

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

Raymond J. Carroll

Department of Statistics, Texas A&M University, E-mail: carroll@stat.tamu.edu

Xiaohong Chen

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

Yingyao Hu

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

Abstract

This paper considers identification and estimation of a general nonlinear Errors-in-Variables (EIV) model using two samples. Both samples consist of a dependent variable, some error-free covariates, and an error-prone covariate, for which the measurement error has unknown distribution and could be arbitrarily correlated with the latent true values; and neither sample contains an accurate measurement of the corresponding true variable. We assume that the regression model of interest — the conditional distribution of the dependent variable given the latent true covariate and the error-free covariates — is the same in both samples, but the distributions of the latent true covariates vary with observed error-free discrete covariates. We first show that the general latent nonlinear model is nonparametrically identified using the two samples when both could have nonclassical errors, without either instrumental variables or independence between the two samples. When the two samples are independent and the nonlinear regression model is 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, with easily estimated standard errors. A Monte Carlo simulation and two data applications are presented to show the power of the approach.

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

1 INTRODUCTION

We consider measurement error problems when (a) there are no validation data, i.e., no data where the error-prone covariate is known exactly; (b) there is no knowledge of the measurement error distribution; and (c) there is no instrumental variable. As far as we know, it is thus the first paper to allow estimation in the absence of knowledge about the measurement error distribution, of an instrumental variable and of validation data. Of course, some assumptions must be made, and ours includes assumptions that (i) there are two independent data sets with the same distribution for the response given the true covariates; (ii) the measurement error is nondifferential; (iii) there is a discrete-valued covariate that is not exogenous and that the distribution of the error-prone covariate given the discrete covariate differs in the two data sets; and (iv) certain technical conditions about the invertibility of operators hold.

The paper has two main results. First, in Section 2, we show that with our data setup, and under our assumptions, the distributions of the regression model, the measurement error model and the model for the error-prone covariate are all nonparametrically identified. Armed with identification, we then take up the question of how to do practical estimation and inference (Section 3) in the case that the distribution of the response given the true predictors is specified parametrically, but that the measurement error model and the model for the error-prone covariate are nonparametric.

Measurement error problems are frequently encountered by researchers conducting empirical studies in the social and natural sciences. A measurement error is called *classical* if it is independent of the latent true values; otherwise, it is called *nonclassical*. There have been many studies on identification and estimation of linear, nonlinear, and even nonparametric models with classical measurement errors, see, e.g., Cheng and Van Ness (1999) and Carroll, et al. (2006) for detailed reviews. However, numerous validation studies in survey data sets indicate that the errors in self-reported variables, such as earnings, are typically correlated with the true values, and hence, are nonclassical, e.g., Bound, et al. (2001). This motivates many recent studies of Errors-In-Variables (EIV) problems allowing for nonclassical measurement errors. In this paper, we provide one solution to the nonparametric identification of a general nonlinear EIV model by combining two samples, where both samples contain mismeasured covariates and neither contains an accurate measurement of the latent true variable. Our identification strategy does not require the existence of instrumental

variables or repeated measurements, both samples could have nonclassical measurement errors and the two samples could be arbitrarily correlated.

There are three broad approaches to identification of general nonlinear EIV models. The first imposes parametric restrictions on measurement error distributions, see, e.g., Fan (1991), Liang, et al. (1999) and Hong and Tamer (2003). The second 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., Hausman, et al. (1991), Buzas and Stefanski (1996), Li and Vuong (1998), Li (2002), Wang (2004), Carroll, et al. (2004), Schennach (2004), Hu (2008), Hu and Schennach (2008) and Zinde-Walsh (2007). The third approach to identifying nonlinear EIV models with nonclassical errors is to combine two samples, see, e.g., Carroll and Wand (1991), Lee and Sepanski (1995), Chen, et al. (2005). Additional references and discussions about these existing methods can be found in Carroll, et al. (2006) and Chen, et al. (2007).

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 contains an accurate measurement of the true value; such a sample might be difficult to find in some applications.

Identification 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.

Estimation and Inference Our identification result allows for fully nonparametric EIV models and also allows for two correlated samples. However, in most empirical applications, the latent models of interest are parametric nonlinear models, and the two samples are regarded as independent. Within this framework, in Section 3 we propose a sieve Quasi-Maximum Likelihood Estimation (Q-MLE) approach. 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. Easily computed standard errors are also provided.

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

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. Denote $W_j = \{j\}$ for $j = 0, 1$, then all the probability distributions $f_{X,Y|W_j}$, $f_{Y|X^*,W_j}$, $f_{X^*|W_j}$ and $f_{X|X^*}$ can be equivalently represented in terms of matrices

$L_{X,Y|W_j}$, $L_{Y|X^*,W_j}$, $L_{X^*|W_j}$ and $L_{X|X^*}$:

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

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

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

We first state an identification result for the dichotomous case.

Proposition 2.1. *Suppose that the random variables X^*, X, W, Y and X_a^*, X_a, W_a, Y_a all have supports $\{0, 1\}$, and the following conditions hold: (1) $f_{X|X^*,W,Y} = f_{X|X^*}$; (2) $f_{X_a|X_a^*,W_a,Y_a} = f_{X_a|X_a^*}$; (3) $f_{Y_a|X_a^*,W_a} = f_{Y|X^*,W}$; (4) for $j = 0, 1$, $L_{X,Y|W_j}$ and $L_{X_a,Y_a|W_j}$ are invertible, and $f_{X^*|W_j}(0)$, $f_{X_a^*|W_j}(0) \in (0, 1)$; (5) $k_{X_a^*}(0) \neq k_{X_a^*}(1)$; (6) $f_{X_a|X_a^*}(x^*|x^*) > 0.5$ for $x^* = 0, 1$. Then: $f_{X,W,Y}$ and f_{X_a,W_a,Y_a} uniquely determine $f_{Y|X^*,W}$, $f_{X|X^*}$, $f_{X_a|X_a^*}$, $f_{X^*|W_j}$ and $f_{X_a^*|W_j}$.*

Proposition 2.1 can be viewed as a special case of the general identification theorem 2.1; hence we shall discuss its conditions in the next subsection. Nevertheless, we sketch a proof of Proposition 2.1 here to illustrate our general identification strategy. Conditions (1), (2) and (3) imply that for $j = 0, 1$, and for all $x, y \in \{0, 1\}$,

$$f_{X,Y|W=j}(x, y) = \sum_{x^*=0,1} f_{X|X^*}(x|x^*) f_{Y|X^*,W=j}(y|x^*) f_{X^*|W=j}(x^*), \quad (2.1)$$

$$f_{X_a,Y_a|W_a=j}(x, y) = \sum_{x^*=0,1} f_{X_a|X_a^*}(x|x^*) f_{Y|X^*,W=j}(y|x^*) f_{X_a^*|W_a=j}(x^*). \quad (2.2)$$

Since all the variables are 0-1 dichotomous and probabilities sum to one, Equations (2.1) and (2.2) involve 12 distinct known probability values of $f_{X,Y|W=j}$ and $f_{X_a,Y_a|W_a=j}$, and 12 distinct unknown values of $f_{X|X^*}$, $f_{Y|X^*,W=j}$, $f_{X^*|W=j}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a=j}$, which makes exact identification (unique solution) of the 12 distinct unknown values possibly. However, equations (2.1) and (2.2) are nonlinear in the unknown values, we need additional restrictions (such as conditions (4), (5) and (6)) to ensure the existence of unique solution.

Using the matrix notations, equations (2.1) and (2.2) can be respectively expressed as

$$L_{X,Y|W_j} = L_{X|X^*}L_{X^*,Y|W_j} = L_{X|X^*}L_{X^*|W_j}L_{Y|X^*,W_j} \quad \text{for } j = 0, 1, \quad (2.3)$$

and

$$L_{X_a,Y_a|W_j} = L_{X_a|X_a^*}L_{X_a^*,Y_a|W_j} = L_{X_a|X_a^*}L_{X_a^*|W_j}L_{Y|X^*,W_j} \quad \text{for } j = 0, 1. \quad (2.4)$$

Condition (6) implies that $L_{X_a|X_a^*}$ is invertible, this, condition (4) and equations (2.3) - (2.4) imply that $L_{Y|X^*,W_j}$, $L_{X|X^*}$, $L_{X^*|W_j}$ and $L_{X_a^*|W_j}$ are invertible. Thus,

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

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

$$\begin{aligned} L_{X_a,X_a} &\equiv \left(L_{X_a,Y_a|W_1}L_{X,Y|W_1}^{-1} \right) \left(L_{X_a,Y_a|W_0}L_{X,Y|W_0}^{-1} \right)^{-1} \\ &= L_{X_a|X_a^*} \begin{pmatrix} k_{X_a^*}(0) & 0 \\ 0 & k_{X_a^*}(1) \end{pmatrix} L_{X_a^*|X_a^*}^{-1}. \end{aligned} \quad (2.5)$$

