Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments

Yingyao Hu
Department of Economics
The University of Texas at Austin
1 University Station C3100
BRB 1.116
Austin, TX 78712
hu@eco.utexas.edu

S. M. Schennach*
Department of Economics
University of Chicago
1126 East 59th Street
Chicago IL 60637
smschenn@uchicago.edu

First version: December 2004; This version: June 2006.

Accompanying presentation slides for this paper are available at:
http://home.uchicago.edu/∼smschenn/ncmetalk.pdf

Abstract

While the literature on nonclassical measurement error traditionally relies on the availability of an auxiliary dataset containing correctly measured observations, we establish that the availability of instruments enables the identification of a large class of nonclassical nonlinear errors-in-variables models with continuously distributed variables. Our main identifying assumption is that, conditional on the value of the true regressors, some "measure of location" of the distribution of the measurement error (e.g. its mean, mode or median) is equal to zero. The proposed approach relies on the eigenvalue-eigenfunction decomposition of an integral operator associated with specific joint probability densities. The main identifying assumption is used to "order" the eigenfunctions so that the decomposition is unique. We propose a convenient sieve-based estimator, derive its asymptotic properties and investigate its finite-sample behavior through Monte Carlo simulations. An example of application to the relationship between earnings and divorce rates is also provided.

Keywords: Nonclassical measurement error, nonlinear errors-in-variables model, instrumental variable, operator, semiparametric estimator, sieve maximum likelihood.

[*] Corresponding author. S. M. Schennach acknowledges support from the National Science Foundation via grant SES-0452089. The authors would like to thank Maxwell Stinchcombe and Xiaohong Chen, as well as seminar audiences at the University of Illinois at Urbana-Champain, New York University, Northwestern University, Mannheim University, Paris-Crest and the Cemmap/ESRC Econometric Study Group Workshop on Semiparametric Methods for helpful comments.

1 Introduction

In recent years, there has been considerable progress in the development of inference methods that account for the presence of measurement error in the explanatory variables in nonlinear models (see, for instance, Chesher (1991), Lewbel (1996), Chesher (1998), Lewbel (1998), Hausman (2001), Chesher (2001), Chesher, Dumangane, and Smith (2002), Hong and Tamer (2003), Carrasco and Florens (2005)). The case of classical measurement errors, in which the measurement error is either independent from the true value of the mismeasured variable or has zero mean conditional on it, has been thoroughly studied. In this context, approaches that establish identifiability of the model, and provide estimators that are either consistent or root n consistent and asymptotically normal have been devised when either instruments (Hausman, Newey, Ichimura, and Powell (1991), Hausman, Newey, and Powell (1995), Newey (2001), Wang and Hsiao (1995), Schennach (2004b)), repeated measurements (Hausman, Newey, Ichimura, and Powell (1991), Hausman, Newey, and Powell (1995), Li (2002), Schennach (2004a), Schennach (2004c)) or validation data (Hu and Ridder (2004)) are available.

However, the are a number of practical applications where the assumption of classical measurement error is not appropriate (Bound, Brown, and Mathiowetz (2001)). In the case of discretely distributed regressors, instrumental variable estimators that are robust to the presence of such "nonclassical" measurement error have been developed for binary regressors (Mahajan (2006), Lewbel (2006)) and general discrete regressors (Hu (2005)). Unfortunately, these results cannot trivially be extended to continuously distributed variables, because the number of nuisance parameters needed to describe the measurement error distribution (conditional on given values of the observable variables) becomes infinite. Identifying these parameters thus involves solving operator equations that exhibit potential ill-defined inverse problems (similar to those discussed in Carrasco, Florens, and Renault (2005), Darolles, Florens, and Renault (2002), and Newey and Powell (2003)).

In the case of continuously distributed variables (in both linear or nonlinear models), the only approach capable of handling nonclassical measurement errors proposed so far has been the use of an auxiliary dataset containing correctly measured observations (Chen, Hong, and Tamer (2005), Chen, Hong, and Tarozzi (2005)). Unfortunately, the availability of such a clean data set is the exception rather than the rule. Our interest in instrumental variables is driven by the fact that instruments suitable for the proposed approach are conceptually similar to the ones used in conventional instrumental variable methods and researchers will have little difficulty identifying appropriate instrumental variables in typical datasets.

Our approach relies on the observation that, even though the measurement error may not have zero mean conditional on the true value of the regressor, perhaps some other measure of location, such as the median or the mode, could still be zero. This type of nonclassical measurement error has been observed, for instance, in the self-reported income found in the Current Population Survey (CPS).[1] Thanks to the availability of validation data for one of the years of the survey, it was found that, although measurement error is correlated with true income, the median of misreported income conditional on true income is in fact equal to the true income (Bollinger (1998)). In another study on the same dataset, it was found that the mode of misreported income conditional on true income is also equal to the true income (see Bound and Krueger (1991) and Figure 1 in Chen, Hong, and Tarozzi (2005)).

There are numerous plausible settings where the conditional mode, median, or some other quantile, of the error could be zero even though its conditional mean is not. First, if respondents are more likely to report values close to the truth than any particular value far from the truth, then the mode of the measurement error would be zero. This is a very plausible form of measurement error that even allows for systematic over- or underreporting. In addition, data truncation usually preserves the mode, but not the mean, provided the truncation is not so severe that the mode itself is deleted. This assumption regarding the mode can be viewed as a generalization of the assumption, used by Mahajan (2006) and Lewbel (2006) in the simple misclassified binary variable case, that survey respondents are

[1] Bureau of Labor Statistics and Bureau of Census, http://www.bls.census.gov/cps/cpsmain.htm

more likely to report the truth than to lie. Of course, in the continuous case covered here, this assumption is particularly weak, since there are an infinite number of alternatives and respondents would literally have to collude on misreporting in a similar way in order to violate the mode assumption.

Second, if respondents are equally likely to over- or under-report, but not by the same amounts on average, then the median of the measurement error is zero. This could occur perhaps because the observed regressor is a nonlinear monotonic function (e.g., a logarithm) of some underlying mismeasured variable with symmetric errors. Such a nonlinear function would preserve the zero median, but not the zero mean of the error. Another important case is data censoring, which also preserves the median, as long as the upper censoring point is above the median and the lower censoring point is below the median.

Third, in some cases, a quantile other than the median might be appropriate. For instance, tobacco consumption is likely to be either truthfully reported or under-reported and, in that case, the topmost quantile of the error conditional on the truth would plausibly equal true consumption.

In order to encompass practically relevant cases such as these, which so far could only have been analyzed in the presence of auxiliary correctly measured data, our approach relies on the general assumption that some given "measure of location" (e.g. the mean, the mode, the median, or some other quantile) characterizing the distribution of the observed regressor conditional on the true regressor is left unaffected by the presence of measurement error. This framework is also sufficiently general to include measurement error models in which the true regressor and the errors enter the model in a nonseparable fashion.

The paper is organized as follows. We first provide a general proof of identification before introducing a semiparametric sieve estimator that is shown to be root n consistent and asymptotically normal. Our identification is fully nonparametric and therefore establishes identification in the presence of measurement error of any model that would be identified in the absence of measurement error. Our estimation framework encompasses models which, when expressed in terms of the measurement error-free variables, take the form of either parametric likelihoods or (conditional or unconditional) moment restrictions and automatically provides a corresponding measurement error-robust semiparametric instrumental variable estimator. This framework therefore addresses nonclassical measurement error issues in most of the widely used models, including probit, logit, tobit and duration models, in addition to conditional mean and quantile regressions, as well as nonseparable models (thanks to their relationship with quantile restrictions). The finite sample properties of the estimator are investigated via Monte Carlo simulations, while the usefulness of our approach is motivated through a simple example of an application to the study of the relationship between divorce rates and income, which is measured with possibly nonclassical error.

2 Identification

The "true" model is defined by the density of the dependent variable y conditional on the true regressor x^*, denoted $f_{y|x^*}(y|x^*)$. However, x^* is not observed, only its error-contaminated counterpart, x, is observed. In this section, we rely on the availability of an instrument (or a repeated measurement) z to show that $f_{y|x^*}(y|x^*)$ and, more generally, $f_{yx^*}(y, x^*)$, is identified from the knowledge of the joint density of all observed variables $f_{yxz}(y, x, z)$. Our treatment can be straightforwardly extended to allow for the presence of a vector w of additional correctly measured regressors, merely by conditioning all densities on w. Although we consider scalar-valued x^* in the sequel, for the sake of simplicity of exposition, our general approach is clearly applicable to multivariate settings, and we will note whenever the multivariate extension requires special attention. Also, the instrument z is considered univariate here, but multivariate instruments Z can easily be used, for instance, simply by defining z as the predicted value of the least-squares projection of x on Z.

2.1 Basic integral relationships

To state our identification result, we start by making natural assumptions regarding the conditional densities of all the variables of the model. Let $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{X}^*$ and $\mathcal{Z}$ denote the supports of the densities of the random variables y, x, x^* and z, respectively.

Assumption 1 *(i)* $f_{y|xx^*z}(y|x, x^*, z) = f_{y|x^*}(y|x^*)$ *for all* $(y, x, x^*, z) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$ *and (ii)* $f_{x|x^*z}(x|x^*, z) = f_{x|x^*}(x|x^*)$ *for all* $(x, x^*, z) \in \mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$*.*

Remark: Assumption 1(i) indicates that x and z do not provide any more information about y than x^* already provides, while Assumption 1(ii) specifies that z does not provide any more information about x than x^* already provides. The first assumption could be interpreted as a standard exclusion restriction, that is, z does not affect y directly, but only through its effect on x^*. The second assumption implies that the instrument contains no information regarding the measurement error, once the value of x^* is known. Conditional independence restrictions are widely used in the recent econometrics literature (e.g. Holderlein and Mammen (2006), Heckman and Vytlacil (2005), Altonji and Matzkin (2005)). Our assumptions regarding the instrument z are sufficiently general to encompass both the repeated measurement and the instrumental variable cases in a single framework. In the repeated measurement case, having the measurement error on the two measurements z and x be mutually independent conditional on x^* will be sufficient to satisfy Assumption 1. Note that while we will refer to y as the “dependent variable”, it should be clear that it could also contain another error-contaminated measurement of x^* or even a type of instrument that is “caused by” x^*, as discussed further in Section 2.4 and in Chalak and White (2006). Finally, note that our assumptions allow for the measurement error $(x - x^*)$ to be correlated with x^*, which is crucial in the presence of potentially nonclassical measurement error.

Assumption 1 implies that

$$\begin{aligned}
f_{yx|z}(y,x|z) &= \int f_{yxx^*|z}(y,x,x^*|z)\,dx^* \\
&= \int f_{y|xx^*z}(y|x,x^*,z)\,f_{xx^*|z}(x,x^*|z)\,dx^* \\
&= \int f_{y|x^*}(y|x^*)\,f_{xx^*|z}(x,x^*|z)\,dx^* \\
&= \int f_{y|x^*}(y|x^*)\,f_{x|x^*z}(x|x^*,z)\,f_{x^*|z}(x^*|z)\,dx^* \\
&= \int f_{y|x^*}(y|x^*)\,f_{x|x^*}(x|x^*)\,f_{x^*|z}(x^*|z)\,dx^*
\end{aligned}$$

or

$$f_{yx|z}(y,x|z) = \int f_{x|x^*}(x|x^*)\,f_{y|x^*}(y|x^*)\,f_{x^*|z}(x^*|z)\,dx^*. \tag{1}$$

To facilitate the proof of identification, is it useful to note that any function of two variables can be associated with an integral operator. For instance, the function $f_{yx|z}(y,x|z)$ (for a fixed y) can be associated with the operator $L_{y;x|z}$, defined as

$$L_{y;x|z}g = \int f_{yx|z}(y,\cdot|z)\,g(z)\,dz.$$

The notation emphasizes that y is regarded as a parameter on which $L_{y;x|z}$ depends, while the operator itself maps functions of z onto functions of x. More specifically, this operator maps the function $g(z)$ onto the function $\left[L_{y;x|z}g\right](x) = \int f_{yx|z}(y,x|z)\,g(z)\,dz$. Similarly, we define the operators $L_{x|z}, L_{x|x^*}, L_{x^*|z}$, and $L_{y;x^*|x^*}$ as

$$\begin{aligned}
L_{x|z}g &= \int f_{x|z}(\cdot|z)\,g(z)\,dz \\
L_{x|x^*}g &= \int f_{x|x^*}(\cdot|x^*)\,g(x^*)\,dx^* \\
L_{x^*|z}g &= \int f_{x^*|z}(\cdot|z)\,g(z)\,dz \\
L_{y;x^*|x^*}g &= f_{y|x^*}(y|\cdot)\,g(\cdot)\,.
\end{aligned}$$

Note that $L_{y;x^*|x^*}$ operator is a "diagonal" operator[2] since it is just a multiplication by a function (for a given y), i.e. $\left[L_{y;x^*|x^*}g\right](x^*) = f_{y|x^*}(y|x^*)\,g(x^*)$. By calculating $L_{y;x|z}g$ for

[2] The rationale behind the notation $L_{y;x^*|x^*}$ is that this operator can also be written as $\left[L_{y;x^*|x^*}g\right](u) = \int f_{yx^*|x^*}(y,u|x^*)\,g(x^*)\,dx^* = \int f_{y|x^*}(y|x^*)\,\delta(u-x^*)\,g(x^*)\,dx^* = f_{y|x^*}(y|u)\,g(u)$, where $\delta(\cdot)$ denotes a Dirac delta function.

an arbitrary absolutely integrable[3] function $g(\cdot)$, we can find an operator equation that is equivalent to Equation (1):

$$\begin{aligned}
\left[L_{y;x|z}g\right](x) &= \int f_{yx|z}(y,x|z)\,g(z)\,dz \\
&= \int\int f_{x|x^*}(x|x^*)\,f_{y|x^*}(y|x^*)\,f_{x^*|z}(x^*|z)\,dx^* g(z)\,dz \\
&= \int f_{x|x^*}(x|x^*)\,f_{y|x^*}(y|x^*)\int f_{x^*|z}(x^*|z)\,g(z)\,dzdx^* \\
&= \int f_{x|x^*}(x|x^*)\,f_{y|x^*}(y|x^*)\left[L_{x^*|z}g\right](x^*)\,dx^* \\
&= \int f_{x|x^*}(x|x^*)\left[L_{y;x^*|x^*}L_{x^*|z}g\right](x^*)\,dx^* \\
&= \left[L_{x|x^*}L_{y;x^*|x^*}L_{x^*|z}g\right](x), \qquad (2)
\end{aligned}$$

where we have used, (i) Equation (1), (ii) an interchange of the order of integration (justified by the absolute integrability of the integrand, by Fubini's Theorem), (iii) the definition of $L_{x^*|z}$, (iv) the definition of $L_{y;x^*|x^*}$ operating on the function $\left[L_{x^*|z}g\right]$ and (v) the definition of $L_{x|x^*}$ operating on the function $\left[L_{y;x^*|x^*}L_{x^*|z}g\right]$.

