}$, $\widehat{f}_{X_a|X_a^*}$;
$k_{2,n} = 4$ for $\widehat{f}_{X^*|W}$, $\widehat{f}_{X_a^*|W_a}$.

5. Conclusion. This paper considers nonparametric identification and semiparametric estimation of a general nonlinear model using two random samples. Both samples consist of a dependent variable, some error-free covariates and an error-ridden covariate, in which the measurement error has unknown distribution and could be arbitrarily correlated with the latent true values. We provide reasonable conditions so that the latent nonlinear model is nonparametrically identified using the two samples. The advantage of our identification strategy is that, in addition to allowing for nonclassical measurement errors in both samples, neither sample is required to contain an accurate measurement of the latent true covariate, and only one measurement of the error-ridden covariate is assumed in each sample. Moreover, our identification result does not require that the primary sample contain an IV excluded from the nonlinear model of interest, nor does it require that the two samples be independent.

Since the latent nonlinear model is nonparametrically identified without imposing two independent samples, we could estimate the latent nonlinear model nonparametrically via two potentially correlated samples, provided that we impose some structure on the correlation of the two samples. In particular, the panel data structure in Horowitz and Markatou (1996) could be borrowed to model two correlated samples. We shall investigate this in future research.

6. Appendix: Mathematical Proofs.

Proof : (Theorem 2.1) Under assumption 2.1, the probability density of the observed vectors equals

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

For each value w_j of W, assumption 2.1 implies that

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

in the primary sample. Similarly, assumptions 2.2 and 2.3 imply that

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

in the auxiliary sample.

By equation (A.2) and the definition of the operators, we have, for any function h,

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

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 ($L_{Y|X^*,W_j}$).

Assumptions 2.4 and 2.5 imply that all the operators involved in equations (A.4) and (A.5) are invertible. Under assumptions 2.4 and 2.5, for any given W_j we can eliminate $L_{Y|X^*,W_j}$ in equations (A.4) and (A.5) to obtain

$$L_{X_a,Y_a|W_j} L_{X,Y|W_j}^{-1} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{X^*|W_j}^{-1} L_{X|X^*}^{-1}. \tag{A.6}$$

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

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

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 $L_{X_a^*}^{ij} \equiv \left(L_{X_a^*|W_j} L_{X^*|W_j}^{-1} L_{X^*|W_i} L_{X_a^*|W_i}^{-1} \right)$ is a diagonal operator defined as

$$\left(L_{X_a^*}^{ij} h \right)(x^*) \equiv k_{X_a^*}^{ij}(x^*)\, h(x^*) \tag{A.8}$$

with

$$k_{X_a^*}^{ij}(x^*) \equiv \frac{f_{X_a^*|W_j}(x^*)\, f_{X^*|W_i}(x^*)}{f_{X^*|W_j}(x^*)\, f_{X_a^*|W_i}(x^*)}.$$

Equation (A.7) implies a diagonalization 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; see, e.g., Dunford and Schwartz (1971). 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 2.6 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 2.7 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. Moreover, assumption 2.8 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 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:

$$f_{Y_a,X_a^*|W_a} = L_{X_a|X_a^*}^{-1} f_{X_a,Y_a|W_a}.$$

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

$$L_{X|X^*} L_{X^*|W_j} = L_{X,Y|W_j} L_{Y|X^*,W_j}^{-1}. \tag{A.9}$$

The left-hand side of equation (A.9) 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.9) 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) The proof is a simplified version of that for theorem 4.1 in Ai and Chen (2007). Recall the neighborhoods $\mathcal{N}_{0n} = \{\alpha \in \mathcal{A}_{0sn} : \|\alpha - \alpha_0\|_2 = o([n+n_a]^{-1/4})\}$ and $\mathcal{N}_0 = \{\alpha \in \mathcal{A}_{0s} : \|\alpha - \alpha_0\|_2 = o([n+n_a]^{-1/4})\}$. 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