Equation (2.5) provides an eigenvalue-eigenvector decomposition of the observed (or known) matrix L_{X_a,X_a} . Condition (5) implies that the eigenvalues are distinct. Notice that each eigenvector is a column in $L_{X_a|X_a^*}$, which is a conditional density; hence each eigenvector is automatically normalized. Therefore, from the observed L_{X_a,X_a} , we can compute its eigenvalue-eigenvector decomposition as follows:

$$\begin{aligned} L_{X_a,X_a} &= \begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix} \begin{pmatrix} k_{X_a^*}(x_1^*) & 0 \\ 0 & k_{X_a^*}(x_2^*) \end{pmatrix} \times \\ &\quad \times \begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix}^{-1}, \end{aligned} \quad (2.6)$$

that is, the value of each entry on the right-hand side of equation (2.6) can be directly computed from the observed matrix L_{X_a,X_a} . The only ambiguity left is the value of the indices x_1^* and x_2^* , or the indexing of the eigenvalues and eigenvectors. Since for $j = 1, 2$, the values of $f_{X_a|X_a^*}(0|x_j^*)$ and $f_{X_a|X_a^*}(1|x_j^*)$ are known in equation (2.6), condition (6) pins down the index x_j^* to be: $x_j^* = 0$ if $f_{X_a|X_a^*}(0|x_j^*) > 0.5$ and $x_j^* = 1$ if $f_{X_a|X_a^*}(1|x_j^*) > 0.5$. Thus we have identified $L_{X_a|X_a^*}$ (i.e., $f_{X_a|X_a^*}$) from the decomposition of the observed matrix L_{X_a,X_a} . Next, we can identify $L_{X_a^*,Y_a|W_j}$ ($f_{X_a^*,Y_a|W_j}$) from equation (2.4) as $L_{X_a^*,Y_a|W_j} = L_{X_a^*|X_a^*}^{-1}L_{X_a,Y_a|W_j}$; hence the conditional density $f_{Y|X^*,W_j} = f_{Y_a|X_a^*,W_j}$ and the marginal density $f_{X_a^*|W_j}$ are identified. We can then identify $L_{X,X^*|W_j}$ ($f_{X,X^*|W_j}$) from equation (2.3) as $L_{X,X^*|W_j} = L_{X,Y|W_j}L_{Y|X^*,W_j}^{-1}$; hence the densities $f_{X|X^*}$ and $f_{X^*|W_j}$ are identified.

2.2 General Case

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

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

A formal treatment of identifiability results in technical conditions that are seemingly algebraically and conceptually complex. The precise conditions are given in Section A.1. Here we state the conditions in a form more suitable for understanding.

Assumption 2.1. *The regression model for the responses given the true covariates is the same in both samples. That is, the distribution of Y given (X^*, W) in the main sample is the same as that of Y_a given (X_a^*, W_a) in the auxiliary sample. In addition, W is not exogenous, i.e., the distribution of Y given (X^*, W) depends on W .*

Assumption 2.2. *Measurement error is nondifferential. That is, the distribution of X given X^* , W and Y is independent of W and Y , and similarly the distribution of X_a given X_a^* , W_a and Y_a is independent of W_a and Y_a . In such cases, X and X_a are surrogates for X^* and X_a^* , respectively.*

Assumption 2.3. *The measurement error distributions of X given X^* and X_a given X_a^* are not pathological. For example, these distributions should be complete or boundedly complete, see Section A.1 and Assumption A.4 for more details.*

Assumption 2.4. *The samples are not random samples from the same population, nor are they derived by splitting a single random sample into two. In addition, the distributions of X^* given W and X_a^* given W_a are not identical.*

Assumption 2.5. *The surrogate X_a is targeted for the true X_a^* . Specifically, either the mean, mode or a fixed quantile of the distribution of X_a given X_a^* is equal to X_a^* . This condition is not required in the primary data set.*

With these essential assumptions, we have the following result ensuring nonparametric identifiability.

Theorem 2.1. *Suppose Assumptions 2.1-2.5 hold, see Assumptions A.1–A.6 for the precise technical description. Then, nonparametrically, the observed data uniquely determine the underlying common regression model, as well as the measurement error model for X given X^* and the distribution of X^* given W . Similarly, the error model and the latent variable model are identified in the auxiliary sample. The identification theorem does not require that the two samples be independent of each other.*

Unfortunately, for understanding at least, the proof of Theorem 2.1 relies heavily on operator theory. However, to see what is going on, it is instructive to consider a simple example.

2.2.1 Example

More complex examples, both simulated and data-based, are given in Sections 4 and 5. Suppose that W and W_a are binary. Suppose further that given (X^*, W) , Y has mean $\beta_0 + \beta_1 X^*$ and variance σ_ϵ^2 , and also in the auxiliary sample. This is Assumption 2.1.

Next suppose that measurement error is nondifferential (Assumption 2.2), and that given (X^*, W) , the observed surrogate X has mean X^* and variance σ_u^2 , while in the auxiliary sample, X_a given (X_a^*, W_a) has mean X_a^* and variance σ_{ua}^2 , i.e., different measurement error variances. This satisfies Assumption 2.5 with mean "targeting" in the auxiliary data set.

Finally, suppose that X^* given W has mean μ_x independent of W , with variance σ_x^2 , but that X_a^* given W_a has mean $\alpha_0 + \alpha_1 W$ and variance σ_{xa}^2 . This satisfies Assumption 2.4 if either $\alpha_1 \neq 0$ or $\sigma_x^2 \neq \sigma_{xa}^2$.

Theorem 2.1 now asserts that if the measurement error does not have a pathological distribution, Assumption 2.3, then all the parameters listed are identified from the observed data, as are the unspecified distributions. For example, if $\alpha_1 \neq 0$, note that $E(Y_a|W_a = 1) - E(Y_a|W_a = 0) = \beta_1 \alpha_1$ and $E(X_a|W_a = 1) - E(X_a|W_a = 0) = \alpha_1$, so that β_1 is readily identified.

2.3 What Does Nonparametric Identification Tell Us?

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

Under the intuitive version of the assumptions 2.1-2.5, the point is that all aspects of the problem can be identified: regression model, measurement error model and latent variable model. Crucially, this says that whatever one's favorite paradigm, be it parametric, semiparametric, or nonparametric, be it Bayesian or frequentist, consistent estimation is possible. If one were to chose a Bayesian approach, then our result suggests that inference will not depend crucially upon the prior.

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

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

2.4 Why Does Splitting a Sample Into Two Not Work?

The somewhat informal assumptions given in Section 2.2 make it seem possible that one can achieve identifiability in a single random sample from a population by somehow cleverly splitting it into two. Assumption 2.4 actually prohibits this by a proscription, but the actual details arise in the technically precise conditions of Section A.1. In the appendix Section A.2, we describe in more detail, using the technically precise conditions of Section A.1, why splitting a single random sample into two will not lead to identifiability.

3 Sieve Quasi Likelihood Estimation

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

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 A.1–A.6. Nevertheless, in empirical studies, we typically have either a semiparametric or a parametric specification of the conditional density $f_{Y|X^*,W}$ as the model of interest. In this section, we treat the other densities $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$, and $f_{X_a^*|W_a}$ as unknown nuisance functions, but consider a parametrically specified conditional density of Y given (X^*, W) :

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

Define

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

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

Let $\alpha_0 \equiv (\theta_0^T, f_{01}, f_{01a}, f_{02}, f_{02a})^T \equiv (\theta_0^T, f_{X|X^*}, f_{X_a|X_a^*}, f_{X^*|W}, f_{X_a^*|W_a})^T$ denote the true parameter values, in which θ_0 is really “pseudo-true” when the parametric model $g(y|x^*, w; \theta)$ is incorrectly specified for the unknown true density $f_{Y|X^*,W}$. We first introduce a sieve MLE estimator $\hat{\alpha}$ for α_0 , and in later subsections establish the asymptotic normality of $\hat{\theta}$.

3.1 Sieve Likelihood Under Possible Misspecification

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

Of course, we need to impose some mild smoothness restrictions on the unknown densities. To do this, for concreteness we consider the widely used Hölder space of functions. Let $\xi = (\xi_1, \xi_2)^T \in \mathbb{R}^2$, $a = (a_1, a_2)^T$, and $\nabla^a h(\xi) \equiv \frac{\partial^{a_1+a_2} h(\xi_1, \xi_2)}{\partial \xi_1^{a_1} \partial \xi_2^{a_2}}$ denote the $(a_1 + a_2)^{\text{th}}$ derivative. Let $\|\cdot\|_E$ denote the

Euclidean norm. Let $\mathcal{V} \subseteq \mathbb{R}^2$ and $\underline{\gamma}$ be the largest integer satisfying $\gamma > \underline{\gamma}$. The Hölder space $\Lambda^\gamma(\mathcal{V})$ of order $\gamma > 0$ is a space of functions $h : \mathcal{V} \mapsto \mathbb{R}$, such that the first $\underline{\gamma}$ derivatives are continuous and bounded, and the $\underline{\gamma}^{\text{th}}$ derivative is Hölder continuous with the exponent $\gamma - \underline{\gamma} \in (0, 1]$. We define a Hölder ball as $\Lambda_c^\gamma(\mathcal{V}) \equiv \{h \in \Lambda^\gamma(\mathcal{V}) : \|h\|_{\Lambda^\gamma} \leq c < \infty\}$, in which

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

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

$$\begin{aligned} \mathcal{F}_1 &= \{f_1(\cdot|\cdot) \in \Lambda_c^{\gamma_1}(\mathcal{X} \times \mathcal{X}^*) : f_1(\cdot|x^*) \text{ is a positive density function for all } x^* \in \mathcal{X}^*\}, \\ \mathcal{F}_{1a} &= \left\{ \begin{array}{l} f_{1a}(\cdot|\cdot) \in \Lambda_c^{\gamma_{1a}}(\mathcal{X}_a \times \mathcal{X}^*) : \text{Assumption A.6 holds,} \\ f_{1a}(\cdot|x^*) \text{ is a positive density function for all } x^* \in \mathcal{X}^* \end{array} \right\}, \end{aligned}$$

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

$$\mathcal{F}_2 = \left\{ \begin{array}{l} f_2(\cdot|w), f_{2a}(\cdot|w) \in \Lambda_c^{\gamma_2}(\mathcal{X}^*) : \text{Assumption A.5 holds,} \\ f_2(\cdot|w), f_{2a}(\cdot|w) \text{ are positive density functions for all } w \in \mathcal{W} \end{array} \right\}.$$

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

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

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

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

in which

$$\begin{aligned} \ell(Z_i; \alpha) &\equiv S_i \ell_p(Z_i; \theta, f_1, f_2) + (1 - S_i) \ell_a(Z_i; f_{1a}, f_{2a}), \\ \ell_p(Z_i; \theta, f_1, f_2) &= \log \left\{ \int f_1(X_i | x^*) g(Y_i | x^*, W_i; \theta) f_2(x^* | W_i) dx^* \right\} + \log f_W(W_i), \\ \ell_a(Z_i; f_{1a}, f_{2a}) &= \log \left\{ \int f_{1a}(X_i | x_a^*) g(Y_i | x_a^*, W_i; \theta) f_{2a}(x_a^* | W_i) dx_a^* \right\} + \log \{f_{W_a}(W_i)\}. \end{aligned}$$

Let $E(\cdot)$ denote the expectation with respect to the underlying true data generating process for Z_i .