Equation (2) thus implies the following operator equivalence

$$L_{y;x|z} = L_{x|x^*}L_{y;x^*|x^*}L_{x^*|z}. \qquad (3)$$

By integration over y we similarly get

$$L_{x|z} = L_{x|x^*}L_{x^*|z}, \qquad (4)$$

since $\int L_{y;x|z}dy = L_{x|z}$ and $\int L_{y;x^*|x^*}dy = I$, the identity operator.

2.2 Injectivity

Our method of proof will require the following assumption.

Assumption 2 *$L_{x|z}$ and $L_{x|x^*}$ are injective.*

[3] It is sufficient to consider absolutely integrable functions because, in the case of an integral operator having a probability density as its kernel, such as $L_{y;x|z}$, we have $f_{yx|z}(y,x|z_0) = \lim_{n\to\infty} L_{y;x|z}g_{n,z_0}$ where $g_{n,z_0}(z) = n1\left(|z-z_0| \le n^{-1}\right)$, a sequence of absolutely integrable functions. The kernel $f_{yx|z}(y,x|z_0)$ of this integral operator is therefore uniquely determined by evaluating this limit for all values of z_0.

An operator L is said to be *injective* if its inverse L^{-1} is defined over the range of the operator L (see Section 3.1 in Carrasco, Florens, and Renault (2005)). In a finite-dimensional space, the qualifier "injective" is synonymous with "invertible", but in an infinite-dimensional space the distinction is needed to account for the fact that inverses are often defined only over a restricted domain. As discussed in Carrasco, Florens, and Renault (2005), the weaker notion of injectivity is the concept needed to establish identification. In our setup, the inverses are guaranteed to be defined over a sufficiently large domain because the results of the inversions (such as $L_{x|x^*}^{-1}L_{x|z} = L_{x^*|z}$, from Equation (4)) always yield a well-defined integral operator. Assumption 2 could also be stated in terms of the injectivity of $L_{x^*|z}$ and $L_{x|x^*}$: Since $L_{x|z} = L_{x|x^*}L_{x^*|z}$ under Assumption 1, injectivity of $L_{x^*|z}$ and $L_{x|x^*}$ implies injectivity of $L_{x|z}$ and $L_{x|x^*}$.

Intuitively, $L_{x|x^*}$ (or $L_{x|z}$) will be injective if there is enough variation in the density of x for different values of x^* (or z). For instance, a simple case where Assumption 2 is violated is when $f_{x|x^*}(x|x^*)$ or $f_{x|z}(x|z)$ are uniform. In general, however, Assumption 2 is quite weak and numerous results enabling its verification under more primitive conditions exist in the literature.

First, Assumption 2 is related to the identification conditions employed in Newey and Powell (2003) (see Proposition 2.1). Newey and Powell's assumption has the general form "for all $g(z)$ (for which $E[g(z)|x]$ is defined) $E[g(z)|x] = 0$ implies that $g(z) = 0$." If the densities of x and z are bounded and nonvanishing over the interior of their respective supports, then this condition is equivalent[4] to $\int g(z) f_{x|z}(x|z)\,dz = 0$ implies that $g(z) = 0$, which is equivalent to $L_{x|z}$ being injective. A similar reasoning applies to $L_{x|x^*}$, provided that the marginal densities of x and x^* are bounded and nonvanishing on their respective supports. A nice consequence of this connection is that known results regarding the so-called completeness of exponential families of distributions can be used to formulate primitive conditions for operators to be injective (as in Newey and Powell (2003)). Under the assumption

[4] $E[g(z)|x] = f_x^{-1}(x)\int g(z) f_z(z) f_{x|z}(x|z)\,dz = 0 \Leftrightarrow \int (g(z) f_z(z)) f_{x|z}(x|z)\,dz = 0$ if $0 < f_x(x) < \infty$ and $0 < f_z(z) < \infty$ over the interior of their respective supports.

that all conditional densities involved are bounded, the weaker notion of bounded completeness (as discussed in Blundell, Chen, and Kristensen (2003)) can also be used to find more general families of distributions leading to injective operators.

An alternative way to verify Assumption 2 under primitive conditions is to follow the approach taken in Darolles, Florens, and Renault (2002) by constructing a so-called singular value decomposition of the operators of interest and by verifying that none of the singular values vanish. We illustrate the approach for the $L_{x|z}$ operator — a similar treatment will apply to $L_{x|x^*}$. Let $\mathcal{H}_q$ denote the Hilbert space associated with the inner product

$$\langle g, h\rangle_q = \int g(z)\, h(z)\, (q(z))^{-2}\, dz$$

where g, h and q are functions from $\mathbb{R}$ to $\mathbb{R}$ and $q(z)$ is nonvanishing. The idea is then to note that $L_{x|z}$ is a compact operator, when viewed as a mapping from $\mathcal{H}_q$ to $\mathcal{H}_1$, where $q(z)$ is selected so that

$$\int\int f_{x|z}^2(x|z)\, q^2(z)\, dxdz < \infty. \tag{5}$$

The condition (5) implies that $L_{x|z}$ is a Hilbert-Schmidt operator, which is necessarily compact (see Theorem 2.32 in Carrasco, Florens, and Renault (2005)). This in turn implies the existence of a singular value decomposition,

$$L_{x|z}g = \sum_{i=1}^{\infty} \phi_i \mu_i \langle \psi_i, g\rangle_q$$

where $\{\mu_i\}$ is a sequence of non-negative[5] real numbers, $\{\phi_i\}$ is an orthonormal basis of $\mathcal{H}_1$ and $\{\psi_i\}$ is an orthonormal basis of $\mathcal{H}_q$. With this representation in hand, the inverse is simply given by

$$L_{x|z}^{-1}g = \sum_{i=1}^{\infty} \mu_i^{-1} \psi_i \langle \phi_i, g\rangle_1$$

and a sufficient condition of injectivity is that $\mu_i > 0$ for all i. Note that having positive singular values μ_i does not exclude that $\mu_i \to 0$ as $i \to \infty$ and the inverses of $L_{x|x^*}$ or $L_{x|z}$ will generally not be continuous. However, as mentioned earlier, for identification purpose,

[5] A negative μ_i can always be avoided by replacing ψ_i by $-\psi_i$.

injectivity is sufficient, whether or not the inverse is continuous (Carrasco, Florens, and Renault (2005)).

It is tempting to draw an analogy between injectivity of an operator and invertibility of a matrix. However, this analogy is dangerous if taken too literally. For instance, the fact that a matrix needs to be square to be invertible does not in any way imply that the support of the kernel of the $L_{x|x^*}$ operator needs to be square for $L_{x|x^*}$ to be injective. Fundamentally, this can be the case, because, for instance, the cardinality of the $[-1, 1]$ interval is the same as the $[-2, 2]$ interval, since they can be put into a one-to-one correspondence through the mapping $x \mapsto 2x$. The same reasoning does not apply in the discrete case associated with a rectangular matrix.

In the case where x^* and x are multivariate (and of the same dimension, by construction), the assumption of injectivity of $L_{x|x^*}$ generalizes very naturally. Injectivity of $L_{x|z}$ in multivariate settings is also natural if the dimensions of x and z are the same. If the dimension of z is less than the dimension of x^* or if z contains too many colinear elements, identification will not be possible, as expected. If the dimension of z exceeds the dimension of x, some elements of z can be dropped for the purpose of establishing identification, since identification with a subset of the available instruments trivially implies identification for the full set of instruments.[6]

2.3 Eigenvalue-Eigenfunction decomposition

Having motivated the assumption that $L_{x|x^*}$ and $L_{x|z}$ are injective, we are ready to prove identification of our model. Since $L_{x|x^*}$ is injective, Equation (4) can be written as

$$L_{x^*|z} = L_{x|x^*}^{-1} L_{x|z} \tag{6}$$

[6] If one wishes to state an identification result that explicitly allows for overidentification (i.e. allowing for the dimension of z to exceed the dimension of x), the assumption of injectivity of $L_{x|z}$ must be replaced by the assumption of injectivity of $L_{x|z} L_{x|z}^*$, where $*$ denotes the adjoint, it order to be able to invoke generalized inverses.

and this expression for $L_{x^*|z}$ can be substituted into Equation (3) to yield

$$L_{y;x|z} = L_{x|x^*} L_{y;x^*|x^*} L_{x|x^*}^{-1} L_{x|z} \tag{7}$$

Since $L_{x|z}$ is injective we also have[7]

$$L_{y;x|z} L_{x|z}^{-1} = L_{x|x^*} L_{y;x^*|x^*} L_{x|x^*}^{-1}. \tag{8}$$

The operator $L_{y;x|z} L_{x|z}^{-1}$ is defined in terms of densities of the observable variables x, y and z and can therefore be considered known. Equation (8) states that the known operator $L_{y;x|z} L_{x|z}^{-1}$ admits a spectral decomposition taking the form of an eigenvalue-eigenfunction decomposition.[8] The eigenvalues of the $L_{y;x|z} L_{x|z}^{-1}$ operator are given by the "diagonal elements" of the $L_{y;x^*|x^*}$ operator (i.e. $\left\{f_{y|x^*}\left(y|x^*\right)\right\}$ for a given y and for all x^*) while the eigenfunctions of the $L_{y;x|z} L_{x|z}^{-1}$ operator are given by the kernel of the integral operator $L_{x|x^*}$, i.e. $\left\{f_{x|x^*}\left(\cdot|x^*\right)\right\}$ for all x^*. Although Equation (8) establishes the existence of an eigenvalue-eigenfunction decomposition (which is no trivial matter since, in general, $L_{y;x|z} L_{x|z}^{-1}$ is a nonnormal and noncompact operator), it does not prove that this decomposition is unique. Fortunately, only a few more assumptions are sufficient to guarantee a unique decomposition, thereby establishing that the model is identified.

Theorem XV.4.5 in Dunford and Schwartz (1971) provides necessary and sufficient conditions for the existence of a unique representation of the so-called spectral decomposition of a linear operator. If a bounded operator T can be written as $T = A + N$ where A is an

[7] To allow for overidentification (i.e. the dimension of z exceeding the dimension of x), Equation (8) must be slightly modified. Applying $L_{x|z}^*$ from the right on each side of Equation (7) yields $L_{y;x|z} L_{x|z}^* = L_{x|x^*} L_{y;x^*|x^*} L_{x|x^*}^{-1} L_{x|z} L_{x|z}^*$. If $L_{x|z} L_{x|z}^*$ is invertible we also have $L_{y;x|z} L_{x|z}^* \left(L_{x|z} L_{x|z}^*\right)^{-1} = L_{x|x^*} L_{y;x^*|x^*} L_{x|x^*}^{-1}$. The remainder of the treatment thus follows, replacing each occurence of $L_{y;x|z} L_{x|z}^{-1}$ by $L_{y;x|z} L_{x|z}^* \left(L_{x|z} L_{x|z}^*\right)^{-1}$.

[8] A spectral decomposition of an operator T takes the form of an eigenvalue-eigenfunction decomposition when $(T - \lambda I)$ is not one-to-one for all eigenvalues λ in the spectrum. This can be verified to be the case here, because all eigenfunctions are well-behaved functions that are mapped to 0 under $(T - \lambda I)$. An example of a spectral decomposition that is not an eigenvalue-eigenfunction decomposition would be one where some of the eigenfunctions lie outside the space of functions considered (e.g. can only be reached by a limiting process).

operator of the form

$$A = \int_{\sigma} \lambda P\left(d\lambda\right) \tag{9}$$

where P is a projection-valued measure[9] supported on the spectrum σ, a subset of the complex plane, and N is a "quasi-nilpotent" operator commuting with A, then this representation is unique. The result is applicable to our situation (with $T = L_{y;x|z} L_{x|z}^{-1}$), in the special case where $N = 0$ and $\sigma \subset \mathbb{R}$. The spectrum σ is simply the range of $f_{y|x^*}\left(y|x^*\right)$, that is, $\left\{f_{y|x^*}\left(y|x^*\right) : x^* \in \mathcal{X}^*\right\}$. The projection-valued measure P assigned to any subset Λ of $\mathbb{R}$ is

$$P\left(\Lambda\right) = L_{x|x^*} I_{\Lambda} L_{x|x^*}^{-1} \tag{10}$$

where the operator I_{Λ} is defined via

$$\left[I_{\Lambda} g\right]\left(x^*\right) = 1\left(f_{y|x^*}\left(y|x^*\right) \in \Lambda\right) g\left(x^*\right).$$

Note that it can easily be verified that $P\left(\Lambda\right)$ is idempotent using Equation (10). An equivalent way to define $P\left(\Lambda\right)$ is by introducing the subspace

$$\mathcal{S}\left(\Lambda\right) = \text{span}\left\{f_{x|x^*}\left(\cdot|x^*\right) : x^* \text{ such that } f_{y|x^*}\left(y|x^*\right) \in \Lambda\right\} \tag{11}$$

for any subset Λ of the spectrum σ. The projection $P\left(\Lambda\right)$ is then uniquely defined by specifying that its range is $\mathcal{S}\left(\Lambda\right)$ and that its null space is $\mathcal{S}\left(\sigma \backslash \Lambda\right)$.