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

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

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

In the following we will show that:

$$\begin{aligned}
\text{(A.10)} \qquad & \frac{1}{n+n_a}\sum_{t=1}^{n+n_a}\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^* - \upsilon^*] \\
&= o_P\left(\frac{1}{\sqrt{n+n_a}}\right); \\
\text{(A.11)} \qquad & E\left(r[Z_t;\widehat{\alpha},\alpha_0] - r[Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*,\alpha_0]\right) \\
&= \pm\varepsilon_n\left\langle\widehat{\alpha}-\alpha_0,\upsilon^*\right\rangle_2 + \varepsilon_n o_P\left(\frac{1}{\sqrt{n+n_a}}\right); \\
\text{(A.12)} \qquad & \mu_n\left(r[Z_t;\widehat{\alpha},\alpha_0] - r[Z_t;\widehat{\alpha}\pm\varepsilon_n\upsilon_n^*,\alpha_0]\right) \\
&= \varepsilon_n \times o_P\left(\frac{1}{\sqrt{n+n_a}}\right).
\end{aligned}$$

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.10) - (A.12) we have:

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

Hence

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

with

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

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

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

To complete the proof, it remains to establish (A.10) - (A.12). Notice that (A.10) is implied by the Chebyshev inequality, i.i.d. data, and assumptions 3.10 and 3.13. For (A.11) and (A.12) we notice that

$$\begin{aligned}
& r[Z_t;\widehat{\alpha},\alpha_0] - r[Z_t;\widehat{\alpha} \pm \varepsilon_n \upsilon_n^*, \alpha_0] \\
&= \ell(Z_t;\widehat{\alpha}) - \ell(Z_t;\widehat{\alpha} \pm \varepsilon_n \upsilon_n^*) - \frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\mp\varepsilon_n \upsilon_n^*] \\
&= \mp\varepsilon_n \times \left(\frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}[\upsilon_n^*] - \frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\upsilon_n^*]\right) \\
&= \mp\varepsilon_n \times \left(\frac{d^2\ell(Z_t;\overline{\alpha})}{d\alpha d\alpha^T}[\widetilde{\alpha} - \alpha_0, \upsilon_n^*]\right)
\end{aligned}$$

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.11), by the definition of inner product $\langle\cdot,\cdot\rangle_2$, we have:

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

in which the last two equalities hold due to the definition of $\widetilde{\alpha}$, assumptions 3.10 and 3.12, and

$$\langle \widehat{\alpha} - \alpha_0, \upsilon_n^* - \upsilon^*\rangle_2 = o_P\left(\frac{1}{\sqrt{n+n_a}}\right) \text{ and } ||\upsilon_n^*||_2^2 \to ||\upsilon^*||_2^2 < \infty.$$

Hence, (A.11) is satisfied. For (A.12), 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.12) holds.

For the sake of completeness, we write down the expressions of $\frac{d\ell_p(Z;\theta_0,f_{01},f_{02})}{d\alpha}[\alpha - \alpha_0]$ and $\frac{d\ell_a(Z;f_{01a},f_{02a})}{d\alpha}[\alpha - \alpha_0]$ that are needed in the calculation of the Riesz representer and the asymptotic efficient variance of the sieve MLE $\widehat{\theta}$ in subsection 3.1.4:

$$\begin{aligned}
& f_{X,Y|W}(X, Y|W; \theta_0, f_{01}, f_{02}) \times \frac{d\ell_p(Z; \theta_0, f_{01}, f_{02})}{d\alpha} [\alpha - \alpha_0] \\
= & \int_{\mathcal{X}^*} f_{01}(X|x^*) \frac{dg(Y|x^*, W; \theta_0)}{d\theta^T} f_{02}(x^*|W) dx^* [\theta - \theta_0] \\
& + \int_{\mathcal{X}^*} [f_1(X|x^*) - f_{01}(X|x^*)] \, g(Y|x^*, W; \theta_0) f_{02}(x^*|W) dx^* \\
& + \int_{\mathcal{X}^*} f_{01}(X|x^*) g(Y|x^*, W; \theta_0) \, [f_2(x^*|W) - f_{02}(x^*|W)] \, dx^*,
\end{aligned}$$