To stress that our combined data set consists of two samples, sometimes we let $Z_{pi} = (X_i, W_i, Y_i)^T$

denote the i^{th} observation in the primary data set, and $Z_{aj} = (X_{aj}, W_{aj}, Y_{aj})^T$ denote j^{th} observation in the auxiliary data set. Then

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

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

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

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

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

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

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

$$\mathcal{F}_1^n = \left\{ f_1(x|x^*) = [p_1^{k_{1,n}}(x, x^*)^T \beta_1]^2 \in \mathcal{F}_1 \right\}, \quad \mathcal{F}_{1a}^n = \left\{ f_{1a}(x_a|x_a^*) = [p_{1a}^{k_{1a,n}}(x_a, x_a^*)^T \beta_{1a}]^2 \in \mathcal{F}_{1a} \right\},$$

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

The results of our simulation study and real data examples are not sensitive to these different choices of sieves spaces. See Section 3.6 for detailed discussion of implementation.

3.2 Consistency Under a Strong Norm

In Section A.4 in the Appendix, we state conditions under which the two-sample sieve quasi- MLE $\hat{\alpha}_n = (\hat{\theta}^T, \hat{f}_1, \hat{f}_{1a}, \hat{f}_2, \hat{f}_{2a})^T$ is consistent for $\alpha_0 \equiv (\theta_0^T, f_{01}, f_{01a}, f_{02}, f_{02a})^T \equiv (\theta_0^T, f_{X|X^*}, f_{X_a|X_a^*}, f_{X^*|W}, f_{X_a^*|W_a})^T$ under a strong norm (defined in the Appendix).

3.2.1 Estimation of the Nonparametric Components

Our identification result (Theorem 2.1) says that not only is the latent regression model $(g(y|x^*, w; \theta_0))$ identified, but so too are the measurement error densities $(f_{X|X^*}, f_{X_a|X_a^*})$ and the densities of the latent variables (X^*, X_a^*) given (W, W_a) (or $f_{X^*|W}, f_{X_a^*|W_a}$ for two independent samples). Because these latter densities are modeled nonparametrically, their estimation falls into the realm of deconvolution, for which rates of convergence under the strong norm are often miserable, e.g., logarithmic when the measurement error is normally distributed.

In one of the examples, Section 5.1, we show the sieve estimated densities. While rough, they do give a sense of the form of the measurement error and latent variable density functions. This is about all one can expect given the slow rates of convergence of the estimates.

3.2.2 Rates of Convergence Under a Weaker Norm

The two-sample sieve quasi-ML estimation of unknown densities $f_{X|X^*}, f_{X_a|X_a^*}, f_{X^*|W}, f_{X_a^*|W_a}$ is ill-posed; hence their convergence rates under the strong norm are very slow. However, In Section A.5 in the Appendix, we derive its faster than $(n+n_a)^{-1/4}$ rate of convergence under a weaker norm (which is the Fisher-norm when the latent parametric regression model is correctly specified). We show that the consistency under the strong norm and faster rate of convergence under the weaker norm is what we need for root- $(n+n_a)$ consistency of $\hat{\theta}_n$ for θ_0 (the parametric part). The most important sufficient condition for the faster rate under the weaker norm is that each of the sieve number of terms k_{1n}, k_{1an}, k_{2n} must grow neither too fast nor too slowly, so that $\frac{\max(k_{1n}, k_{1an}, k_{2n})}{n+n_a} = o([n+n_a]^{-1/2})$ for the variance part and $\max(k_{1n}^{-\gamma_1/2}, k_{1an}^{-\gamma_{1a}/2}, k_{2n}^{-\gamma_2}) = o([n+n_a]^{-1/4})$ for the bias part.

3.3 Asymptotic Normality Under Possible Misspecification

Our main asymptotic result is to show that the two-sample sieve quasi ML estimator $\hat{\theta}_n$ of θ_0 is asymptotically normally distributed around θ_0 . We actually derive the asymptotic distribution of $\hat{\theta}_n$ regardless of whether the latent parametric model $g(y|x^*, w; \theta_0)$ is correctly specified or not. The technical details and all assumptions are given in the Appendix Sections A.6 and A.7, with the result summarized here.

Theorem 3.1. *Under Assumptions A.7–A.19 in the Appendix, it follows that $\sqrt{n+n_a}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V_*^{-1}I_*V_*^{-1})$ for matrices (V_*, I_*) defined in the Appendix.*

Remark 3.1. *The matrices (V_*, I_*) are defined in terms of expectations of various pathwise derivatives of the sieve loglikelihood, and hence can be estimated in the usual manner of replacing the expectation by averages across the data and replacing unknown quantities by their estimates. See Ai and Chen (2007) for the formal justification of consistent estimation of asymptotic covariance under possible model misspecification.*

3.4 Semiparametric Efficiency Under Correct Specification

When $g(y|x^*, w; \theta_0)$ correctly specifies the true unknown conditional density $f_{Y|X^*, W}(y|x^*, w)$, then $I_* = V_*$ becomes the semiparametric information bound for θ_0 , and our above estimator $\hat{\theta}_n$ becomes semiparametrically efficient for θ_0 .

Specifically, by combining our Theorem 3.1 and Theorem 4 of Shen (1997), we immediately obtain the following:

Theorem 3.2. *Suppose that $g(y|x^*, w; \theta_0) = f_{Y|X^*, W}(y|x^*, w)$, that I_* is positive definite, and that Assumptions A.7– A.18 hold. Then the sieve MLE $\hat{\theta}_n$ is semiparametrically efficient, and $\sqrt{n+n_a}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_*^{-1})$.*

3.5 Estimation of Standard Error and Confidence Region

There are two ways of estimating the asymptotic covariance matrix of the two-sample sieve MLE $\hat{\theta}_n$. The first is to use the definition of I_*^{-1} (stated in the Appendix), but to replace expectations by sample averages and to replace unknown parameters by their sieve estimates. This is what we did in our examples. As it turns out, this is asymptotically equivalent to using the sieve MLE approximation as if it were a parametric model and estimating I_*^{-1} as the appropriate submatrix of the inverse of the estimated Fisher information.

The above consistent standard error estimation automatically provides an asymptotically valid confidence region. Alternatively, applying theorem B of Chen, Linton and van Keilegom (2003), we know that the standard bootstrap also provide consistent estimate of confidence region. We implemented both in a real data example.

3.6 Computation

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

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

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

More difficult is the targeting Assumption 2.5, see also Assumption A.6, that is applied to (X_a, X_a^*) . Here we use the same form for the density as for that of (X, X^*) , and we consider mean targeting, i.e., $E(X_a|X_a^*) = X_a^*$. Let the Hermite orthogonal series be defined as $H_0(x) = 1$, $H_1(x) = 2x$ and $H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$, and define $p_n(x) = H_n(x) \exp(-x^2/2)(2^n n! \pi^{1/2})^{-1/2}$. Then $\{p_n(\cdot)\}$ is an orthogonal series of the type required, with the property that $\int xp_n(s)p_m(x)dx = \{(n+1)/2\}^{1/2}I(m = n+1) + (n/2)^{1/2}I(m = n-1)$. Let $Q = \{q_1(x_a^*), \dots, q_{K_n}(x_a^*)\}^\top$, $P = \{p_1(x_a - x_a^*), \dots, p_{J_n}(x_a - x_a^*)\}^\top$, and let $B = (\gamma_{jk})$. Then $f(x_a, x_a^*) = (P^\top BQ)^2 = Q^\top B^\top P P^\top BQ$. We require for mean targeting that, for every x_a^* , $0 = \int (x_a - x_a^*)f(x_a - x_a^*)dx_a$, which means $0 = B^\top \int (x_a - x_a^*)P P^\top dx_a B$. However, the latter is $B^\top S B$, where S has all zeros except that its $(k, k+1)$ and $(k+1, k)$ components equal $(k+1)^{1/2}$ for $k = 1, \dots, K_n - 1$. The restriction $0 = B^\top S B$ is readily achieved, e.g., for $J_n = 5$, $K_n = 4$, set $B_* = \text{diag}(1, 0, 1, 0, 1)B$, then $B_*^\top S B_* = 0$ by algebra.

In applications, the sieve MLE method needs to choose the order of the sieve terms. Our experience is that the estimation of the finite dimensional parameters θ are not very sensitive to the order of sieves. Of course if one cares about estimation of nonparametric density functions, then one could apply either the AIC or the generalized cross-validation, or the more recent covariance penalty methods suggested in Efron (2004) and Shen and Huang (2006), among others.

4 Simulation and Comparisons

4.1 The Simulation Study

Ours is the first paper to show nonparametric identifiability in the context of two samples, and the first to derive an estimator of the parametric part with no assumptions made about the distribution of the measurement error or the latent variable. In this section, we are going to compare 5 estimators, as follows.