The fact that $\int_{\sigma} \lambda P\left(d\lambda\right) = L_{x|x^*} L_{y;x^*|x^*} L_{x|x^*}^{-1}$, thus connecting Equation (8) with Equation (9), can be shown by noting that

$$P\left(d\lambda\right) \equiv \left(\frac{d}{d\lambda} P\left(\left[-\infty, \lambda\right]\right)\right) d\lambda = L_{x|x^*} \left(\frac{dI_{[-\infty,\lambda]}}{d\lambda} d\lambda\right) L_{x|x^*}^{-1}$$

and that

$$\int_{\sigma} \lambda P\left(d\lambda\right) = L_{x|x^*} \left(\int_{\sigma} \lambda \frac{dI_{[-\infty,\lambda]}}{d\lambda} d\lambda\right) L_{x|x^*}^{-1},$$

[9] Just like a real-valued measure assigns a real number to each set in some field, a projection-valued measure, assigns a projection operator to each set in some field (here, the Borel σ-field). A projection operator Q, is one that is *idempotent*, i.e. $QQ = Q$.

where the operator in parenthesis can be obtained by calculating its effect on some function $g(x^*)$, as follows

$$\begin{aligned}\left[\int_\sigma \lambda \frac{dI_{[-\infty,\lambda]}}{d\lambda} d\lambda g\right](x^*) &= \int_\sigma \lambda \frac{d}{d\lambda} 1\left(f_{y|x^*}(y|x^*) \in [-\infty,\lambda]\right) g(x^*)\, d\lambda \\ &= \int_\sigma \lambda \delta\left(\lambda - f_{y|x^*}(y|x^*)\right) g(x^*)\, d\lambda \\ &= f_{y|x^*}(y|x^*)\, g(x^*) \\ &= \left[L_{y;x^*|x^*} g\right](x^*).\end{aligned}$$

where we have used that the differential of a step function $1(\lambda \leq 0)$ is a Dirac delta $\delta(\lambda)$, which has the property that $\int \delta(\lambda)\, h(\lambda)\, d\lambda = h(0)$ for any function $h(\lambda)$ continuous at $\lambda = 0$, and, in particular, for $h(\lambda) = \lambda$. Hence, we can indeed conclude that $\int_\sigma \lambda P(d\lambda) = L_{x|x^*} L_{y;x^*|x^*} L_{x|x^*}^{-1}$.

The result that the representation $T = \int_\sigma \lambda P(d\lambda)$ is unique requires that the operator T be bounded. Since the operator T is bounded (in a suitably defined operator norm) if the largest element of the spectrum is bounded,[10] the following condition is sufficient to ensure that T is bounded in our case:

Assumption 3 $\sup_{y\in\mathcal{Y}} \sup_{x^*\in\mathcal{X}^*} f_{y|x^*}(y|x^*) < \infty$.

Note that, for the purpose of establishing identification, there is no requirement that $f_{y|x^*}(y|x^*)$ be nonzero or bounded away from zero, because our approach does not involve inverting the T operator.

2.4 Uniqueness

Having established uniqueness of the decomposition (9) does not yet imply that the representation (8) is unique. The situation is analogous to standard matrix diagonalization, where eigenvectors are (i) unique only up to scale (or up to a linear combination when eigenvalues

[10] This follows from Lemma XVIII.2.2 in Dunford and Schwartz (1971), setting σ to be the whole spectrum, so that the restriction of the operator to the subspace of its domain associated with σ is, in fact, the whole domain of the operator (in Dunford and Schwartz's notation $E(\sigma)\mathfrak{X} = \mathfrak{X}$ and $T|E(\sigma)\mathfrak{X} = T|\mathfrak{X} = T$).

are degenerate) and (ii) can be "pasted" in any order to form a transformation matrix. In the present, more complex, context of operator diagonalization, these issues can be summarized as follows:

1. Each eigenvalue λ is associated with a unique subspace $\mathcal{S}(\{\lambda\})$, for $\mathcal{S}(\cdot)$ as defined in Equation (11). However, there are multiple ways to select a basis of functions whose span defines that subspace.

    (a) Each basis function can always be multiplied by a constant.

    (b) Also, if $\mathcal{S}(\{\lambda\})$ has more than one dimension (i.e. if λ is degenerate), a new basis can be defined in terms of linear combinations of functions of the original basis.

2. There is a unique mapping between λ and $\mathcal{S}(\{\lambda\})$, but one is free to index the eigenvalues by some other variable (here x^*) and represent the diagonalization by a function $\lambda(x^*)$ and the family of subspaces $\mathcal{S}(\{\lambda(x^*)\})$. The choice of the mapping $\lambda(x^*)$ is not unique.[11]

We first address issue 1a, namely that the kernel of the operator $L_{x|x^*}$ could be replaced by $f_{x|x^*}(x|x^*)\,s(x^*)$ for some nonvanishing function $s(x^*)$ without changing the value of $P(\Lambda)$ in Equation (10). Fortunately, the fact that $\int f_{x|x^*}(x|x^*)\,dx = 1$ requires the function $s(x^*)$ to be equal to 1 everywhere and this ambiguity is therefore avoided.

The potential presence of degenerate eigenvalues (issue 1b above), which introduces an ambiguity among the various possible linear combinations between the eigenfunctions associated with duplicate eigenvalues, can be avoided under the following, relatively weak, assumption.

Assumption 4 *For all $x_1^*, x_2^* \in \mathcal{X}^*$, the set $\left\{y : f_{y|x^*}(y|x_1^*) \neq f_{y|x^*}(y|x_2^*)\right\}$ has positive probability whenever $x_1^* \neq x_2^*$.*

[11] This nonuniqueness is even more severe than in the matrix diagonalization case. For matrices, it is sufficient to place the eigenvectors in the correct order. For operators, once the order of the eigenfunctions is set, it is still possible to parametrize them in multiple ways (e.g. index them by x^* or by $(x^*)^3$), as shown in Appendix A.

Remark: This assumption is weaker than the monotonicity assumptions typically made in the nonseparable error literature (e.g., Chernozhukov, Imbens, and Newey (2006), Matzkin (2003)), since the whole conditional distribution of y at different values of the regressors would have to agree perfectly in order for this condition to be violated. In particular, the presence of conditional heteroskedasticity can be sufficient in the absence of monotonicity. Assumption 4 circumvents the duplicate eigenvalues issue by simultaneously making use of more than one value of the dependent variable y. The idea is that the operator $L_{x|x^*}$ defining the eigenfunctions does not depend on y while the eigenvalues given by $f_{y|x^*}(y|x^*)$ do depend on y. Hence, if there is an eigenvalue degeneracy involving two eigenfunctions $f_{x|x^*}(\cdot|x_1^*)$ and $f_{x|x^*}(\cdot|x_2^*)$ for some value of y, we can look for another value of y that does not exhibit this problem to resolve the ambiguity. By piecing together the information regarding $f_{x|x^*}(x|x^*)$ obtained for different values of y it is possible to uniquely reconstruct $L_{x|x^*}$.

Formally, this can be shown as follows. Consider a given eigenfunction $f_{x|x^*}(\cdot|x^*)$ and let $D(y, x^*) = \left\{\tilde{x}^* : f_{y|x^*}(y|\tilde{x}^*) = f_{y|x^*}(y|x^*)\right\}$, the set of other values of x^* indexing eigenfunctions sharing the same eigenvalue. Any linear combination of functions $f_{x|x^*}(\cdot|\tilde{x}^*)$ for $\tilde{x}^* \in D(y, x^*)$ is a potential eigenfunction of $L_{y;x|z}L_{x|z}^{-1}$. However, if there exists a set Y such that $v(x^*) = \cap_{y\in Y}\text{span}\Big(\left\{f_{x|x^*}(\cdot|\tilde{x}^*)\right\}_{\tilde{x}^*\in D(y,x^*)}\Big)$ is one dimensional, then the set $v(x^*)$ will uniquely specify the eigenfunction $f_{x|x^*}(\cdot|x^*)$ (after normalization to integrate to 1). We now proceed by contradiction and show that if, for any possible choice of the set Y, $v(x^*)$ is never one dimensional, then Assumption 4 is violated. Indeed, if $v(x^*)$ has more than one dimension, it must contain at least two eigenfunctions, say $f_{x|x^*}(\cdot|x^*)$ and $f_{x|x^*}(\cdot|\tilde{x}^*)$. This implies that $\cap_{y\in Y}D(y, x^*)$ must at least contain the two points x^* and $\tilde{x}^*$. By the definition of $D(y, x^*)$, we must have that $f_{y|x^*}(y|x^*) = f_{y|x^*}(y|\tilde{x}^*)$ for all $y \in Y$. Since this would have to hold for any set Y, we have that $f_{y|x^*}(y|x^*) = f_{y|x^*}(y|\tilde{x}^*)$ almost everywhere under $f_y(y)$,[12] thus violating Assumption 4.

[12] Two densities can differ on a set of probability zero and still define the same probability measure.

Remark: In the special case of binary y, Assumption 4 amounts to a monotonicity assumption (e.g. $P[y=0|x^*]$ is strictly monotone in x^*). When x^* is multivariate, while the outcome variable is still binary, it will be necessary to define y to be a vector containing auxiliary variables in addition to the binary outcome, in order to allow for enough variation in the distribution of y conditional on x^* to satisfy Assumption 4. Each of these additional variables need not be part of the model of interest per se, but does need to be affected by x^* is some way. In that sense, such a variable is a type of "instrument", although it differs conceptually from conventional instruments, as it would typically be "caused by x^*" instead of "causing x^*". See Chalak and White (2006) for a discussion of this type of instrument.

Finally, we address issue 2, namely that the way one chooses to index the eigenvalues and eigenfunctions is not unique. Instead of indexing them by x^*, one could have chosen another variable $\tilde{x}^*$ related to x^* by some one-to-one piecewise differentiable function R, that is, $x^* = R(\tilde{x}^*)$. The kernels of the operators defining the eigenvalues and the eigenfunctions would then become $f_{y|x^*}(y|R(\tilde{x}^*))$ and $f_{x|x^*}(\cdot|R(\tilde{x}^*))$, respectively. This counterexample is fully developed in Appendix A. Fortunately, the issues of the uniqueness of the indexing of the eigenfunctions can be resolved with the following assumption.

Assumption 5 *There exists a known functional M such that $M\left[f_{x|x^*}(\cdot|x^*)\right] = x^*$ for all $x^* \in \mathcal{X}^*$.*

M is a very general functional that maps a univariate density to a real number (or a vector, if x^* is multivariate) and that defines some measure of location. Examples of M include, but are not limited to, the mean, the mode, or the τ quantile, corresponding to the following definitions of M, respectively,

$$\begin{aligned} M[f] &= \int x f(x) dx \\ M[f] &= \arg\max_x f(x) \qquad (12) \\ M[f] &= \inf\left\{x^* : \int 1(x \leq x^*) f(x) dx \geq \tau\right\}. \qquad (13) \end{aligned}$$

Assumption 5 resolves the ordering/indexing ambiguity because

$$M\left[f_{x|\tilde{x}^*}\left(\cdot|\tilde{x}^*\right)\right]=M\left[f_{x|x^*}\left(\cdot|R\left(\tilde{x}^*\right)\right)\right]=R\left(\tilde{x}^*\right),$$

which is only equal to $\tilde{x}^*$ if R is the identity function.

2.5 Summary

We now have all the ingredients needed to establish identification. Assumption 1 lets us obtain the integral Equation (1) relating the joint densities of the observable variables to the joint densities of the unobservable variables. This equation admits an equivalent operator representation (3). Under regularity conditions implying injectivity of some of the operators involved, the identification problem can be cast into the form of an operator diagonalization problem (Equation (8)), in which the operator to be diagonalized is defined in terms of observable densities, while the resulting eigenvalues and eigenfunctions provide the unobserved joint densities of interest. To ensure uniqueness of the eigenvalue-eigenfunction decomposition, we employ four techniques. First, a powerful result from spectral analysis (Theorem XV 4.5 in Dunford and Schwartz (1971)) guarantees a unique representation of an operator as a linear combination of projections, under a weak boundedness assumption. Second, the *a priori* arbitrary scale of the eigenfunctions is fixed by the requirement that densities must integrate to one. Third, to avoid any ambiguity in the definition of the eigenfunctions when degenerate eigenvalues are present, we use the fact that the eigenfunctions found must be consistent across different values of the dependent variable y. Finally, in order to uniquely determine the ordering of the eigenvalues and eigenfunctions, we invoke the assumption that some measure of location is left unaffected by the measurement error. These steps ensure that the diagonalization operation uniquely specifies the unobserved densities $f_{y|x^*}\left(y|x^*\right)$ and $f_{x|x^*}\left(x|x^*\right)$ of interest. We can also show that $f_{yx^*}\left(y,x^*\right)$ is identified by noting that, (i) by Equation (6), $f_{x^*|z}\left(x^*|z\right)$ is identified, (ii) $f_{x^*}\left(x^*\right)=\int f_{x^*|z}\left(x^*|z\right)f_z\left(z\right)dz$ where $f_z\left(z\right)$ is observed and that (iii) $f_{y,x^*}\left(y,x^*\right)=f_{y|x^*}\left(y|x^*\right)f_{x^*}\left(x^*\right)$. We can then summarize the results of this section in the following Theorem.