and

$$\begin{aligned}
& f_{X_a,Y_a|W_a}(X_a, Y_a|W_a; f_{01a}, f_{02a}) \times \frac{d\ell_a(Z; f_{01a}, f_{02a})}{d\alpha} [\alpha - \alpha_0] \\
= & \int_{\mathcal{X}^*} f_{01a}(X|x^*) \frac{dg(Y|x^*, W; \theta_0)}{d\theta^T} f_{02a}(x^*|W_a) dx^* \\
& + \int_{\mathcal{X}^*} [f_{1a}(X|x^*) - f_{01a}(X|x^*)] \, g(Y|x^*, W; \theta_0) f_{02a}(x^*|W_a) dx^* \\
& + \int_{\mathcal{X}^*} f_{01a}(X|x^*) g(Y|x^*, W; \theta_0) \, [f_{2a}(x^*|W_a) - f_{02a}(x^*|W_a)] \, dx^*.
\end{aligned}$$

Proof : (Theorem 3.4) Under stated assumptions, we have, for model $j = 1, 2$,

$$\frac{1}{\sqrt{n + n_a}} \sum_{t=1}^{n+n_a} \left(\begin{array}{c} \{\ell_j(Z_t; \widehat{\alpha}_j) - \ell_j(Z_t; \alpha_{0j})\} \\ -E\{\ell_j(Z_t; \widehat{\alpha}_j) - \ell_j(Z_t; \alpha_{0j})\} \end{array} \right) = o_P(1),$$

and

$$E\{\ell_j(Z_t;\widehat{\alpha}_j)-\ell_j(Z_t;\alpha_{0j})\}\asymp||\widehat{\alpha}_j-\alpha_{0j}||_2^2=o_P\left(\frac{1}{\sqrt{n+n_a}}\right)$$

thus

$$\begin{aligned}
&\frac{1}{\sqrt{n+n_a}}\sum_{t=1}^{n+n_a}\left(\{\ell_j(Z_t;\widehat{\alpha}_j)-E[\ell_j(Z_t;\alpha_{0j})]\}\right)\\
=\;&\frac{1}{\sqrt{n+n_a}}\sum_{t=1}^{n+n_a}\left(\{\ell_j(Z_t;\widehat{\alpha}_j)-\ell_j(Z_t;\alpha_{0j})\}-E\{\ell_j(Z_t;\widehat{\alpha}_j)-\ell_j(Z_t;\alpha_{0j})\}\right)\\
&+\frac{1}{\sqrt{n+n_a}}\sum_{t=1}^{n+n_a}\{\ell_j(Z_t;\alpha_{0j})-E[\ell_j(Z_t;\alpha_{0j})]\}\\
&+\sqrt{n+n_a}E\{\ell_j(Z_t;\widehat{\alpha}_j)-\ell_j(Z_t;\alpha_{0j})\}\\
=\;&\frac{1}{\sqrt{n+n_a}}\sum_{t=1}^{n+n_a}\{\ell_j(Z_t;\alpha_{0j})-E[\ell_j(Z_t;\alpha_{0j})]\}+o_P(1).
\end{aligned}$$

Under stated conditions, it is obvious that $\hat{\sigma}^2=\sigma^2+o_P(1)$. Suppose models 1 and 2 are non-nested, then $\sigma>0$. Thus,

$$\frac{1}{\hat{\sigma}\sqrt{n+n_a}}\sum_{t=1}^{n+n_a}\left(\begin{array}{c}\{\ell_2(Z_t;\widehat{\alpha}_2)-\ell_1(Z_t;\widehat{\alpha}_1)\}\\ -E\{\ell_2(Z_t;\alpha_{02})-\ell_1(Z_t;\alpha_{01})\}\end{array}\right)\xrightarrow{d}N(0,1).$$

REFERENCES

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

Ai, C., and X. Chen (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," forthcoming in Journal of Econometrics.