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

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

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

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

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

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

true value of θ :	$\theta_1 = 1$	$\theta_2 = 1$	$\theta_3 = 1$
ignoring measurement error:			
– mean estimate	0.175	0.307	0.595
– standard error	0.084	0.123	0.188
– root MSE	0.829	0.703	0.446
2-sample sieve MLE:			
– mean estimate	1.026	1.026	1.005
– standard error	0.182	0.158	0.192
– root MSE	0.184	0.160	0.192
Correctly specified parametric model:			
– mean estimate	0.919	1.060	0.973
– standard error	0.053	0.066	0.084
– root MSE	0.096	0.089	0.088
Parametric model with misspecified measurement error distribution:			
– mean estimate	1.435	1.382	1.426
– standard error	0.143	0.121	0.262
– root MSE	0.458	0.401	0.500
Parametric model with measurement error and latent variable models misspecified:			
– mean estimate	1.124	1.478	1.540
– standard error	0.127	0.156	0.292
– root MSE	0.177	0.503	0.614

the unknown true conditional density $f_{X_a^*|W_a}$ according to

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

We let the mismeasured value X_a be $X_a = X_a^* + 0.5e^{-0.1X_a^*}\nu$ with $\nu \sim N(0, 1)$, which implies that x_a^* is the mode of the conditional density $f_{X_a|X_a^*}(\cdot|x_a^*)$. The simulation was repeated 400 times.

We used the simple sieve expression $[p_1^{k_1,n}(x_1, x_2)^T \beta_1]^2 = [\sum_{j=0}^{J_n} \sum_{k=0}^{K_n} \gamma_{jk} p_j(x_1 - x_2) q_k(x_2)]^2$ to approximate $f_{X|X^*}(x_1|x_2)$ and $f_{X_a|X_a^*}(x_1|x_2)$, with $k_{1,n} = (J_n + 1)(K_n + 1)$, $J_n = 5$, $K_n = 3$. We also use $[p_2^{k_2,n}(x^*)^T \beta_2(w)]^2 = [\sum_{k=1}^{k_2,n} \gamma_k(w) q_k(x^*)]^2$ to approximate $f_{X^*|W_j=w}$ and $f_{X_a^*|W_j=w}$ with $W_j = -1, 0, 1$ and $k_{2,n} = 4$. The sieve bases $\{p_j(\cdot)\}$ and $\{q_k(\cdot)\}$ are Hermite polynomials bases, and we also impose the integration to one and the mean targeting constraints as described in subsection 3.6.

For the parametric model with incorrectly specified measurement error distribution, we computed the parametric MLE when it was assumed that the measurement errors in the primary and auxiliary samples were homoscedastic with standard deviations 0.6079 and 0.6202, respec-

tively. For the parametric model with measurement error and latent variable models misspecified, we did the following. Define $(\gamma, \gamma_{1a}, \gamma_{2a}, \gamma_{3a}, \gamma_{4a}) = (1.0, 1.0, 2.0, 1.05, 0.25)$, and define $\varphi(x) = \exp\{x - \exp(x)\}$. Then set $f_{X^*|W}(X^*|W; \gamma) = \gamma\varphi\{\gamma X^*\}$, and for $W_a = (-1, 0, 1)$, set $f_{X_a^*|W_a}(X_a^*|W_a; \gamma_a) = \gamma_{1a}\varphi(\gamma_{1a}X_a^*)$, $\gamma_{2a}\varphi(\gamma_{2a}X_a^*)$ and $\gamma_{3a}\varphi(\gamma_{3a}X_a^* - \gamma_{4a})$, respectively.

The simulation results shown in Table 1, which shows what one might expect. First, accounting for measurement error matters: the 2-sample sieve MLE has a much smaller bias and MSE than the estimator ignoring measurement error. Second, there are substantial costs for being nonparametric: compared to a correctly specified parametric model, 2-sample sieve MLE is simply more variable, a not very surprising result. Third, there are costs for model misspecification of either the measurement error distribution or the distribution for (X^*, X_a^*) .

4.2 Testing Assumption 2.1 and Assumption A.3

Huang, et al. (2006) develop a method that can allow the testing of assumptions about latent variable distributions in measurement error models. Here we show that a modification of their basic idea is capable of detecting violations of Assumption 2.1 or its more technical version Assumption A.3.

We performed 500 simulations of the following experiment. For the main data set we had $n = 1000$, $W = (W_1, W_2)$, where $W_1 = \text{Bernoulli}(0.5)$ and $W_2 = \text{Bernoulli}(0.3)$ are independent of one another, $X^* = W_1 + W_2 + N(0, 0.25)$, $X = X^* + N(0, 0.25)$ and finally $Y = 1.5X^* + 0.3W_1 + 0.7W_2 + N(0, 0.01)$. For the auxiliary data set, we had $n_a = 1000$, $W_a = (W_{1a}, W_{2a})$, where $W_{1a} = \text{Bernoulli}(0.3)$ and $W_{2a} = \text{Bernoulli}(0.7)$ are independent of one another, $X^* = 1.5W_1 + 0.5W_2 + N(0, 0.25)$, $X = X^* + N(0, 0.25)$ and finally $Y = 0.5X^* + 1.3W_1 + 1.7W_2 + N(0, 0.01)$. Note how Assumption 2.1 and its more technical version Assumption A.3 are badly violated in this case because the regression models are very different.

The fitting method was as follows. We assumed that Assumption 2.1 holds and the homoscedastic model that has $E(Y|X^*, W) = \beta_0 + \beta_1 X^* + \beta_2 W_1 + \beta_3 W_2$. In addition, we assumed that the distribution of X^* given W had different means and variances depending on the four levels of W . The same thing but with different parameters was assumed for the distribution of X_a^* given W_a . We also assumed that the measurement error in X and X_a was additive and homoscedastic but with possibly different variances. Normal-theory maximum likelihood, which is also method of moments

in this context, was used to fit the simulated data sets.

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

The results were as follows. For β_1 , the mean difference between the original and perturbed estimates was -0.16 with a standard deviation 0.02, and thus an effect size of -8.38 . For β_2 , the mean difference between the original and perturbed estimates was 0.17 with a standard deviation 0.03, and thus an effect size of 5.04. For β_3 , the mean difference between the original and perturbed estimates was 0.14 with a standard deviation 0.03, and thus an effect size of 4.75. Clearly, this calculation shows that the Huang, et al (2006) method can detect serious departures from Assumption 2.1 and Assumption A.3.

5 Illustrative Examples

5.1 Nutritional Epidemiology

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

With EATS as the primary study and OPEN as the auxiliary study, the assumption of non-differential measurement error in the 24hr recalls (Assumption 2.2) is standard in this context. While our 2-sample sieve Q-MLE does not make this assumption, it makes sense to believe that the measurement error distributions are the same in the two studies. Both studies took place in the

Table 2: Estimates and Bootstrap analysis of the OPEN and EATS data sets.

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

United State, and thus the stability Assumption 2.5 also seems reasonable. The main difference between EATS and OPEN is that the former was a national study, while the latter took place in the relatively affluent Montgomery County Maryland. Thus, one would expect the distributions of X^* given W and X_a^* given W_a to be different, and of course one would expect that the distribution of true transformed beta-carotene intake will depend on gender and age. Thus, assumption 2.4 seems reasonable in this context. Indeed, for those aged under 50, Wilcoxon rank tests comparing the two transformed 24-hour recalls between the two data sets are statistically significant both for men and for women. Within OPEN, the Wilcoxon rank test is also statistically significant when comparing genders or when comparing age categories, while no such differences are observed for EATS. However, in EATS the 24-hour recalls for women had statistically significantly more variability than those for men, using a Wilcoxon test on the absolute differences from the means.

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

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

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

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

- The naive OLS estimator with measurement errors ignored.

- A parametric maximum likelihood estimator under the additional Assumptions: $\epsilon_{ij} = N(0, \sigma_\epsilon^2)$, $U_{ij} = N(0, \sigma_u^2)$, $X_{i1}^* = a_0 + a_1W_{ij1} + a_2W_{ij2} + \nu_{i1}$, and $X_{i2}^* = b_0 + b_1W_{ij1} + b_2W_{ij2} + \nu_{i2}$, with $\nu_{ij} = N(0, \sigma_{\nu,j}^2)$. Note that for this parametric MLE, the measurement error status is assumed to not depend on j .
- The sieve MLE under the additional restriction that the latent model of interest is (4.1) with $\epsilon_{ij} = \text{Normal}(0, \sigma_\epsilon^2)$.
- The sieve MLE with no assumptions about the distribution of ϵ_{ij} .

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

We also implemented 500 bootstraps by resampling (Y, X, W) within each population. The results are given in Table 2. We see that the measurement errors cause significant attenuation in the estimation of θ_1 . The corrected estimators have much greater variability than the naive estimator, with variability increasing as assumptions are relaxed.

In Figure 1, we plot the sieve estimated density functions for the measurement error models in EATS and OPEN, as well as the density functions for the latent covariates. The measurement error density estimates are indeed rough, as expected from the rates of convergence described in Section 3.2, but they appear somewhat vaguely centered at the true value of the latent variable and clearly depend upon it, the latter being the point of most interest. The latent variable density estimates are easier to visualize because W and W_a have only 4 levels: there is more skewness in the EATS data than in the OPEN data.

5.1.1 Testing Assumption 2.1 and Assumption A.3

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

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

The results were as follows. For β_1 , the mean difference between the original and perturbed estimates was 0.035 with a standard deviation 0.035. For β_2 , the mean difference between the

original and perturbed estimates was 0.01 with a standard deviation 0.014. For β_3 , the mean difference between the original and perturbed estimates was 0.01 with a standard deviation 0.015. It appears then that while Assumption A.3 may be violated in this example, the size of that violation is not likely to be large.