Theorem 1 *Under Assumptions 1-5, the knowledge of the conditional density $f_{yx|z}(y,x|z)$ uniquely determines $f_{y|x^*}(y|x^*)$, $f_{x|x^*}(x|x^*)$, and $f_{x^*|z}(x^*|z)$. Moreover, the knowledge of $f_{yxz}(y,x,z)$ uniquely determines $f_{yx^*}(y,x^*)$.*

While Theorem 1 establishes identification, we can also show that the model is actually overidentified, thus permitting a test of the model. Equation (1), upon which Theorem 1 is based, relates a function of 3 variables to a triplet of functions of 2 variables. Since the set of functions of 3 variables is much "larger" than the set of triplets of functions of 2 variables, there exist densities $f_{yx|z}(y,x|z)$ that cannot be generated by Equation (1), a telltale sign of an overidentifying restriction. The availability of more than one valid instrument offers further opportunities to test the model's assumptions, in particular, Assumption 5.

It is important to note that, although our proof of identification relies on the relatively abstract operation of finding an eigenvalue-eigenfunction decomposition of an operator, the estimation procedure need not parallel this approach. The diagonalization identity (8) in fact provides the same information as the initial Equation (1) and a valid estimation procedure can be based on solving Equation (1) for the unknown $f_{x|x^*}(x|x^*)\, f_{y|x^*}(y|x^*)$ and $f_{x^*|z}(x^*|z)$ under the constraints imposed by Assumption 5. Our proof is, however, essential to establish that this solution exists and is unique, thus justifying such a simplified estimation procedure.

3 Estimation using sieve maximum likelihood

3.1 Definitions

Having shown that all the conditional densities $f_{y|x^*}$, $f_{x|x^*}$, and $f_{x^*|z}$ are identified from the observed conditional density $f_{yx|z}(y,x|z)$, we now propose a sieve-based estimator (e.g. Grenander (1981), Shen (1997), Chen and Shen (1998), Ai and Chen (2003)) and derive its asymptotic properties. For simplicity, we consider y, x, x^*, and z to be scalars, although our treatment can easily be extended to multivariate settings. The support of all variables y, x^*, x, z is allowed to be unbounded, i.e., to be the whole real line.

Consider a latent model in the form of a conditional density as follows:

$$f_{y|x^*}(y|x^*;\theta_0). \tag{14}$$

The model also could be conditional on any number of other, correctly measured, variables, although this is not explicitly considered here, for simplicity. The model depends on a potentially infinite-dimensional parameter $\theta_0 \in \Theta = \mathcal{B} \times \mathcal{M}$, which is decomposed as $\left(b_0^T, \eta_0\right)^T$, where $b_0 \in \mathcal{B} \subset \mathbb{R}^{d_b}$ is the parameter vector of interest and $\eta_0 \in \mathcal{M}$ is a potentially infinite-dimensional nuisance parameter. Naturally, we assume $\left(b_0^T, \eta_0\right)^T$ are identified if $f_{y|x^*}$ is identified, i.e., that the parametrization (14) does not include redundant degrees of freedom. The sets $\mathcal{B}$ and $\mathcal{M}$ will be defined formally below.

This framework nests two main subcases of interest. First, setting $\theta_0 \equiv b_0^T$ covers the parametric likelihood case (which then becomes semiparametric once we account for measurement error). Second, models defined via moment restrictions $E\left[m\left(y, x^*, b\right)|x^*\right] = 0$ can be considered by defining instead a family of densities $f_{y|x^*}\left(y|x^*; b, \eta\right)$ such that $\int f_{y|x^*}\left(y|x^*; b, \eta\right) m\left(y, x^*, b\right) dy = 0$ for all b and η, which is clearly equivalent to imposing a moment condition. For example, in a nonlinear regression model $y = g\left(x^*, b\right) + \epsilon$ with $E\left(\epsilon|x^*\right) = 0$, we have $f_{y|x^*}\left(y|x^*; b, \eta\right) = f_{\epsilon|x^*}\left(y - g\left(x^*, b\right)|x^*\right)$. The infinite-dimensional nuisance parameter η is the conditional density $f_{\epsilon|x^*}\left(\cdot|\cdot\right)$, constrained to have zero mean. Another important example is the quantile regression case (where the conditional density $f_{\epsilon|x^*}\left(\cdot|\cdot\right)$ is constrained to have its conditional τ-quantile equal to 0). Quantile restrictions are useful, as they provide the fundamental concept enabling a natural treatment of nonseparable models (e.g. Chesher (2003)). More examples of a partition of θ into $\left(b^T, \eta\right)^T$ can be found in Shen (1997). In this paper, we consider η to be a function defined as $\eta\left(\cdot, \cdot\right) : \mathcal{U} \mapsto \mathbb{R}$ with $\mathcal{U} \subset \mathbb{R}^2$. Such a setup is reasonable because $f_{y|x^*}$ itself can be treated as an infinite-dimensional unknown parameter and $f_{y|x^*}$ was shown to be nonparametrically identified. Any user-specified $f_{y|x^*}\left(y|x^*; b, \eta\right)$ is a particular case of this fully nonparametric case.

Our sieve estimator is based on the following expression for the observed density (from

Equation (1))[13]

$$f_{yx|z}\left(y,x|z;\alpha_0\right)=\int_{\mathcal{X}^*}f_{y|x^*}(y|x^*;\theta_0)f_{x|x^*}(x|x^*)f_{x^*|z}(x^*|z)dx^*. \tag{15}$$

The unknown α_0 in the density function $f_{yx|z}$ includes θ_0 and density functions $f_{x|x^*}$ and $f_{x^*|z}$, i.e., $\alpha_0=\left(\theta_0,f_{x|x^*},f_{x^*|z}\right)^T$. The estimation procedure basically consists of replacing $f_{x|x^*}$, $f_{x^*|z}$ (and $f_{y|x^*}$ if it contains an infinite dimensional nuisance parameter η) by truncated series approximations and optimizing all parameters within a semiparametric maximum likelihood framework. The number of terms kept in the series approximations is allowed to grow with sample size at a controlled rate.

Our asymptotic analysis relies on standard smoothness restrictions (e.g. Ai and Chen (2003)) on the unknown functions η, $f_{x|x^*}$ and $f_{x^*|z}$. To describe them, let $\xi\in\mathcal{V}\subset\mathbb{R}^d$, $a=(a_1,\ldots,a_d)^T$, and

$$\nabla^a g(\xi)\equiv\frac{\partial^{a_1+\ldots+a_d}g(\xi)}{\partial\xi_1^{a_1}\ldots\partial\xi_d^{a_d}}$$

denote the $(a_1+\ldots+a_d)$-th derivative. Let $\|\cdot\|_E$ denote the Euclidean norm. Let $\underline{\gamma}$ denote the largest integer satisfying $\gamma>\underline{\gamma}$. The Hölder space $\Lambda^\gamma(\mathcal{V})$ of order $\gamma>0$ is a space of functions $g:\mathcal{V}\mapsto\mathbb{R}$ such that the first $\underline{\gamma}$ derivatives are bounded, and the $\underline{\gamma}$-th derivative are Hölder continuous with the exponent $\gamma-\underline{\gamma}\in(0,1]$, i.e.,

$$\max_{a_1+\ldots+a_d=\underline{\gamma}}\left|\nabla^a g(\xi)-\nabla^a g(\xi')\right|\leq c\left(\|\xi-\xi'\|_E\right)^{\gamma-\underline{\gamma}}$$

for all ξ, $\xi'\in\mathcal{V}$ and some constant c. The Hölder space becomes a Banach space with the Hölder norm as follows:

$$\|g\|_{\Lambda^\gamma}=\sup_{\xi\in\mathcal{V}}|g(\xi)|+\max_{a_1+\ldots+a_d=\underline{\gamma}}\sup_{\xi\neq\xi'\in\mathcal{V}}\frac{\left|\nabla^a g(\xi)-\nabla^a g(\xi')\right|}{\left(\|\xi-\xi'\|_E\right)^{\gamma-\underline{\gamma}}}.$$

To facilitate the treatment of functions defined on noncompact domains, we follow the technique suggested in Chen, Hong, and Tamer (2005), introducing a weighting function of the

[13] After multiplication by $f_z(z)$ on each side of Equation (15), one obtains an alternative expression, $f_{yxz}\left(y,x,z;\alpha_0\right)=\int_{\mathcal{X}^*}f_{yx^*}(y,x^*;\theta_0)f_{x|x^*}(x|x^*)f_{z|x^*}(z|x^*)dx^*$, which proves useful if the model specifies $f_{yx^*}\left(y,x^*\right)$ instead $f_{y|x^*}\left(y|x^*\right)$. The remainder of our treatment can be easily adapted to cover this case as well.

form $\omega(\xi) = \left(1 + \|\xi\|_E^2\right)^{-\varsigma/2}$, $\varsigma > \gamma > 0$ and defining a weighted Hölder norm as $\|g\|_{\Lambda^{\gamma,\omega}} \equiv \|\tilde{g}\|_{\Lambda^{\gamma}}$ for $\tilde{g}(\xi) \equiv g(\xi)\,\omega(\xi)$. The corresponding weighted Hölder space is denoted by $\Lambda^{\gamma,\omega}(\mathcal{V})$ while a weighted Hölder ball can be defined as $\Lambda_c^{\gamma,\omega}(\mathcal{V}) \equiv \{g \in \Lambda^{\gamma,\omega}(\mathcal{V}) : \|g\|_{\Lambda^{\gamma,\omega}} \leq c < \infty\}$.

We assume the functions η, $f_{x|x^*}$, and $f_{x^*|z}$ belong to the sets $\mathcal{M}$, $\mathcal{F}_1$, and $\mathcal{F}_2$ respectively, defined below.

Assumption 6 *$\eta \in \Lambda_c^{\gamma_1,\omega}(\mathcal{U})$ with $\gamma_1 > 1$;*[14]

Assumption 7 *$f_1 \in \Lambda_c^{\gamma_1,\omega}(\mathcal{X} \times \mathcal{X}^*)$ with $\gamma_1 > 1$ and $\int_{\mathcal{X}} f_1(x|x^*)dx = 1$ for all $x^* \in \mathcal{X}^*$;*

Assumption 8 *$f_2 \in \Lambda_c^{\gamma_1,\omega}(\mathcal{X}^* \times \mathcal{Z})$ with $\gamma_1 > 1$ and $\int_{\mathcal{X}^*} f_2(x^*|z)dx^* = 1$ for all $z \in \mathcal{Z}$.*

$$\begin{aligned}
\mathcal{M} &= \{\eta(\cdot,\cdot) : \textit{Assumption 6 holds.}\}, \\
\mathcal{F}_1 &= \{f_1(\cdot|\cdot) : \textit{Assumptions 2, 5, and 7 hold.}\}, \\
\mathcal{F}_2 &= \{f_2(\cdot|\cdot) : \textit{Assumptions 2, 8 hold.}\},
\end{aligned}$$

The condition $\|f\|_{\Lambda^{\gamma_1,\omega}} \leq c < \infty$ is necessary for the method of sieve, which we will use in the next step. In principle, one can solve for the true value $\alpha_0 = \left(\theta_0, f_{x|x^*}, f_{x^*|z}\right)^T$ as follows

$$\alpha_0 = \underset{\alpha=(\theta,f_1,f_2)^T \in \mathcal{A}}{\arg\max}\; E\left(\ln \int_{\mathcal{X}^*} f_{y|x^*}(y|x^*;\theta) f_1(x|x^*) f_2(x^*|z) dx^*\right), \tag{16}$$

where $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_2$ with $\Theta = \mathcal{B} \times \mathcal{M}$. Let $p_j^{k_n}(\cdot)$ be a sequence of known univariate basis functions, such as power series, splines, Fourier series, etc. To approximate functions of two variables, we use tensor-product linear sieve basis, denoted by $p^{k_n}(\cdot,\cdot) = (p_1^{k_n}(\cdot,\cdot), p_2^{k_n}(\cdot,\cdot)$, ..., $p_{k_n}^{k_n}(\cdot,\cdot))^T$. In the sieve approximation, we consider η, f_1 and f_2 in finite dimensional spaces $\mathcal{M}_n$, $\mathcal{F}_{1n}$ and $\mathcal{F}_{2n}$, where[15]

$$\begin{aligned}
\mathcal{M}_n &= \left\{\eta(\xi_1,\xi_2) = p^{k_n}(\xi_1,\xi_2)^T \delta \text{ for all } \delta \textit{ s.t. assumption 6 holds.}\right\} \\
\mathcal{F}_{1n} &= \left\{f(x|x^*) = p^{k_n}(x,x^*)^T \beta \text{ for all } \beta \textit{ s.t. assumptions 2, 5, and 7 hold.}\right\}, \\
\mathcal{F}_{2n} &= \left\{f(x^*|z) = p^{k_n}(x^*,z)^T \gamma \text{ for all } \gamma \textit{ s.t. assumptions 2, 8 hold.}\right\}.
\end{aligned}$$

[14] If η is a density function, certain restrictions should be added to assumption 6 analogous to those in assumptions 8 and 7.

[15] For simplicity, the notation $p^{k_n}(\cdot,\cdot)$ implicitly assumes that the sieve for η, $f(x|x^*)$ and $f(x^*|z)$ are the same, although this can be easily relaxed.