Amemiya, Y., and W. A. Fuller (1988): "Estimation for the Nonlinear Functional Relationship," Annals of Statistics 16, 147–160.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semiparametric Engel Curves with Endogenous Expenditure," forthcoming in Econometrica.

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.

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

Carrasco, M., J.-P. Florens, and E. Renault (2006): "Linear Inverse Problems and Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," in Handbook of Econometrics, vol. 6, ed. by J. J. Heckman, and E. Leamer, Elsevier Science.

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.

Carroll, R. J., D. Ruppert, and L. A. Stefanski (1995): Measurement Error in Nonlinear Models: A Modern Perspective. New York: Chapman & Hall.

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

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.

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.

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.

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

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

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

Chernozhukov, V., G. Imbens, and W. Newey (2007): "Instrumental Variable Identification and Estimation of Nonseparable Models via Quantile Conditions," Journal of Econometrics.

Dunford, N., and J. T. Schwartz (1971): Linear Operators. New York: John Wiley & Sons.

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

Fuller, W., 1987, Measurement error models. New York: John Wiley & Sons.

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.

Hoeffding, W. (1977): “Some Incomplete and Boundedly Complete Families of Distributions,” Annals of Statistics, 5, 278-291.

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

Horowitz, J., and M. Markatou (1996): “Semiparametric Estimation of Regression Models for Panel Data,” Review of Economic Studies 63, 145–168.

Hsiao, C. (1989): “Consistent Estimation for Some Nonlinear Errors-in-Variables Models,” Journal of Econometrics 41, 159–185.

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

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

Hu, Y., and S. M. Schennach (2006): “Identification and Estimation of Nonclassical Nonlinear Errors-in-Variables Models with Continuous Distributions Using Instruments,” Cemmap working paper (Centre for Microdata Methods and Practice).

Ichimura, H., and E. Martinez-Sanchis (2006): “Identification and Estimation of GMM Models by Combining Two Data Sets,” working paper, University College London.

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).

Lehmann, E.L. (1986): Testing Statistical Hypothesis, 2nd ed. Wiley: New York.

Lewbel, A. (2007): “Estimation of Average Treatment Effects with Misclassification,” Econometrica, 75, 537-551.

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

Li, T. (2002): “Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models,” Journal of Econometrics 110, 1–26.

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.

Mattner, L. (1993): “Some Incomplete but Bounded Complete Location Families,” Annals of Statistics, 21, 2158-2162.

Murphy, S. A. and Van der Vaart, A. W. 1996, ”Likelihood inference in the errors-in-variables model.” J. Multivariate Anal. 59, no. 1, 81-08.

Newey, W. (2001): “Flexible Simulated Moment Estimation of Nonlinear

Errors-in-Variables Models," Review of Economics and Statistics 83, 616–627.

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

Ridder, R., and R. Moffitt (2006): "The Econometrics of Data Combination," in Handbook of Econometrics, vol. 6, ed. by J. J. Heckman, and E. Leamer, Elsevier Science.

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

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

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

Taupin, M. L. (2001): "Semi-parametric Estimation in the Nonlinear Structural Errors-in-Variables Model," Annals of Statistics 29, 66–93.

Van de Geer, S. (1993), "Hellinger-Consistency of Certain Nonparametric Maximum Likelihood Estimators," The Annals of Statistics, 21, 14-44.

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

Vuong, Q. (1989): "Likelihood Ratio Test for Model Selection and Nonnested Hypotheses," Econometrica 57, 307–333.

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

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

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.

Wansbeek, T., and E. Meijer (2000): Measurement Error and Latent Variables in Econometrics, North Holland.

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

Department of Economics
Yale University
E-mail: xiaohong.chen@yale.edu

Department of Economics
Johns Hopkins University
E-mail: yhu@jhu.edu