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

5.2 Voting Behavior with a Binary Response

Although our theory assumes a continuous response, the methodology also applies to binary responses. Here we apply the 2-sample sieve MLE to estimate the effect of earnings on voting behavior. The sieve basis functions and the number of sieve terms are chosen in the same way as those in the simulation study. The population we consider consists of all the individuals with jobs who were eligible to vote in the presidential election on Tuesday, November 2, 2004. The dependent variable is a dichotomous variable which equals 1 if an individual voted, equals 0 otherwise. We use the probit model to estimate the effect of earnings with covariates such as years of schooling, age, gender, and marital status. We use a random sample from the Current Population Survey (CPS) in November 2004. The major concern regarding this sample is that the self-reported earnings may have nonclassical measurement errors. In order to consistently estimate the model using our new estimator, we use an auxiliary random sample from the Survey of Income and Program Participation (SIPP). The questionnaire of SIPP has more income-related questions than that of CPS. In the probit model, we use log earnings rather than the original ones. We consider four subpopulations: single males, married males, single females, and married females.

Suppose the true parametric model for $f_{V|X^*,U,W}(v|x^*, u, w)$ is a probit model:

$$g(v|x^*, u, w; \theta) = [\Phi(\beta_1 x^* + u^T \beta_2 + w^T \beta_3)]^v [1 - \Phi(\beta_1 x^* + u^T \beta_2 + w^T \beta_3)]^{1-v}$$

in which V stands for voting behavior, X^* denotes the latent true log earning, U contains education and age variables, and W includes gender and marital status. Let Y denote the predicted log

earning using U and W , hence, a measurable function of (U, W) . Then, Theorem 2.1 implies that the distribution $f_{X|X^*}$ may be identified from the joint distribution of (Y, X, W) in the two samples under the identification assumptions. Moreover, the identification of $f_{V, X^*, U, W}$ follows from that of $f_{X|X^*}$ under the assumption X is independent of (V, U, W) conditional on X^* .

We consider four subpopulations: single males, married males, single females, and married females. The CPS sample contains 6689 individuals who have jobs and are eligible to vote. In this sample, 54% of the individuals are married and 56% are male. The average education level in each subsample is a little higher than the high school level. The average age of married males is about 2 years higher than that of married females, while the average age of single males is 2 years lower than that of single females. The CPS sample also shows that married people are more likely to vote than unmarried ones, and females are more likely to vote than males. In both samples, married individuals have a higher average earning than unmarried individuals and males have a higher average earning than females. In the SIPP sample, there are 11683 individuals, 30.4% of whom are married males, 18.1% of whom are single males, and 23.4% of whom are married females. The average ages of single males or females are about the same as those in the CPS sample. The married males or females in the SIPP sample are younger on average than those in the CPS sample. The average earnings are higher in the SIPP sample than in the CPS sample, except in the subsample of single males. The average education levels are similar in the four subsamples of the SIPP sample and in the CPS sample.

We consider two estimators. The first one ignores the measurement error and is the standard probit estimator using the CPS sample only; the results in Table 3 show that every variable has a significant impact on voting behavior if the CPS data were accurate. The second estimator is our proposed 2-sample sieve MLE using the two samples from CPS and SIPP. The results are given Table 3: note that none of the signs of the parameter estimates are changed. The correction for measurement error in log earnings is what one might get from additive normally distributed measurement error that was $1/3$ the variability of true log earnings. This suggests a modest effect due to measurement error.

Table 3: Estimates and analysis of voting behavior example.

voted	MLE ignoring m. error		2-sample sieve MLE	
	mean	std.dev	mean	std.dev
log weekly earning	0.063	0.0264	0.087	0.0294
years of schooling	0.164	0.0078	0.151	0.0486
age	0.020	0.0015	0.011	0.0149
male	-0.175	0.0379	-0.229	0.1297
married	0.256	0.0366	0.343	0.1035
constant	0.724	0.0315	0.793	0.0845

6 Summary

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

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

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

Acknowledgment

The authors would like to thank P. Cross, S. Donald, E. Mammen, M. Stinchcombe, and conference participants at the 2006 North American Summer Meeting of the Econometric Society and the 2006 Southern Economic Association annual meeting for valuable suggestions. We are also grateful to

Arthur Schatzkin, Amy Subar and Victor Kipnis for making the data in our example available to us. Chen acknowledges support from the National Science Foundation (SES-0631613). Carroll's research was supported by grants from the National Cancer Institute (CA57030, CA104620). Part of Carroll's work was supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

References

- [1] Ai, C., and X. Chen (2007): "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables," *Journal of Econometrics* 141, 5-43.
- [2] Bissantz, N., T. Hohage, A. Munk and F. Ruymgaart (2007): "Convergence Rates of General Regularization Methods for Statistical Inverse Problems and Applications," *SIAM Journal on Numerical Analysis*, forthcoming.
- [3] 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.
- [4] Buzas, J., and L. Stefanski (1996): "Instrumental Variable Estimation in Generalized Linear Measurement Error Models," *Journal of the American Statistical Association* 91, 999–1006.
- [5] 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.
- [6] Carroll, R. J., D. Ruppert, L. A. Stefanski and C. Crainiceanu, 2006, *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, CRI.
- [7] 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.
- [8] Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," in *Handbook of Econometrics*, vol. 6, ed. by J. J. Heckman and E. Leamer, Elsevier Science.
- [9] Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies* 72, 343–366.
- [10] Chen, X., Hong, H. and Nekipelov, D., 2007, "Nonlinear Models of measurement errors." survey article prepared for *Journal of Economic Literature*.
- [11] Chen, X., O. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models When the Criterion Function is not Smooth," *Econometrica* 71, 1591-1608.
- [12] Cheng, C. L., Van Ness, J. W., 1999, *Statistical Regression with Measurement Error*, Arnold, London.
- [13] Dunford, N., and J. T. Schwartz (1971): *Linear Operators, Part 3: Spectral Operators*. New York: John Wiley & Sons.

- [14] Efron, B. (2004): “The Estimation of Prediction Error: Covariance Penalties and Cross-Validation,” *Journal of the American Statistical Association* 99, 619-642.
- [15] Fan, J. (1991): “On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems,” *Annals of Statistics* 19, 1257–1272.
- [16] Hausman, J., H. Ichimura, W. Newey, and J. Powell (1991): “Identification and Estimation of Polynomial Errors-in-variables Models,” *Journal of Econometrics* 50, 273–295.
- [17] Hong, H., and E. Tamer (2003): “A Simple Estimator for Nonlinear Error in Variable Models,” *Journal of Econometrics* 117(1), 1–19.
- [18] Hu, Y. (2008): “Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables,” *Journal of Econometrics* 144, 27–61.
- [19] Hu, Y., and S. M. Schennach (2008): “Instrumental Variable Treatment of Nonclassical Measurement Error Models,” *Econometrica*, 76, 195-216.
- [20] Huang, X., Stefanski, L. A. and Davidian, M. (2006): “Latent-model robustness in structural measurement error models,” *Biometrika*, 93, 53-64.
- [21] Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. Bingham, S., Schoeller, D. A., Schatzkin, A. and Carroll, R. J. (2003). “The structure of dietary measurement error: results of the OPEN biomarker study.” *American Journal of Epidemiology*, 158, 14-21.
- [22] Lee, L.-F., and J. H. Sepanski (1995): “Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data,” *Journal of the American Statistical Association* 90 (429).
- [23] Li, T., and Q. Vuong (1998): “Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators,” *Journal of Multivariate Analysis* 65, 139–165.
- [24] Li, T. (2002): “Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models,” *Journal of Econometrics* 110, 1–26.
- [25] Liang, H., W. Hardle, and R. Carroll, 1999, “Estimation in a Semiparametric Partially Linear Errors-in-Variables Model,” *The Annals of Statistics*, 27, No. 5, 1519-1535.
- [26] Mattner, L. (1993): “Some Incomplete but Bounded Complete Location Families,” *Annals of Statistics*, 21, 2158-2162.
- [27] Schennach, S. (2004): “Estimation of Nonlinear Models with Measurement Error,” *Econometrica* 72, 33–76.
- [28] Shen, X. (1997): “On Methods of Sieves and Penalization,” *Annals of Statistics* 25, 2555–2591.
- [29] Shen, X. and H. Huang (2006): “Optimal Model Assessment, Selection, and Combination,” *Journal of the American Statistical Association* 101, 554-568.
- [30] Shen, X., and W. Wong (1994) “Convergence Rate of Sieve Estimates,” *The Annals of Statistics* 22, 580–615.

- [31] Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A. and Rosenfeld, S. (2001) “Comparative validation of the Block, Willett and National Cancer Institute food frequency questionnaires: The Eating at America’s Table Study,” *American Journal of Epidemiology*, 154, 1089-1099.
- [32] Wang, L., 2004, ”Estimation of nonlinear models with Berkson measurement errors,” *Annals of Statistics* 32, 2559–2579.
- [33] Zinde-Walsh, V., 2007, Errors-in-variables models: a generalized functions approach. Working paper, McGill University and CIREQ.

Appendix: Mathematical Proofs

A.1 Detailed Technical Conditions for Identifiability

In this section, we state the detailed technical assumptions that are described intuitively in Section 2. Suppose the supports of X, W, Y , and X^* are $\mathcal{X} \subseteq \mathbb{R}$, $\mathcal{W} = \{w_1, w_2, \dots, w_J\}$, $\mathcal{Y} \subseteq \mathbb{R}$, and $\mathcal{X}^* \subseteq \mathbb{R}$, respectively. We assume

Assumption A.1. (i) $f_{Y, X, X^*, W}(y, x, x^*, w)$ is positive, bounded on its support $\mathcal{Y} \times \mathcal{X} \times \mathcal{X}^* \times \mathcal{W}$, and is continuous in $(y, x, x^*) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{X}^*$; (ii) $f_{X|X^*, W, Y}(x|x^*, w, y) = f_{X|X^*}(x|x^*)$ on $\mathcal{X} \times \mathcal{X}^* \times \mathcal{W} \times \mathcal{Y}$.