Therefore, we replace $\mathcal{M} \times \mathcal{F}_1 \times \mathcal{F}_2$ with $\mathcal{M}_n \times \mathcal{F}_{1n} \times \mathcal{F}_{2n}$ in the optimization problem, and then estimate α_0 by $\widehat{\alpha}_n$ as follows:

$$\widehat{\alpha}_n = \left(\widehat{\theta}_n, \widehat{f}_{1n}, \widehat{f}_{2n}\right)^T = \underset{\alpha=(\theta,f_1,f_2)^T \in \mathcal{A}_n}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \ln \int_{\mathcal{X}^*} f_{y|x^*}(y_i|x^*;\theta) f_1(x_i|x^*) f_2(x^*|z_i) dx^*. \quad (17)$$

where $\mathcal{A}_n = \Theta_n \times \mathcal{F}_{1n} \times \mathcal{F}_{2n}$ with $\Theta_n = \mathcal{B} \times \mathcal{M}_n$. In practice, the above integral can be conveniently carried out though either one of a number of numerical techniques, including Gaussian quadrature, Simpson's rules, Importance Sampling or Markov Chain Monte Carlo. In the sequel, we simply assume that this integral can be evaluated, for a given sample and a given truncated sieve, with a numerical accuracy that is far better than the statistical noise associated with the estimation procedure.

This setup is the same as in Shen (1997). We also use techniques described in Ai and Chen (2003) to state more primitive regularity conditions. In their paper, there are two sieve approximations: One is used to directly estimate the conditional mean as a function of the unknown parameter, the other is the sieve approximation of the infinite-dimensional parameter estimated through the maximization procedure. Our setup is, in some ways, simpler than in Ai and Chen (2003), because all the unknown parameters in α are estimated through a single-step semiparametric sieve MLE (Maximum Likelihood Estimator). Since our estimator takes the form of a semiparametric sieve estimator, the very general treatment of Shen (1997) and Chen and Shen (1998) can be used to establish asymptotic normality and root n consistency under a very wide variety of conditions, including dependent and nonidentically distributed data. However, for the purpose of simplicity and conciseness, this section provides specific sufficient regularity conditions for the i.i.d. case.

The restrictions in the definitions of $\mathcal{F}_{1n}$ and $\mathcal{F}_{2n}$ are easy to impose on a sieve estimator. We have the sieve expressions of f_1 and f_2 as follows:

$$f_1(x|x^*) = \sum_{i=0}^{i_n}\sum_{j=0}^{j_n} \beta_{ij} p_i(x - x^*) p_j(x^*), \quad f_2(x^*|z) = \sum_{i=0}^{i_n}\sum_{j=0}^{j_n} \gamma_{ij} p_i(x^* - z) p_j(z). \quad (18)$$

where $p_i(.)$ are user-specified basis functions. Define $k_n = (i_n + 1)(j_n + 1)$ and assume that i_n/j_n is bounded and bounded away from zero for all n. We also define the projection of the true value α_0 onto the space $\mathcal{A}_n$ associated with k_n:

$$\Pi_n \alpha \equiv \alpha_n \equiv \underset{\alpha_n=(\theta,f_1,f_2)^T \in \mathcal{A}_n}{\arg\max} E\left(\ln \int_{\mathcal{X}^*} f_{y|x^*}(y|x^*;\theta) f_1(x|x^*) f_2(x^*|z) dx^*\right).$$

and we let the smoothing parameter $k_n \to \infty$ as the sample size $n \to \infty$. The restriction $\int_{\mathcal{X}} f_1(x|x^*)dx = 1$ in the definition of $\mathcal{F}_{1n}$ implies $\sum_{j=0}^{j_n}\left(\sum_{i=0}^{i_n}\beta_{ij}\int_{\mathcal{E}} p_i(\varepsilon)d\varepsilon\right)p_j(x^*) = 1$ for all x^*, where $\varepsilon = x - x^*$. Suppose $p_0(.)$ is the only constant in $p_j(.)$. That equation implies that $\sum_{i=0}^{i_n}\beta_{i0}\int_{\mathcal{E}} p_i(\varepsilon)d\varepsilon = 1$ and $\sum_{i=0}^{i_n}\beta_{ij}\int_{\mathcal{E}} p_i(\varepsilon)d\varepsilon = 0$ for $j = 1, 2, ..., j_n$. Similar restrictions can be found for $\int_{\mathcal{X}^*} f_2(x^*|z)dx^* = 1$. Moreover, the identification assumption *5* also implies restrictions on the sieve coefficients. For example, consider the zero mode case. If the mode is unique and not at a boundary, we then have $\frac{\partial}{\partial x} f_{x|x^*}(x|x^*) = 0$ if and only if $x = x^*$. The restriction $\frac{\partial}{\partial x} f_{x|x^*}(x|x^*)\big|_{x=x^*} = 0$ in the definition of $\mathcal{F}_{1n}$ implies $\sum_{j=0}^{j_n}\left(\sum_{i=0}^{i_n}\beta_{ij}\frac{\partial p_i(0)}{\partial x}\right)q_j(x^*) = 0$. Since it must hold for all x^*, we have additional j_n constraints $\sum_{i=0}^{i_n}\beta_{ij}\frac{\partial p_i(0)}{\partial x} = 0$ for $j = 1, 2, ..., j_n$. Similar restrictions can be found for the zero mean and the zero median cases. In all three cases, the assumption *5* can be expressed as linear restrictions on β, which are easy to implement. See Appendix C for an explicit expression for the restrictions in the case where Fourier series are used in the sieve approximation.

3.2 Consistency

We use the results in Newey and Powell (2003) to show consistency of the sieve estimator. Define $D \equiv (y, x, z)$ for $y \in \mathcal{Y}$, $x \in \mathcal{X}$, and $z \in \mathcal{Z}$. The random variables x, y and z can have unbounded support $\mathbb{R}$. Following Ai and Chen (2003), we first show consistency under a strong norm $\|\cdot\|_s$ as a stepping stone to establishing a convergence rate under a suitably constructed weaker norm. Let

$$\|\alpha\|_s = \|b\|_E + \|\eta\|_{\infty,\omega} + \|f_1\|_{\infty,\omega} + \|f_2\|_{\infty,\omega}$$

where $\|g\|_{\infty,\omega} \equiv \sup_\xi |g(\xi)\omega(\xi)|$ with $\omega(\xi) = \left(1 + \|\xi\|_E^2\right)^{-\varsigma/2}$, $\varsigma > \gamma_1 > 0$. We make the following assumptions:

Assumption 9 *i) the data* $\{(Y_i, X_i, Z_i)_{i=1}^n\}$ *are i.i.d.; ii) the density of* $D \equiv (y, x, z)$, f_D, *satisfies* $\int \omega(D)^{-2} f_D(D) dD < \infty$.

Assumption 10 *i)* $b_0 \in \mathcal{B}$, *a compact subset of* $\mathbb{R}^b$. *ii) assumptions 6-8 hold for* (b, η, f_1, f_2) *in a neighborhood of* α_0 *(in the norm* $\|\cdot\|_s$*).*

Assumption 11 *i)* $E\left[\left(\ln f_{yx|z}(D)\right)^2\right]$ *is bounded; ii) there exists a measurable function* $h_1(D)$ *with* $E\left\{(h_1(D))^2\right\} < \infty$ *such that, for any* $\overline{\alpha} = (\overline{\theta}, \overline{f}_1, \overline{f}_2)^T \in \mathcal{A}$,

$$\left| \frac{f^{|1|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})}{f_{yx|z}(D, \overline{\alpha})} \right| \leq h_1(D),$$

where $f^{|1|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})$ *is defined as* $\frac{d}{dt} f_{yx|z}(D; \overline{\alpha} + t\bar{\omega})\big|_{t=0}$ *with each linear term, i.e.,* $\frac{d}{d\theta} f_{y|x^*}$, $\overline{f}_1$, *and* $\overline{f}_2$, *replaced by its absolute value, and* $\overline{\omega}(\xi, x, x^*, z) = [1,\ \omega^{-1}(\xi),\ \omega^{-1}((x, x^*)^T),\ \omega^{-1}((x^*, z)^T)]^T$ *with* $\xi \in \mathcal{U}$. *(The explicit expression of* $f^{|1|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})$ *can be found in equation 47 in the proof.)*

Assumption 12 $\|\Pi_n \alpha_0 - \alpha_0\|_s = o(1)$ *(as* $k_n \to \infty$*) and* $k_n/n \to 0$.

Assumption 9 is commonly used in cross-sectional analyses. Assumption 9(ii) is a typical condition on the tail behavior on the density, analogous to Assumption 3.2 in Chen, Hong, and Tamer (2005). Assumption 10 imposes restrictions on the parameter space. Detailed discussions on this assumption can be found in Gallant and Nychka (1987). Assumption 11 imposes an envelope condition on the first derivative of the log likelihood function, and guarantees a Hölder continuity property for the log likelihood. Assumption 12 states that the sieve can approximate the true α_0 arbitrarily well, in order the control the bias, while ensuring that the number of terms in the sieve grows slower than the sample size, in order to control the variance. We show consistency in the following Lemma.

Lemma 2 *Under assumptions 1-5 and 9-12, we have* $\|\widehat{\alpha}_n - \alpha_0\|_s = o_p(1)$.

Proof. See the appendix. ■

Consistency under the norm $\|\cdot\|_s$ is the first step needed to obtain the asymptotic properties of the estimator. In order to proceed towards our main semiparametric asymptotic normality and root n consistency result, we then need to establish convergence at the rate $o_p\left(n^{-1/4}\right)$ in a suitable norm. In order to achieve this convergence rate under relatively weak assumptions, we employ a device introduced by Ai and Chen (2003) and employ a weaker norm $\|\cdot\|$, under which $o_p\left(n^{-1/4}\right)$ convergence is easier to establish.

We now recall the concept of pathwise derivative, which is central to the asymptotics of sieve estimators. Consider $\alpha_1,\ \alpha_2 \in \mathcal{A}$, and assume the existence of a continuous path $\{\alpha(\tau) : \tau \in [0,1]\}$ in $\mathcal{A}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. If $\ln f_{yx|z}(D, (1-\tau)\alpha_0 + \tau\alpha)$ is continuously differentiable at $\tau = 0$ for almost all D and any $\alpha \in \mathcal{A}$, the pathwise derivative of $\ln f_{yx|z}(D, \alpha_0)$ at α_0 evaluated at $\alpha - \alpha_0$ can be defined as

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right] \equiv \left.\frac{d\ln f_{yx|z}(D,(1-\tau)\alpha_0+\tau\alpha)}{d\tau}\right|_{\tau=0} \tag{19}$$

almost everywhere (under the probability measure of D). The pathwise derivative is a linear functional that approximates $\ln f_{yx|z}(D, \alpha_0)$ in the neighborhood of α_0, i.e. for small values of $\alpha - \alpha_0$. Note that this functional can also be evaluated for other values of the argument. For instance, by linearity,

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha_1-\alpha_2\right] \equiv \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha_1-\alpha_0\right] - \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha_2-\alpha_0\right]. \tag{20}$$

In our setting, the pathwise derivative at α_0 is as follows (from Equation (15)):

$$\begin{aligned} & \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right] \qquad\qquad (21) \\ = \;& \frac{1}{f_{yx|z}(D,\alpha_0)}\left\{\int_{\mathcal{X}^*}\frac{d}{d\theta}f_{y|x^*}(y|x^*;\theta_0)\left[\theta-\theta_0\right]f_{x|x^*}(x|x^*)f_{x^*|z}(x^*|z)dx^* + \right. \\ & + \int_{\mathcal{X}^*} f_{y|x^*}(y|x^*;\theta_0)\left[f_1(x|x^*)-f_{x|x^*}(x|x^*)\right]f_{x^*|z}(x^*|z)dx^* + \\ & \left. + \int_{\mathcal{X}^*} f_{y|x^*}(y|x^*;\theta_0)f_{x|x^*}(x|x^*)\left[f_2(x^*|z)-f_{x^*|z}(x^*|z)\right]dx^*\right\}. \end{aligned}$$

Note that the denominator $f_{yx|z}(D, \alpha_0)$ is nonzero with probability 1. We use the Fisher norm $\|\cdot\|$ defined as

$$\|\alpha_1 - \alpha_2\| \equiv \sqrt{E\left\{\left(\frac{d \ln f_{yx|z}(D, \alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)^2\right\}} \tag{22}$$

for any α_1, $\alpha_2 \in \mathcal{A}$. In order to establish the asymptotic normality of $\widehat{b}_n$, one typically needs that $\widehat{\alpha}_n$ converges to α_0 at a rate faster than $n^{-1/4}$. We need the following assumptions to obtain this rate of convergence:

Assumption 13 $\|\Pi_n \alpha_0 - \alpha_0\| = O\left(k_n^{-\gamma_1/d_1}\right) = o\left(n^{-1/4}\right)$ *with* $d_1 = 2$ *and* $\gamma_1 > d_1$,[16] *for* γ_1 *as in Assumptions 6-8.*

Assumption 14 *i) there exists a measurable function* $c(D)$ *with* $E\left\{c(D)^4\right\} < \infty$ *such that* $\left|\ln f_{yx|z}(D; \alpha)\right| \leq c(D)$ *for all* D *and* $\alpha \in \mathcal{A}_n$*; ii)* $\ln f_{yx|z}(D; \alpha) \in \Lambda_c^{\gamma,\omega}(\mathcal{Y} \times \mathcal{X} \times \mathcal{Z})$ *for some constant* $c > 0$ *with* $\gamma > d_D/2$*, for all* $\alpha \in \mathcal{A}_n$*, where* d_D *is the dimension of* D*.*

Assumption 15 $\mathcal{A}$ *is convex in* α_0*, and* $f_{y|x^*}(y|x^*; \theta)$ *is pathwise differentiable at* θ_0*.*

Assumption 16 *For some* $c_1, c_2 > 0$*,*

$$c_1 E\left(\ln \frac{f_{yx|z}(D; \alpha_0)}{f_{yx|z}(D; \alpha)}\right) \leq \|\alpha - \alpha_0\|^2 \leq c_2 E\left(\ln \frac{f_{yx|z}(D; \alpha_0)}{f_{yx|z}(D; \alpha)}\right). \tag{23}$$

holds for all $\alpha \in \mathcal{A}_n$ *with* $\|\alpha - \alpha_0\|_s = o(1)$*.*

Assumption 17 $\left(k_n n^{-1/2} \ln n\right) \sup_{(\xi_1,\xi_2) \in (\mathcal{U} \cup (\mathcal{X} \times \mathcal{X}^*) \cup (\mathcal{X}^* \times \mathcal{Z}))} \left\|p^{k_n}(\xi_1, \xi_2)\right\|_E^2 = o(1)$*.*

Assumption 18 $\ln N(\varepsilon, \mathcal{A}_n) = O(k_n \ln(k_n/\varepsilon))$ *where* $N(\varepsilon, \mathcal{A}_n)$ *is the minimum number of balls (in the* $\|\cdot\|_s$ *norm) needed to cover the set* $\mathcal{A}_n$*.*

Assumption 13 controls the approximation error of $\Pi_n \alpha_0$ to α_0 and the selection of k_n. It is usually satisfied by using sieve functions such as power series, Fourier series, etc. (see

[16] In general, $d_1 = \max\left\{\dim(\mathcal{U}), \dim(\mathcal{X} \times \mathcal{X}^*), \dim(\mathcal{X}^* \times \mathcal{Z})\right\}$.

Newey (1995) and Newey (1997) for more discussion.) Assumption 14 imposes an envelope condition and a smoothness condition on the log likelihood function. Assumption 15 implies that the norm $\|\cdot\|$ is well-defined. Define $K(\alpha,\alpha_0)=E\left(\ln\frac{f_{yx|z}(D;\alpha_0)}{f_{yx|z}(D;\alpha)}\right)$, which is the Kullback-Leibler discrepancy. Assumption 16 implies that $\|\cdot\|$ is a norm equivalent to the $(K(\cdot,\cdot))^{1/2}$ discrepancy on $\mathcal{A}_n$. Under the norm $\|\cdot\|$, the sieve estimator can be shown to converge at the requisite rate $o_p(n^{-1/4})$.

Theorem 3 *Under assumptions 1-5 and 9-18, we have* $\|\widehat{\alpha}_n-\alpha_0\|=o_p(n^{-1/4})$.

Proof. See the appendix. ■

It may appear surprising at first that such a fast convergence rate could be obtained in a nonparametric estimation problem that includes, as a special case, models traditionally handled through deconvolution approaches and that are known to be prone to slow convergence issues (e.g. Fan (1991)). These issues can be circumvented, thanks to the fact that the Fisher norm downweighs each dimension of the estimation error $\hat{\alpha}-\alpha_0$ according to its own standard error. In other words, more error is tolerated along the dimensions that are more difficult to estimate. Assumption 16 does impose a limit on how weak the Fisher norm can be, however. In the limit where the Fisher norm becomes singular (i.e. completely insensitive to some dimensions of α), the local quadratic behavior of the objective function is lost and Assumption 16 no longer holds.

Thanks to the Fisher norm's downweighting property, as the number of terms in the sieve increases, each new degree of freedom that gets included in the estimation problem does not appear increasingly difficult to estimate. A relatively fast convergence in the Fisher norm is therefore possible and does not conflict with slower convergence obtained in some other norm. Naturally, for the same reason, convergence in the Fisher norm is not a very useful concept for the sole purpose of establishing a nonparametric convergence result. In nonparametric settings, convergence in some well-understood L_p norm would be a more useful result. However, our ultimate goal is to establish the asymptotics for some parametric component of our semiparametric model. In that context, the Fisher norm is a very useful device that was employed in Ai and Chen (2003) and that guarantees the important intermediate results of $o_p\left(n^{-1/4}\right)$ convergence under rather weak conditions.

3.3 Asymptotic Normality

We follow the semiparametric MLE framework of Shen (1997) to show the asymptotic normality of the estimator $\widehat{b}_n$. We define the inner product

$$\langle v_1, v_2\rangle = E\left\{\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[v_1]\right)\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[v_2]\right)\right\}. \tag{24}$$

Obviously, the weak norm $\|\cdot\|$ defined in Equation (22) can be induced by this inner product. Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A}-\{\alpha_0\}$ under the norm $\|\cdot\|$ (i.e., $\overline{\mathbf{V}} = \mathbb{R}^{d_b}\times\overline{\mathcal{W}}$ with $\overline{\mathcal{W}} \equiv \overline{\mathcal{M}\times\mathcal{F}_1\times\mathcal{F}_2} - \left\{\left(\eta_0, f_{x|x^*}, f_{x^*|z}\right)^T\right\}$) and define the Hilbert space $\left(\overline{\mathbf{V}}, \langle\cdot,\cdot\rangle\right)$ with its inner product defined in Equation (24).

As shown above, we have

$$\begin{aligned}\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[\alpha-\alpha_0] &= \frac{d\ln f_{yx|z}(D,\alpha_0)}{db}[b-b_0] + \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[\eta-\eta_0] + \\ &\quad + \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[f_1 - f_{x|x^*}\right] + \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[f_2 - f_{x^*|z}\right].\end{aligned} \tag{25}$$

For each component b_j of b, $j = 1, 2, ..., d_b$, we define $w_j^* \in \overline{\mathcal{W}}$ as follows:

$$\begin{aligned} w_j^* &\equiv \left(\eta_j^*, f_{1j}^*, f_{2j}^*\right)^T \\ &= \underset{(\eta,f_1,f_2)^T\in\overline{\mathcal{W}}}{\arg\min} E\left\{\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{db_j} - \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[\eta] + \right.\right. \\ &\quad \left.\left. - \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}[f_1] - \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}[f_2]\right)^2\right\}.\end{aligned} \tag{26}$$

Define $w^* = \left(w_1^*, w_2^*, ..., w_{d_b}^*\right)$,

$$\begin{aligned}\frac{d\ln f_{yx|z}(D,\alpha_0)}{df}\left[w_j^*\right] &= \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}\left[\eta_j^*\right] + \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[f_{1j}^*\right] + \\ &\quad + \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[f_{2j}^*\right], \\ \frac{d\ln f_{yx|z}(D,\alpha_0)}{df}[w^*] &= \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df}\left[w_1^*\right], ..., \frac{d\ln f_{yx|z}(D,\alpha_0)}{df}\left[w_{d_b}^*\right]\right),\end{aligned} \tag{27}$$

and the row vector

$$G_{w^*}(D) = \frac{d\ln f_{yx|z}(D,\alpha_0)}{db^T} - \frac{d\ln f_{yx|z}(D,\alpha_0)}{df}\left[w^*\right]. \tag{28}$$

We want to show that $\widehat{b}_n$ has a multivariate normal distribution asymptotically. It is well known that if $\lambda^T b$ has a normal distribution for all λ, then b has a multivariate normal distribution. Therefore, we consider $\lambda^T b$ as a functional of α. Define $s(\alpha) \equiv \lambda^T b$ for $\lambda \in \mathbb{R}^{d_b}$ and $\lambda \neq 0$. If $E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$ is finite positive definite, then the function $s(\alpha)$ is bounded, and the Riesz representation theorem implies that there exists a representor v^* such that

$$s(\alpha) - s(\alpha_0) \equiv \lambda^T\left(b - b_0\right) = \langle v^*, \alpha - \alpha_0\rangle \tag{29}$$

for all $\alpha \in \mathcal{A}$. Here, $v^* \equiv \binom{v_b^*}{v_f^*}$, $v_b^* = J^{-1}\lambda$, $v_f^* = -w^* v_b^*$, with $J = E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$. Under suitable assumptions made below, the Riesz representor v^* exists and is bounded.

As mentioned in Begun, Hall, Huang, and Wellner (1983), v_f^* corresponds to a worst possible direction of approach to $\left(\eta_0, f_{x|x^*}, f_{x^*|z}\right)$ for the problem of estimating b_0. In the language of Stein (1956), v_f^* yields the most difficult one-dimensional sub-problem. Equation (29) implies that it is sufficient to find the asymptotic distribution of $\langle v^*, \widehat{\alpha}_n - \alpha_0\rangle$ to obtain that of $\lambda^T\left(\widehat{b}_n - b_0\right)$ under suitable conditions. We denote

$$\frac{d\ln f_{yx|z}(D,\alpha)}{d\alpha}\left[v\right] \equiv \left.\frac{d\ln f_{yx|z}(D,\alpha+\tau v)}{d\tau}\right|_{\tau=0} \quad \text{a.s. } D \text{ for any } v \in \overline{\mathbf{V}}. \tag{30}$$

For a sieve MLE, we have that

$$\langle v^*, \widehat{\alpha}_n - \alpha_0\rangle = \frac{1}{n}\sum_{i=1}^{n}\frac{d\ln f_{yx|z}(D_i,\alpha_0)}{d\alpha}\left[v^*\right] + o_p\left(n^{-1/2}\right) \tag{31}$$

Note that $\left(\frac{d\ln f_{yx|z}(D,\alpha)}{d\alpha}\left[v^*\right]\right) = G_{w^*}(D)J^{-1}\lambda$. Thus, by the classical central limit theorem, the asymptotic distribution of $\sqrt{n}\left(\widehat{b}_n - b_0\right)$ is $N\left(0, J^{-1}\right)$. In fact, the matrix J is the efficient information matrix in this semiparametric estimation, under suitable regularity conditions given in Shen (1997).

We now present the sufficient conditions for the $\sqrt{n}-$normality of $\widehat{b}_n$. Define

$$\mathcal{N}_{0n} \equiv \left\{\alpha \in \mathcal{A}_n : \left\|\alpha - \alpha_0\right\|_s \leq \upsilon_n,\ \ \left\|\alpha - \alpha_0\right\| \leq \upsilon_n n^{-1/4})\right\} \tag{32}$$

with $\upsilon_n = o(1)$ and $\mathcal{N}_0$ the same way with $\mathcal{A}_n$ replaced by $\mathcal{A}$. Note that $\mathcal{N}_0$ still depends on n. For $\alpha \in \mathcal{N}_{0n}$ we define a local alternative $\alpha^*(\alpha, \varepsilon_n) = (1-\varepsilon_n)\alpha + \varepsilon_n (v^* + \alpha_0)$ with $\varepsilon_n = o\left(n^{-1/2}\right)$. Let $\Pi_n \alpha^*(\alpha, \varepsilon_n)$ be the projection of $\alpha^*(\alpha, \varepsilon_n)$ onto $\mathcal{A}_n$.

Assumption 19 *i) $E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$ exists, is bounded and is positive-definite; ii) $b_0 \in$ $int(\mathcal{B})$.*

Assumption 20 *There exists a measurable function $h_2(D)$ with $E\{h_2(D))^2\} < \infty$ such that, for any $\overline{\alpha} = (\overline{\theta}, \overline{f}_1, \overline{f}_2)^T \in \mathcal{N}_0$,*

$$\left|\frac{f^{|1|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})}{f_{yx|z}(D, \overline{\alpha})}\right|^2 + \left|\frac{f^{|2|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})}{f_{yx|z}(D, \overline{\alpha})}\right| < h_2(D), \tag{33}$$

where $f^{|2|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})$ is defined as $\left.\frac{d^2}{dt^2} f_{yx|z}(D; \overline{\alpha} + t\bar{\omega})\right|_{t=0}$ with each linear term, i.e., $\frac{d}{d\theta} f_{y|x^*}$, $\frac{d^2}{d\theta^2} f_{y|x^*}$, $\overline{f}_1$, and $\overline{f}_2$, replaced by its absolute value. (The explicit expression of $f^{|2|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})$ can be found in equation 63 in the proof.)

We introduce the following notations for the next assumption: for $\widetilde{f} = \eta$, f_1, or f_2,

$$\frac{d\ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}\left[p^{k_n}\right] = \left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}\left[p_1^{k_n}\right], \frac{d\ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}\left[p_2^{k_n}\right], ..., \frac{d\ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}\left[p_{k_n}^{k_n}\right]\right)^T, \tag{34}$$

$$\frac{d\ln f_{yx|z}(D, \alpha_0)}{db} = \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{db_1}, \quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{db_2}, \quad ..., \quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{db_{d_b}}\right)^T,$$

$$\begin{aligned}\frac{d\ln f_{yx|z}(D, \alpha_0)}{d\alpha}\left[p^{k_n}\right] \quad = \quad & \left(\left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{db}\right)^T, \left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{d\eta}\left[p^{k_n}\right]\right)^T,\right. \\ & \left., \left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{df_1}\left[p^{k_n}\right]\right)^T, \left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{df_2}\left[p^{k_n}\right]\right)^T\right)^T,\end{aligned}$$

and

$$\Omega_{k_n} = E\left\{\left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{d\alpha}\left[p^{k_n}\right]\right)\left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{d\alpha}\left[p^{k_n}\right]\right)^T\right\}.$$

Assumption 21 *The smallest eigenvalue of the matrix Ω_{k_n} is bounded away from zero, and $\left\|p_j^{k_n}\right\|_{\infty,\omega} < \infty$ for $j = 1, 2, ..., k_n$ uniformly in k_n.*

Assumption 22 *There is a* $v_n^* = \binom{v_b^*}{-(\Pi_n w^*)v_b^*} \in \mathcal{A}_n - \{\Pi_n \alpha_0\}$ *such that* $\|v_n^* - v^*\| = o(n^{-1/4})$.