Assumption A.2. (i) $f_{Y_a, X_a, X_a^*, W_a}(y, x, x^*, w)$ is positive, bounded on its support $\mathcal{Y} \times \mathcal{X}_a \times \mathcal{X}^* \times \mathcal{W}$, and is continuous in $(y, x, x^*) \in \mathcal{Y} \times \mathcal{X}_a \times \mathcal{X}^*$; (ii) $f_{X_a|X_a^*, W_a, Y_a}(x|x^*, w, y) = f_{X_a|X_a^*}(x|x^*)$ on $\mathcal{X}_a \times \mathcal{X}^* \times \mathcal{W} \times \mathcal{Y}$.

Assumption A.3. (i) $f_{Y_a|X_a^*, W_a}(y|x^*, w) = f_{Y|X^*, W}(y|x^*, w)$ on $\mathcal{Y} \times \mathcal{X}^* \times \mathcal{W}$; (ii) $f_{Y|X^*, W=w}$ changes with w .

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

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

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

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

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

Assumptions A.1, A.2 and A.3 imply that $L_{X, Y|W_j} = L_{X|X^*} L_{X^*|W_j} L_{Y|X^*, W_j}$ and $L_{X_a, Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{Y|X^*, W_j}$, where the operators $L_{X, Y|W_j}$ and $L_{X_a, Y_a|W_j}$ are observed given the data. We assume

Assumption A.4. (i) $L_{X_a|X_a^*}$ is injective, i.e., the set $\{h \in \mathcal{L}^p(\mathcal{X}^*) : L_{X_a|X_a^*} h = 0\} = \{0\}$; (ii) $L_{X, Y|W_j}$ has a right-inverse (denoted as $A = (L_{X, Y|W_j})^{-1}$), i.e., $L_{X, Y|W_j} A = I$. (iii) $L_{X_a, Y_a|W_j}$ has a right-inverse.

Assumption A.4(i), the precise version of the more intuitive Assumption 2.3, is commonly imposed in general deconvolution problems; see, e.g., Bissantz, et al. (2007). Assumption A.4(i) is the same as the *completeness* of the conditional density $f_{X_a^*|X_a}$, which is satisfied, for example, when $f_{X_a^*|X_a}$ belongs to an exponential family. 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 A.4(i) is the *bounded completeness* of the conditional density $f_{X_a^*|X_a}$; see, e.g., Mattner (1993).

Assumption A.4 implies that $L_{Y|X^*, W_j}$ and $L_{X|X^*}$ are invertible. In Section A.3 we establish the diagonalization of an observed operator $L_{X_a, X_a}^{ij} : L_{X_a, X_a}^{ij} = L_{X_a|X_a^*} L_{X_a^*}^{ij} L_{X_a|X_a^*}^{-1}$, where the operator $L_{X_a^*}^{ij} \equiv (L_{X_a^*|W_j} L_{X^*|W_j}^{-1} L_{X^*|W_i} L_{X_a^*|W_i}^{-1})$ is a diagonal operator defined as: $(L_{X_a^*}^{ij} h)(x^*) = k_{X_a^*}^{ij}(x^*) h(x^*)$ with

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

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

Assumption A.5. For any $x_1^* \neq x_2^*$, there exist $i, j \in \{1, 2, \dots, J\}$, such that $k_{X_a^*}^{ij}(x_1^*) \neq k_{X_a^*}^{ij}(x_2^*)$ and $\sup_{x^* \in \mathcal{X}^*} k_{X_a^*}^{ij}(x^*) < \infty$.

Assumption A.5 is the precise version of the more intuitive Assumption 2.4. Notice that the subsets $W_1, W_2, \dots, W_J \subset \mathcal{W}$ do not need to be collectively exhaustive. We may only consider those subsets in \mathcal{W} in which these assumptions are satisfied. Since the indices i, j are exchangeable, the condition $\sup_{x^* \in \mathcal{X}^*} k_{X_a^*}^{ij}(x^*) < \infty$ may be replaced by $\inf_{x^* \in \mathcal{X}^*} k_{X_a^*}^{ij}(x^*) > 0$.

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 $f_{X_a|X_a^*}(\cdot|x^*)$. This leads to the intuitive Assumption 2.5. 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. Normalization is needed to accomplish this task.

Assumption A.6. *One of the followings holds for all $x^* \in \mathcal{X}^*$: (i) (mean) $\int x f_{X_a|X_a^*}(x|x^*) dx = x^*$; or (ii) (mode) $\arg \max_x f_{X_a|X_a^*}(x|x^*) = x^*$; or (iii) (quantile) there is an $\gamma \in (0, 1)$ such that $\inf\{z : \int_{-\infty}^z f_{X_a|X_a^*}(x|x^*) dx \geq \gamma\} = x^*$.*

A.2 Why Does Splitting a Sample Into Two Not Work?

Here we describe why our conditions mean that one cannot achieve identifiability in a single random sample from a population by somehow cleverly splitting it into two.

First suppose that we split the sample into two based on whether $W < c$ or $W > c$. This fails for two reasons. First, in Assumptions A.1 and A.2, we assume that W and W_a have a combined support W and that the densities of (Y, X, X^*, W) and (Y_a, X_a, X_a^*, W_a) are positive on that support. In addition, to achieve Assumption A.3, splitting the sample in this way would require W to be exogenous, something excluded in Assumption A.1.

Another way to split the sample into two is to obtain different distributions for W and W_a by splitting the single random sample into two based on probability weighting. This will now violate Assumption A.5, as follows. All densities used are conditional on W and W_a , so that a change in the marginal distribution of W does not change the mapping between $k_{X_a^*}^{ij}$ and x^* , which is the key to distinguish eigenvalues. In other words, if Assumption A.5 is violated, one can not make it hold only by changing the distribution of W .

Of course, one cannot split a single random sample into two by probability weighting based on Y , or based on another variable, say S , that is related to Y but not part of (Y, X, W) , since this would mean violation of Assumption A.3.

Finally, one cannot achieve identifiability by randomly splitting a single random sample into two because of Assumption A.5.

A.3 Proof of Theorem 2.1

Under Assumption A.1,

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

For each value w_j of W , assumptions A.1-A.3 imply that

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

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

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

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

This means we have the operator equivalence

$$L_{X,Y|W_j} = L_{X|X^*} L_{X^*|W_j} L_{Y|X^*,W_j} \quad (\text{A.4})$$

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

$$L_{X_a, Y_a|W_j} = L_{X_a|X_a^*} L_{X_a^*|W_j} L_{Y|X^*,W_j}. \quad (\text{A.5})$$

Note that the left-hand sides of equations (A.4) and (A.5) are observed. Assumption A.4 implies that all the operators involved in equations (A.4) and (A.5) are invertible. Hence

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

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

$$L_{X_a, X_a}^{ij} \equiv (L_{X_a, Y_a|W_j} L_{X,Y|W_j}^{-1}) (L_{X_a, Y_a|W_i} L_{X,Y|W_i}^{-1})^{-1} = L_{X_a|X_a^*} L_{X_a^*}^{ij} L_{X_a^*|X_a^*}^{-1}. \quad (\text{A.7})$$

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

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

Equation (A.7) implies a diagonalization of an observed operator L_{X_a, X_a}^{ij} . An eigenvalue of L_{X_a, X_a}^{ij} equals $k_{X_a^*}^{ij}(x^*)$ for a value of x^* , which corresponds to an eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$.

We now show the identification of $f_{X_a|X_a^*}$ and $k_{X_a^*}^{ij}(x^*)$. First, we require the operator L_{X_a, X_a}^{ij} to be bounded so that the diagonal decomposition may be unique; see, e.g., Dunford and Schwartz (1971, theorem XV.4.3.5, p. 1939). Equation (A.7) implies that the operator L_{X_a, X_a}^{ij} has the same spectrum as the diagonal operator $L_{X_a^*}^{ij}$. Since an operator is bounded by the largest element of its spectrum, Assumption A.5 guarantees that the operator L_{X_a, X_a}^{ij} is bounded. Second, although it implies a diagonalization of the operator L_{X_a, X_a}^{ij} , equation (A.7) does not guarantee distinctive eigenvalues. However, such ambiguity can be eliminated by noting that the observed operators L_{X_a, X_a}^{ij} for all i, j share the same eigenfunctions $f_{X_a|X_a^*}(\cdot|x^*)$. Assumption A.5 guarantees that, for any two different eigenfunctions $f_{X_a|X_a^*}(\cdot|x_1^*)$ and $f_{X_a|X_a^*}(\cdot|x_2^*)$, one can always find two subsets W_j and W_i such that the two different eigenfunctions correspond to two different eigenvalues $k_{X_a^*}^{ij}(x_1^*)$ and $k_{X_a^*}^{ij}(x_2^*)$ and, therefore, are identified.

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

Fourth, in order to fully identify each eigenfunction, i.e., $f_{X_a|X_a^*}$, we need to identify the exact value of x^* in each eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$. However, note that assumption A.6 identifies the exact value of x^* for each eigenfunction $f_{X_a|X_a^*}(\cdot|x^*)$.

After fully identifying the density function $f_{X_a|X_a^*}$, we now show that the density of interest $f_{Y|X^*, W}$ and $f_{X|X^*}$ are also identified. By equation (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 A.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

$$L_{X|X^*}L_{X^*|W_j} = L_{X,Y|W_j}L_{Y|X^*,W_j}^{-1}. \quad (\text{A.8})$$

The left-hand side of equation (A.8) equals an operator with the kernel function $f_{X,X^*|W=w_j} \equiv f_{X|X^*}f_{X^*|W=w_j}$. Since the right-hand side of equation (A.8) has been identified, the kernel $f_{X,X^*|W=w_j}$ on the left-hand side is also identified. We may then identify $f_{X|X^*}$ through

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

A.4 Consistency Under a Strong Norm

Here we describe the precise conditions that guarantee consistency of the sieve estimator. We make the following assumptions.

Assumption A.7. (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 $w \in \mathcal{W}$.

Define a norm on \mathcal{A} as: $\|\alpha\|_s = \|\theta\|_E + \|f_1\|_{\infty, \omega_1} + \|f_{1a}\|_{\infty, \omega_{1a}} + \|f_2\|_{\infty, \omega_2} + \|f_{2a}\|_{\infty, \omega_{2a}}$ 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 A.8. (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, $\lim_{n \rightarrow \infty} \frac{n}{n+n_a} = p \in (0, 1)$; (ii) $g(y|x^*, w; \theta)$ is continuous in $\theta \in \Theta$, and Θ is a compact subset of \mathbb{R}^{d_θ} ; 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 A.9. (i) $-\infty < E[\ell(Z_i; \alpha_0)] < \infty$, $E[\ell(Z_i; \alpha)]$ is upper semicontinuous on \mathcal{A} under the metric $\|\cdot\|_s$; (ii) there is a finite $\kappa > 0$ and a random variable $U(Z_i)$ with $E\{U(Z_i)\} < \infty$ such that $\sup_{\alpha \in \mathcal{A}_n: \|\alpha - \alpha_0\|_s \leq \delta} |\ell(Z_i; \alpha) - \ell(Z_i; \alpha_0)| \leq \delta^\kappa U(Z_i)$.

Assumption A.10. (i) $p_2^{k_{2,n}}(\cdot)$ is a $k_{2,n} \times 1$ -vector of spline wavelet basis functions on \mathbb{R} , and for $j = 1, 1a$, $p_j^{k_{j,n}}(\cdot, \cdot)$ is a $k_{j,n} \times 1$ -vector of tensor product of spline wavelet basis functions on \mathbb{R}^2 ; (ii) $\min\{k_{1,n}, k_{1a,n}, k_{2,n}\} \rightarrow \infty$ and $\max\{k_{1,n}, k_{1a,n}, k_{2,n}\}/n \rightarrow 0$.

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

Lemma A.1. Under Assumptions A.7–A.10, we have $\|\widehat{\alpha}_n - \alpha_0\|_s = o_p(1)$.

A.5 Rates of Convergence

Given consistency, 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 $\mathcal{A}_{0sn} \equiv \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_0\|_s = o(1), \|\alpha\|_s \leq c_0 < c\}$. We assume that both \mathcal{A}_{0s} and \mathcal{A}_{0sn} are convex parameter spaces, and that $\ell(Z_i; \alpha + \tau v)$ is twice continuously differentiable at $\tau = 0$ for almost all Z_i and any direction $v \in \mathcal{A}_{0s}$.

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

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

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

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

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 A.11. (i) $\varsigma_j > \gamma_j$ for $j = 1, 1a, 2$; (ii) $\max\{k_{1,n}^{-\gamma_1/2}, k_{1a,n}^{-\gamma_{1a}/2}, k_{2,n}^{-\gamma_2}\} = o([n + n_a]^{-1/4})$.

Assumption A.12. (i) \mathcal{A}_{0s} is convex at α_0 and $\theta_0 \in \text{int}(\Theta)$; (ii) $\ell(Z_i; \alpha)$ is twice continuously pathwise differentiable with respect to $\alpha \in \mathcal{A}_{0s}$, and $\log g(y|x^*, w; \theta)$ is twice continuously differentiable at θ_0 .

Assumption A.13. $\sup_{\widetilde{\alpha} \in \mathcal{A}_{0s}} \sup_{\alpha \in \mathcal{A}_{0sn}} \left| \frac{d\ell(Z_i; \widetilde{\alpha})}{d\alpha} \left[\frac{\alpha - \alpha_0}{\|\alpha - \alpha_0\|_s} \right] \right| \leq U(Z_i)$ for a random variable $U(Z_i)$ with $E\{[U(Z_i)]^2\} < \infty$.

Assumption A.14. (i) $\sup_{v \in \mathcal{A}_{0s}: \|v\|_s=1} -E \left(\frac{d^2 \ell(Z_i; \alpha_0)}{d\alpha d\alpha^T} [v, v] \right) \leq C < \infty$; (ii) uniformly over $\tilde{\alpha} \in \mathcal{A}_{0s}$ and $\alpha \in \mathcal{A}_{0sn}$, we have

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

The assumptions are straightforward and standard. The following convergence rate theorem is a direct application of Theorem 3.2 of Shen and Wong (2004) to the local parameter space \mathcal{A}_{0s} and the local sieve space \mathcal{A}_{0sn} ; hence, we omit its proof.

Theorem A.1. Let $\gamma \equiv \min\{\gamma_1/2, \gamma_{1a}/2, \gamma_2\} > 1/2$. Under assumptions A.7–A.14, if $k_{1,n} = O\left([n + n_a]^{\frac{1}{\gamma_1+1}}\right)$, $k_{1a,n} = O\left([n + n_a]^{\frac{1}{\gamma_{1a}+1}}\right)$, and $k_{2,n} = O\left([n + n_a]^{\frac{1}{2\gamma_2+1}}\right)$, then

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

A.6 Conditions and Definitions for Asymptotic Normality

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

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

where

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

Let $\bar{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A} - \{\alpha_0\}$ under the metric $\|\cdot\|_2$. Then $(\bar{\mathbf{V}}, \|\cdot\|_2)$ is a Hilbert space and we can represent $\bar{\mathbf{V}} = \mathbb{R}^{d_\theta} \times \bar{\mathcal{U}}$ with $\bar{\mathcal{U}} \equiv \overline{\mathcal{F}_1 \times \mathcal{F}_{1a} \times \mathcal{F}_2 \times \mathcal{F}_2} - \{(f_{01}, f_{01a}, f_{02}, f_{02a})\}$.

Let $h = (f_1, f_{1a}, f_2, f_{2a})$ denote all the unknown densities. The pathwise first derivative can be written as

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

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

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

Note that

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

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

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

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

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

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

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

Also denote

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

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

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

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 $v^* \equiv (v_\theta^*, v_h^*)$ with $v_\theta^* \equiv (V_*)^{-1}\lambda$ and $v_h^* \equiv -\mu^* \times v_\theta^*$. Then the Riesz representation theorem implies: $\lambda^T (\theta - \theta_0) = \langle v^*, \alpha - \alpha_0 \rangle_2$ for all $\alpha \in \mathcal{A}$. In Section A.7, we show (A.9):

$$\lambda^T (\hat{\theta}_n - \theta_0) = \langle v^*, \hat{\alpha}_n - \alpha_0 \rangle_2 = \frac{1}{n + n_a} \sum_{i=1}^{n+n_a} \frac{d\ell(Z_i; \alpha_0)}{d\alpha} [v^*] + o_p\{(n + n_a)^{-1/2}\}. \quad (\text{A.9})$$

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 $\hat{\theta}_n$:

Assumption A.15. μ^* exists (i.e., $\mu^{*k} \in \bar{\mathcal{U}}$ for $k = 1, \dots, d_\theta$), and V_* is positive-definite.

Assumption A.16. There is a $v_n^* \in \mathcal{A}_n - \{\alpha_0\}$, such that $\|v_n^* - v^*\|_2 = o(1)$ and $\|v_n^* - v^*\|_2 \times \|\hat{\alpha}_n - \alpha_0\|_2 = o_P(\frac{1}{\sqrt{n+n_a}})$.

Assumption A.17. There is a random variable $U(Z_i)$ with $E\{[U(Z_i)]^2\} < \infty$ and a non-negative measurable function η with $\lim_{\delta \rightarrow 0} \eta(\delta) = 0$, such that, for all $\alpha \in \mathcal{N}_{0n}$,

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

Assumption A.18. Uniformly over $\bar{\alpha} \in \mathcal{N}_0$ and $\alpha \in \mathcal{N}_{0n}$,

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

Assumption A.19. $E\left\{\left(\frac{d\ell(Z_i; \alpha_0)}{d\alpha} [v_n^* - v^*]\right)^2\right\}$ goes to zero as $\|v_n^* - v^*\|_2$ goes to zero.