Assumption 23 *For all* $\alpha \in \mathcal{N}_{0n}$*, there exists a measurable function* $h_4\left(D\right)$ *with* $E\left|h_4\left(D\right)\right| < \infty$ *such that*

$$\left|\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t\left(\alpha-\alpha_0\right))\right|_{t=0} \leq h_4(D)\left\|\alpha-\alpha_0\right\|_s^4. \tag{35}$$

Assumption 19 is essential to obtain root n consistency since it ensures that the asymptotic variance exists and that b_0 is an "interior" solution. Assumption 20 imposes an envelope condition on the second derivative of the log likelihood function. This condition is related to the stochastic equicontinuity condition A in Shen (1997). The condition guarantees the linear approximation of the likelihood function by its derivative near α_0. That condition can be replaced by a stronger condition that $f_{yx|z}(D,\alpha)$ is differentiable in quadratic mean. Assumption 21 is similar to Assumption 2 in Newey (1997). Intuitively, Assumptions 21 and 23 are used to characterizes the local quadratic behavior of the criterion difference, i.e., condition B in Shen (1997) and can be simplified to: For all $\alpha \in \mathcal{N}_{0n}$,

$$E\left(\ln\frac{f_{yx|z}(D,\alpha_0)}{f_{yx|z}(D,\alpha)}\right) = \frac{1}{2}\left\|\alpha-\alpha_0\right\|^2\left(1+o(1)\right). \tag{36}$$

Assumption 22 states that the representor can be approximated by the sieve with an asymptotically negligible error, which is an important necessary condition for the asymptotic bias of the sieve estimator itself to be asymptotically negligible. A detailed discussion of these assumptions can be found in Shen (1997) and Chen and Shen (1998). By theorem 1 in Shen (1997), we show that the estimator for the parametric component b_0 is $\sqrt{n}$ consistent and asymptotically normally distributed.

Theorem 4 *Under assumptions 1-5, 9-16 and 19-23,* $\sqrt{n}\left(\widehat{b}_n - b_0\right) \overset{d}{\to} N\left(0, J^{-1}\right)$ *where* $J = E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$ *for* $G_{w^*}(D)$ *given in Equation (28).*

Proof. See the appendix. ■

Achieving the level of generality provided by Theorem 4 forces us to state some of our regularity conditions is a relatively high-level form, as is often done in the sieve estimation literature (e.g. Ai and Chen (2003), Shen (1997), Chen and Shen (1998)). However, once the type of sieve and the particular form of $f_{y|x^*}(y|x^*;\theta)$ are specified, more primitive assumptions can be formulated, using some of the techniques found in Blundell, Chen, and Kristensen (2003), for instance.

It is known that obtaining a root n consistency and asymptotic normality result for a semiparametric estimator in the context of classical errors-in-variable models demands a balance between the smoothness of the measurement error and of the densities (or regression functions) of interest (e.g. Taupin (1998), Schennach (2004a)). Our treatment, when specialized to classical measurement errors, does not evade this requirement. When the measurement error densities are "too smooth" and the functions of interest are "not smooth enough" to guarantee root n consistency and asymptotic normality, this will manifests itself as a violation of one of our assumptions. If the failure is first-order, i.e. due to the inexistence of an influence function with bounded variance, then a bounded Riesz representor v^* will fail to exist and Assumptions 19 and 22 will not hold. If the failure is of a "higher-order" nature, i.e. when nonlinear remainder terms in the estimator's stochastic expansion are not negligible, then either one of Assumptions 20, 21 or 23 will not hold. Intuitively, this represents a case where the local quadratic behavior of the objective function is lost.

4 Simulations

This section considers the performance of the proposed estimator with simulated data. For simplicity, we set $\theta_0 \equiv b_0$ and consider a parametric probit model

$$f_{y|x^*}(y|x^*) = [\Phi(a+bx^*)]^y \left[1-\Phi(a+bx^*)\right]^{1-y} \tag{37}$$

where (a,b) is the unknown parameter vector and $\Phi(.)$ is the standard normal cdf. In the simulation, we generate the latent variable and instrumental variable as follows: $z \sim N(1,(0.7)^2)$ and $x^* = z + 0.3(e-z)$ with an independent $e \sim N(1,(0.7)^2)$. The distribution of both z and η are truncated on $[0,2]$, for simplicity in the implementation. The conditional density of the measurement error $\varepsilon \equiv x - x^*$ can be written as $f_{\varepsilon|x^*}(\varepsilon|x^*) = f_{x|x^*}(x^* + \varepsilon|x^*)$, which depends on x^*. As shown before, the identification conditions imposed on $f_{x|x^*}$ may allow for correlations between the measurement error and the true value in a very general way. We give five examples below. In the simulation of each example, there are 2000 observations with 1000 repetitions. A Fourier series is used, where each term is of the form $\cos(k\pi\varepsilon/l)$ or $\sin(k\pi\varepsilon/l)$ with $l = 2$. We consider three estimators. First, the model is estimated with the measurement error ignored. This estimator is expected to be inconsistent. Second, we estimate the model using the accurate, measurement error-free data. This estimator is just the standard MLE of the probit model. It should be consistent and efficient but, of course, infeasible since the data is actually measured with error. The third estimator is the proposed sieve MLE, which is consistent and feasible in the presence of measurement error. It should have a larger variance than the second estimator, but a much smaller bias than the first estimator. For each estimator, we present the mean, the standard deviation (std. dev.), and the square root of the mean squared error (RMSE). We are now ready to present the performance of the estimator with five examples.

Example I (a heteroskedastic error with zero mean): Consider a measurement error as follows:

$$x = x^* + \sigma(x^*)\nu \tag{38}$$

with $x^* \perp \nu$, $E(\nu) = 0$, and $\sigma(.) > 0$ being an unknown non-stochastic function. These assumptions can also be written as $E(x - x^*|x^*) = 0$, i.e., the measurement error is conditional mean independent of the true value. The identification condition is also satisfied because it can verified that $x^* = \int x f_{x|x^*}(x|x^*)dx$. The error structure in the simulation is $F_\nu(\nu) = \Phi(\nu)$ with $\sigma(x^*) = 0.5\exp(-x^*)$. The simulation results are in Table 1.

Example II (a heteroskedastic error with mean shift): In this example, we relax the assumption that $E(\nu) = 0$ in (38) so that the measurement error may have a systematic mean shift. We can decompose the observed x as follows:

$$x = x^* + \mu_\nu\sigma(x^*) + \sigma(x^*)(\nu - \mu_\nu) \tag{39}$$

Table 1: Simulation results, a heteroskedastic error with zero mean (n=2000, reps=1000)

	$a=-1$			$b=1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
ignoring meas. error	-0.7601	0.0759	0.2516	0.7601	0.0686	0.2495
accurate data	-0.9974	0.0823	0.0824	0.9989	0.0766	0.0766
Sieve MLE	-0.9556	0.1831	0.1884	0.9087	0.1315	0.1601
smoothing parameters: $i_n = 6, j_n = 6$ in f_1; $i_n = 3, j_n = 2$ in f_2;						

where $\mu_\nu = E(\nu)$ is unknown. The first term is the true value x^*. The second term is the systematic x^*-dependent mean shift of the error. The third term is a heteroskedastic error with zero mean. Because $x^* \perp \nu$, we have $f_{x|x^*}(x|x^*) = \frac{1}{\sigma(x^*)} f_\nu\left(\frac{x-x^*}{\sigma(x^*)}\right)$, where f_ν is the density function of ν. In this setup, the identification restrictions on $f_{x|x^*}(x|x^*)$ can be straightforwardly converted into restrictions on f_ν.

We first consider the zero mode case. The zero mode condition on $f_{x|x^*}$ holds if the density f_ν has its unique mode at zero. In the simulation, we let $f_\nu(\nu) = \exp[\nu - \exp(\nu)]$ with $\sigma(x^*) = 0.5\exp(-x^*)$. The simulation results are in Table 2.

Second, we consider the zero median case, in which the median of the distribution of ν is zero and the density $f_{x|x^*}$ has median at x^*. In the simulation, we let the cdf of ν be

$$F_\nu(\nu) = \frac{1}{\pi}\arctan\left[\frac{1}{2} + \frac{1}{2}\exp(\nu) - \exp(-\nu)\right] + \frac{1}{2} \qquad (40)$$

with $\sigma(x^*) = 0.5\exp(-x^*)$. Note that this distribution is not symmetric around the median zero. The simulation results are in Table 3.

Table 2: Simulation results, a heteroskedastic error with zero mode (n=2000, reps=1000)

	$a=-1$			$b=1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
ignoring meas. error	-0.5676	0.0649	0.4372	0.6404	0.0632	0.3651
accurate data	-1.0010	0.0813	0.0813	1.0030	0.0761	0.0761
Sieve MLE	-0.9575	0.2208	0.2249	0.9825	0.1586	0.1596
smoothing parameters: $i_n = 6, j_n = 3$ in f_1; $i_n = 3, j_n = 2$ in f_2;						

Example III (a nonadditive error with zero mode): An error equation like (38) is usually set up for convenience. The additive structure (38) with $x^* \perp \nu$ may not always be appropriate

Table 3: Simulation results, a heteroskedastic error with zero median (n=2000, reps=1000)

	$a=-1$			$b=1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
ignoring meas. error	-0.6514	0.0714	0.3559	0.6375	0.0629	0.3679
accurate data	-1.0020	0.0796	0.0796	1.0020	0.0747	0.0748
Sieve MLE	-0.9561	0.2982	0.3014	0.9196	0.2734	0.2850

smoothing parameters: $i_n = 8, j_n = 8$ in f_1; $i_n = 3, j_n = 2$ in f_2;

in applications. Therefore, we now consider a nonseparable example, where it is more natural to specify $f_{x|x^*}(x|x^*)$ directly for the purpose of generating the simulated data. Let

$$\begin{aligned} f_{x|x^*}(x|x^*) &= \frac{g(x,x^*)}{\int_{-\infty}^{\infty} g(x,x^*)dx} \\ g(x,x^*) &= \exp\left\{h(x^*)\left[\left(\frac{x-x^*}{\sigma(x^*)}\right) - \exp\left(\frac{x-x^*}{\sigma(x^*)}\right)\right]\right\} \end{aligned} \tag{41}$$

It is easy to show that $f_{x|x^*}$ has the unique mode at x^* for any $h(x^*) > 0$. Thus the model is identified with this error structure. When $h(x^*) = 1$, this density becomes the density generated by equation (38) with ν having an extreme value distribution. Furthermore, the fact that identification holds for a general $h(x^*)$ means the independence assumption $x^* \perp \nu$ in (38) is not necessary. We can deal with more general measurement error using the estimator in this paper. In the simulation, we use $\sigma(x^*) = 0.5\exp(-x^*)$ and $h(x^*) = \exp(-0.1x^*)$. The simulation results are in Table 4.

Table 4: Simulation results, nonadditive error with zero mode (n=2000, reps=1000)

	$a=-1$			$b=1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
ignoring meas. error	-0.5167	0.0611	0.4871	0.5834	0.0590	0.4208
accurate data	-1.0010	0.0813	0.0813	1.0030	0.0761	0.0761
Sieve MLE	-0.9232	0.2010	0.2152	0.9430	0.1440	0.1549

smoothing parameters: $i_n = 7, j_n = 3$ in f_1; $i_n = 3, j_n = 2$ in f_2;

Example IV (a nonadditive error with zero median): Similar to example III, we consider a nonadditive error with zero median. We let the cdf corresponding to $f_{x|x^*}$ be

$$F_{x|x^*}(x|x^*) = \frac{1}{\pi}\arctan\left\{h(x^*)\left[\frac{1}{2} + \frac{1}{2}\exp\left(\frac{x-x^*}{\sigma(x^*)}\right) - \exp\left(-\frac{x-x^*}{\sigma(x^*)}\right)\right]\right\} + \frac{1}{2} \tag{42}$$

with $h(x^*) > 0$. Note $F_{x|x^*}(x^*|x^*) = \frac{1}{2}$ for any $h(x^*)$. Moreover, this distribution is not symmetric around x^*, and x^* is not the mode either. When $h(x^*) = 1$, the error structure is the same as in (38). In the simulation, we use $\sigma(x^*) = 0.5\exp(-x^*)$ and $h(x^*) = \exp(-0.1x^*)$. The simulation results are in Table 5.

Table 5: Simulation results, nonadditive error with zero median (n=2000, reps=1000)

	$a = -1$			$b = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
ignoring meas. error	-0.6351	0.0734	0.3722	0.6219	0.0647	0.3836
accurate data	-1.0010	0.0802	0.0802	1.0020	0.0752	0.0753
Sieve MLE	-0.9741	0.2803	0.2815	0.9342	0.2567	0.2650

smoothing parameters: $i_n = 8, j_n = 8$ in f_1; $i_n = 3, j_n = 2$ in f_2;

Example V (an always-underreporting error): In some applications, it is reasonable to assume that respondents always underreport, i.e., $x \leq x^*$. In other words, x^* is the 100-th percentile of $f_{x|x^*}(x|x^*)$. We have shown that the model is also identified in this case. In the simulation, we consider

$$f_{x|x^*}(x|x^*) = \frac{1}{\sigma(x^*)} \exp\left(\frac{x - x^*}{\sigma(x^*)}\right) I(x \leq x^*) \tag{43}$$

where $I(.)$ is an indicator function and $\sigma(x^*) = 0.5\exp(-x^*)$. The simulation results are in Table 6.