It is easily seen that Assumption A.19 is automatically satisfied when the latent parametric model is correctly specified. The other assumptions are necessary for the proofs. Recall the definitions of Fisher inner product and the Fisher norm:

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

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

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

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

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

A.7 Proof of Theorem 3.1 (Asymptotic Normality)

For any $\alpha \in \mathcal{N}_{0n}$, define $r[Z_i; \alpha, \alpha_0] \equiv \ell(Z_i; \alpha) - \ell(Z_i; \alpha_0) - \frac{d\ell(Z_i; \alpha_0)}{d\alpha} [\alpha - \alpha_0]$. Denote the centered empirical process indexed by any measurable function h as

$$\mu_n\{h(Z_i)\} \equiv \frac{1}{n + n_a} \sum_{i=1}^{n+n_a} \{h(Z_i) - E[h(Z_i)]\}.$$

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

$$\begin{aligned} 0 &\leq \frac{1}{n + n_a} \sum_{i=1}^{n+n_a} [\ell(Z_i; \hat{\alpha}) - \ell(Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*)] \\ &= \mu_n(\ell(Z_i; \hat{\alpha}) - \ell(Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*)) + E(\ell(Z_i; \hat{\alpha}) - \ell(Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*)) \\ &= \mp \varepsilon_n \times \frac{1}{n + n_a} \sum_{i=1}^{n+n_a} \frac{d\ell(Z_i; \alpha_0)}{d\alpha} [v_n^*] + \mu_n(r[Z_i; \hat{\alpha}, \alpha_0] - r[Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*, \alpha_0]) \\ &\quad + E(r[Z_i; \hat{\alpha}, \alpha_0] - r[Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*, \alpha_0]). \end{aligned}$$

In the following we will show that if we denote $c_n = o_p\{(n + n_a)^{-1/2}\}$, then

$$(n + n_a)^{-1} \sum_{i=1}^{n+n_a} \frac{d\ell(Z_i; \alpha_0)}{d\alpha} [v_n^* - v^*] = c_n; \tag{A.10}$$

$$E(r[Z_i; \hat{\alpha}, \alpha_0] - r[Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*, \alpha_0]) = \pm \varepsilon_n \langle \hat{\alpha} - \alpha_0, v^* \rangle_2 + \varepsilon_n c_n; \tag{A.11}$$

$$\mu_n(r[Z_i; \hat{\alpha}, \alpha_0] - r[Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*, \alpha_0]) = \varepsilon_n \times c_n. \tag{A.12}$$

Notice that assumptions A.7, A.8(ii)(iii), and A.12 imply $E\left(\frac{d\ell(Z_i; \alpha_0)}{d\alpha}[v^*]\right) = 0$. Under (A.10) - (A.12) we have:

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

Hence

$$\begin{aligned} \sqrt{n+n_a} \langle \hat{\alpha} - \alpha_0, v^* \rangle_2 &= \sqrt{n+n_a} \mu_n \left(\frac{d\ell(Z_i; \alpha_0)}{d\alpha}[v^*] \right) + o_P(1) \Rightarrow N(0, \sigma_*^2), \\ \sigma_*^2 &\equiv E \left\{ \left(\frac{d\ell(Z_i; \alpha_0)}{d\alpha}[v^*] \right)^2 \right\} = (v_\theta^*)^T E[S_{\theta_0}^T S_{\theta_0}] (v_\theta^*) = \lambda^T (V_*)^{-1} I_*(V_*)^{-1} \lambda. \end{aligned}$$

Thus, assumptions A.8(i), A.13, and A.15 together imply that $\sigma_*^2 < \infty$ and

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

To complete the proof, it remains to establish (A.10) - (A.12). Notice that (A.10) is implied by the Tchebychev inequality, i.i.d. data, and assumptions A.16 and A.19. For (A.11) and (A.12) we notice that $r[Z_i; \hat{\alpha}, \alpha_0] - r[Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*, \alpha_0] = \mp \varepsilon_n \times \frac{d^2\ell(Z_i; \bar{\alpha})}{d\alpha d\alpha^T}[\tilde{\alpha} - \alpha_0, v_n^*]$, in which $\tilde{\alpha} \in \mathcal{N}_{0n}$ is in between $\hat{\alpha}$ and $\hat{\alpha} \pm \varepsilon_n v_n^*$, and $\bar{\alpha} \in \mathcal{N}_0$ is in between $\tilde{\alpha} \in \mathcal{N}_{0n}$ and α_0 . Therefore, for (A.11), by the definition of inner product $\langle \cdot, \cdot \rangle_2$, we have:

$$\begin{aligned} &E(r[Z_i; \hat{\alpha}, \alpha_0] - r[Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*, \alpha_0]) \\ &= \mp \varepsilon_n \times E \left(\frac{d^2\ell(Z_i; \bar{\alpha})}{d\alpha d\alpha^T}[\tilde{\alpha} - \alpha_0, v_n^*] \right) \\ &= \pm \varepsilon_n \times \langle \tilde{\alpha} - \alpha_0, v_n^* \rangle_2 \mp \varepsilon_n \times E \left(\frac{d^2\ell(Z_i; \bar{\alpha})}{d\alpha d\alpha^T}[\tilde{\alpha} - \alpha_0, v_n^*] - \frac{d^2\ell(Z_i; \alpha_0)}{d\alpha d\alpha^T}[\tilde{\alpha} - \alpha_0, v_n^*] \right) \\ &= \pm \varepsilon_n \times \langle \hat{\alpha} - \alpha_0, v_n^* \rangle_2 \pm \varepsilon_n \times \langle \tilde{\alpha} - \hat{\alpha}, v_n^* \rangle_2 + o_P\left(\frac{\varepsilon_n}{\sqrt{n+n_a}}\right) \\ &= \pm \varepsilon_n \times \langle \hat{\alpha} - \alpha_0, v^* \rangle_2 + O_P(\varepsilon_n^2) + o_P\left(\frac{\varepsilon_n}{\sqrt{n+n_a}}\right) \end{aligned}$$

in which the last two equalities hold due to the definition of $\tilde{\alpha}$, assumptions A.16 and A.18, and $\langle \hat{\alpha} - \alpha_0, v_n^* - v^* \rangle_2 = o_P\left(\frac{1}{\sqrt{n+n_a}}\right)$ and $\|v_n^*\|_2^2 \rightarrow \|v^*\|_2^2 < \infty$. Hence, (A.11) is satisfied. For (A.12), we notice

$$\mu_n (r[Z_i; \hat{\alpha}, \alpha_0] - r[Z_i; \hat{\alpha} \pm \varepsilon_n v_n^*, \alpha_0]) = \mp \varepsilon_n \times \mu_n \left(\frac{d\ell(Z_i; \tilde{\alpha})}{d\alpha}[v_n^*] - \frac{d\ell(Z_i; \alpha_0)}{d\alpha}[v_n^*] \right)$$

in which $\tilde{\alpha} \in \mathcal{N}_{0n}$ is in between $\hat{\alpha}$ and $\hat{\alpha} \pm \varepsilon_n v_n^*$. Since the class $\left\{ \frac{d\ell(Z_i; \tilde{\alpha})}{d\alpha} [v_n^*] : \tilde{\alpha} \in \mathcal{A}_{0s} \right\}$ is Donsker under assumptions A.7, A.8, A.12, and A.13, and since

$$E \left\{ \left(\frac{d\ell(Z_i; \tilde{\alpha})}{d\alpha} [v_n^*] - \frac{d\ell(Z_i; \alpha_0)}{d\alpha} [v_n^*] \right)^2 \right\} = E \left\{ \left(\frac{d^2\ell(Z_i; \bar{\alpha})}{d\alpha d\alpha^T} [\tilde{\alpha} - \alpha_0, v_n^*] \right)^2 \right\}$$

goes to zero as $\|\tilde{\alpha} - \alpha_0\|_s$ goes to zero under assumption A.17, we have (A.12) holds.

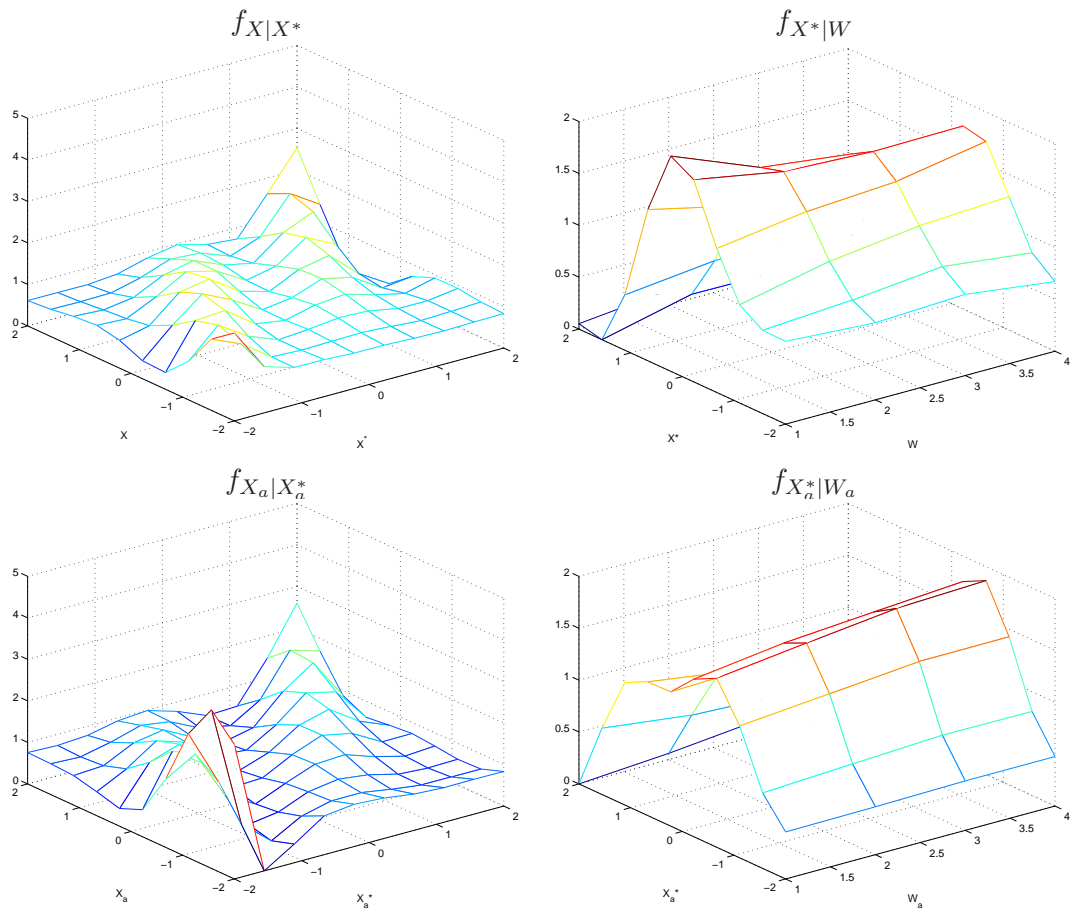


Figure 1: Analysis of the nutrition data set. Left side: sieve estimated measurement error density. Right side: sieve estimated latent variable density.