Table 6: Simulation results, an always-underreporting error (n=2000, reps=1000)

	$a = -1$			$b = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
ignoring meas. error	-0.5562	0.0601	0.4478	0.693	0.0632	0.3134
accurate data	-1.0010	0.0813	0.0813	1.003	0.0761	0.0761
Sieve MLE	-0.9230	0.2389	0.2510	1.071	0.2324	0.2429

smoothing parameters: $i_n = 4, j_n = 6$ in f_1; $i_n = 3, j_n = 2$ in f_2;

The simulation results in Table 1-6 show that out proposed estimator performs well under different identification conditions. The sieve estimator has a smaller bias than the first estimator, which ignores the measurement error. As expected, the sieve estimator has a larger variance than the other two estimators in all the examples. This is due to the nonparametric estimation of the infinite dimensional functions. However, the overall root mean square error (RMSE) for the sieve estimator dominates the RMSE of the other two estimators.

5 Empirical Illustration

The section illustrates the usefulness of our sieve estimator with actual empirical data. We are interested in the impact of earnings on the probability of divorcing. Let y_i be a dichotomous variable equal to 0 if individual i is married and equal to 1 if that individual is divorced or separated. We thus use a probit model as follows

$$f(y_i|x_i^*) = [\Phi(a + bx_i^*)]^{y_i} [1 - \Phi(a + bx_i^*)]^{1-y_i}, \tag{44}$$

where x_i^* is individual i's personal earnings. Since it is widely recognized that earnings, denoted, x_i^* is subject to measurement error that may be nonclassical in nature (e.g. Bollinger (1998), Bound and Krueger (1991), Chen, Hong, and Tamer (2005)), this represents a natural application of the proposed method. The instrumental variable z used is the predicted earnings in the regression of reported income on demographic variables, i.e., education, occupation, race, age, and region. Since z is a predicted value from a regression, it is reasonable to assume that the least-squares projection has purged the instruments from components that would affect divorce rates directly (instead of indirectly through their effect on income). Hence, our exclusions restrictions (Assumption 1) are plausibly satisfied.

The population we study includes men and women who were married and working in 1999-2003. We use a survey sample from the March Supplement of the 1999-2003 Current Population Survey. We keep only individuals who were observed for two consecutive years and who were married during the first year. To avoid the pitfall that changes in marital status can cause changes in income (e.g. women tend to have to go back to work and men may work less after a divorce.), we use personal earnings reported during the first year as a regressor and marital status in the second year as a dependent variable. The descriptive statistics in Table 7 shows that 3.5% of married men with jobs got divorced in the next year. That divorce rate is 5.7% for women.

Table 7: Descriptive statistics.

	male		female	
1999-2004	mean	std. dev.	mean	std. dev.
marital status (divorced=1)	.035	.185	.057	.233
age	45.2	11.3	43.2	10.7
race (white=1)	.89	.31	.88	.33
occupation (labor intensive=0)	.62	.48	.92	.27
earnings (thousands)*	53.3	55.5	27.2	30.5
sample size	50188		41851	

* in 2002 dollars

The parameters of the model are estimated under three identification assumptions, namely, that the measurement error has zero mode, zero mean, or zero median. We apply the model separately to the male and the female subsamples (see Table 8). The empirical results indicate that an increase in earnings decreases the probability of divorcing for both men and women. However, the effect is statistically significant for men only.

The behavior of our various estimates agrees very well with known features of measurement error in earnings. As mentioned in the introduction, Bollinger (1998) has shown that, for men, the median of the measurement error in earnings is close to zero while Bound and Krueger (1991) point out that the mode of the measurement error in earnings is close to zero for men. Our results show that, for men, the zero mode and zero median estimates are indeed very similar (and, in fact, not statistically significantly different from one another). In contrast, the estimate based on a zero mean assumption is statistically significantly different from the estimates based on mode and median restrictions. This strongly supports the view that the estimates based on mode and median assumptions should both be correct but not the one based on the mean. For women, the situation is different: Bollinger (1998) shows that women's reporting errors on earnings are much smaller and nearly classical and that the mean, mode and median restriction are all plausible. Accordingly, the point estimates obtained for women are not statistically significantly different from one another (although the coefficients themselves are not significantly different from zero, so this is not a very stringent test).

It is also possible to test for the presence of measurement error by comparing the point estimates obtained with and without correction for measurement error. For men, the null hypothesis of no measurement error can be rejected at the 5% significant level under the zero mode and zero median assumptions, which are presumably the most plausible. For women, the results are not significant, but this is not surprising given that the measurement error is known to be smaller for women and given that the coefficients themselves are not significantly different from zero.

In summary, this simple empirical illustration illustrates that our estimator performs as it should with real data.

Table 8: Earnings vs marital status.

	a		b		test for meas. error*	
1999-2004	coef.	std. dev.	coef.	std. dev.	statistics	p-value
male (n=50188)						
ignoring meas. error	-1.327	0.1008	-0.0458	0.0096		
zero mode	-0.757	0.2164	-0.1050	0.0247	7.38	0.025
zero mean	-1.387	0.2132	-0.0408	0.0244	0.884	0.643
zero median	-0.710	0.2280	-0.1091	0.0260	16.29	0.00029
female (n=41851)						
ignoring meas. error	-1.484	0.0793	-0.0095	0.0081		
zero mode	-1.355	0.1244	-0.0229	0.0140	1.333	0.513
zero mean	-1.483	0.1723	-0.0099	0.0195	0.074	0.964
zero median	-1.386	0.0961	-0.0198	0.0108	1.045	0.593

smoothing parameters: $i_n = 5, j_n = 5$ in f_1; $i_n = 3, j_n = 2$ in f_2.

*The test statistics is $\left(\widehat{\beta}_{ie} - \widehat{\beta}_{sv}\right)^T V^{-1} \left(\widehat{\beta}_{ie} - \widehat{\beta}_{sv}\right) \sim \mathcal{X}_2^2$, where $\widehat{\beta}_{ie}$ is the estimator with error ignored, $\widehat{\beta}_{sv}$ is the sieve MLE, and V is the variance-covariance matrix of $\left(\widehat{\beta}_{ie} - \widehat{\beta}_{sv}\right)$. The null hypothesis is that there is no error in x.

6 Conclusion

This paper represents the first treatment of wide class of nonclassical nonlinear errors-in-variables models with continuously distributed variables using instruments (or repeated measurements). The instruments must satisfy the intuitive requirement that they provide no more information regarding the variables of interest than the true regressors do. Our main identifying assumption exploits the observation that, even though the measurement error may not have zero mean conditional on the true value of the regressor, perhaps some other measure of location, such as the median or the mode, could still be zero. This type of nonclassical measurement error has been clearly observed, for instance, in the self-reported income found in the Current Population Survey (CPS), thanks to the exceptional availability of validation data for this dataset. More generally, there are numerous plausible settings where the conditional mode, median, or some other quantile, of the error could be zero even though its conditional mean may not.

Under suitable regularity conditions, we show that the identification problem can be cast into the form of an operator diagonalization problem in which the operator to be diagonalized is defined in terms of observable densities, while the resulting eigenvalues and eigenfunctions provide the unobserved joint densities of the variables of interest, including the unobserved error-free regressor. Our main identifying assumption is used to "index" the eigenfunctions so that the decomposition is unique.

We propose a sieve-based semiparametric estimator that is relatively simple to implement. This framework is shown to nest the two main subcases of interest, namely models that, in the absence of measurement error, would take the form of a parametric likelihood or a set of moment conditions. The estimator of the parameters of interest is shown to be root n consistent and asymptotically normal despite the presence of the infinite-dimensional nuisance parameters associated with the measurement error distributions. The finite-sample behavior of the proposed estimator is investigated through Monte Carlo simulations. An example of application to the relationship between earnings and divorce rates is also provided.

References

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

Altonji, J. G., and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1102.

Begun, J., W. Hall, W. Huang, and J. Wellner (1983): "Information and Asymptotic Efficiency in Parametric - Nonparametric Models," *Annals of Statistics*, 11, 432–452.

Blundell, R., X. Chen, and D. Kristensen (2003): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," Working Paper, London School of Economics.

Bollinger, C. R. (1998): "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 16, 576–594.

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

Bound, J., and A. B. Krueger (1991): "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right," *Journal of Labor Economics*, 9, 1–24.

Carrasco, M., and J.-P. Florens (2005): "Spectral method for deconvolving a density," Working Paper, University of Rochester.

Carrasco, M., J.-P. Florens, and E. Renault (2005): "Linear Inverse Problems and Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, vol. Vol. 6. Elsevier Science.

Chalak, K., and H. White (2006): "An Extended Class of Instrumental Variables for the Estimation of Causal Effects," Working Paper, UCSD.

Chen, X., L. Hansen, and J. Scheinkman (1997): "Shape-Preserving Estimation of Diffusions," *working paper.*

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

Chen, X., H. Hong, and A. Tarozzi (2005): "Semiparametric Efficiency in GMM Models with Nonclassical Measurement Error," Working Paper, Duke University.

Chen, X., and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66(2), 289–314.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2006): "Instrumental Variable Identification and Estimation of Nonseparable Models via Quantile Conditions," *Journal of Econometrics*, forthcoming.

Chesher, A. (1991): "The Effect of Measurement Error," *Biometrika*, 78, 451.

——— (1998): "Polynomial Regression with Covariate Measurement Error," Discussion Paper 98/448, University of Bristol.

——— (2001): "Parameter Approximations for Quantile Regressions with Measurement Error," Working Paper CWP02/01, Department of Economics, University College London.

——— (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441.

Chesher, A., M. Dumangane, and R. J. Smith (2002): "Duration response measurement error," *Journal of econometrics*, 111, 169–194.

Darolles, S., J.-P. Florens, and E. Renault (2002): "Nonparametric Instrumental Regression," Working Paper 05-2002, Centre de Recherche et Développement en Économique.

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

Fan, J. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *Annals of Statistics*, 19(3), 1257–1272.

Gallant, A., and D. Nychka (1987): "Semi-Nonparametric Maximum Likelihood Estimators," *econometrica*, 55, 363–390.

Grenander, U. (1981): *Abstract Inference*. Wiley Series, New York.

Hausman, J. (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15, 57–67.

Hausman, J., W. Newey, H. Ichimura, and J. Powell (1991): "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 273–295.

Hausman, J., W. Newey, and J. Powell (1995): "Nonlinear Errors in Variables. Estimation of Some Engel Curves," *Journal of Econometrics*, 65, 205–233.

Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

Holderlein, S., and E. Mammen (2006): "Identification of Marginal Effects in Nonseparable Models without Monotonicity," Working Paper, University of Mannheim.

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

Hu, Y. (2005): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," Working Paper, University of Texas at Austin.

Hu, Y., and G. Ridder (2004): "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," Working Paper, University of Southern California, Department of Economics.

Lewbel, A. (1996): "Demand Estimation with Expenditure Measurement Errors on the Left and Right Hand Side.," *The Review of Economics and Statistics*, 78(4), 718–725.

——— (1998): "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121.

——— (2006): "Estimation of Average Treatment Effects With Misclassification," *Econometrica*, forthcoming.

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

Mahajan, A. (2006): "Identification and Estimation of Single Index Models with Misclassified Regressor," *Econometrica*, forthcoming.

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonparametric Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

Newey, W. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627.

Newey, W. K. (1995): “Convergence Rates for Series Estimators,” in *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao*, ed. by G. Maddalla, P. Phillips, and T. Srinavasan, pp. 254–275. Blackwell, Cambridge, USA.

——— (1997): “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, 79, 147–168.

Newey, W. K., and J. L. Powell (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71, 1565–1578.

Schennach, S. M. (2004a): “Estimation of Nonlinear Models with Measurement Error,” *Econometrica*, 72, 33–75.

——— (2004b): “Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models,” Working Paper, University of Chicago, http://home.uchicago.edu/~smschenn/nlme_iv.pdf.

——— (2004c): “Nonparametric Estimation in the Presence of Measurement Error,” *Econometric Theory*, 20, 1046–1093.

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

Stein, C. (1956): “Efficient Nonparametric Testing and Estimation,” *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1, 187–195.

Taupin, M.-L. (1998): “Estimation in the Nonlinear Errors-in-Variables Model,” *C. R. Acad. Sci. Paris*, 326, Serie I, 885–890.

Wang, L., and C. Hsiao (1995): “Simulation-Based Semiparametric Estimation of Nonlinear Errors-in-Variables Models,” Working Paper, University of Southern California.

A Nonuniqueness of the indexing of the eigenvalues

Let x^* and $\tilde{x}^*$ be related through $x^* = R(\tilde{x}^*)$, where $R(\tilde{x}^*)$ is a given piecewise differentiable function. We now show that, without Assumption 5, models in which x^* or $\tilde{x}^*$ are the unobserved true regressors are observationally equivalent, because

$$L_{x|\tilde{x}^*} L_{y;\tilde{x}^*|\tilde{x}^*} L^{-1}_{x|\tilde{x}^*} = L_{x|x^*} L_{y;x^*|x^*} L^{-1}_{x|x^*},$$

where the operators $L_{y;\tilde{x}^*|\tilde{x}^*}$ and $L_{x|\tilde{x}^*}$ are defined as follows

$$\begin{aligned} \left[L_{y;\tilde{x}^*|\tilde{x}^*} g\right](\tilde{x}^*) &= f_{y|\tilde{x}^*}(y|\tilde{x}^*)\, g(\tilde{x}^*) \\ \left[L_{x|\tilde{x}^*} g\right](x) &= \int f_{x|\tilde{x}^*}(x|\tilde{x}^*)\, g(\tilde{x}^*)\, d\tilde{x}^*. \end{aligned}$$

We first note that the operators $L_{y;\tilde{x}^*|\tilde{x}^*}$ and $L_{x|\tilde{x}^*}$ can also be written in terms of $f_{y|x^*}$ and $f_{x|x^*}$ as

$$\begin{aligned} \left[L_{y;\tilde{x}^*|\tilde{x}^*} g\right](\tilde{x}^*) &= f_{y|x^*}(y|R(\tilde{x}^*))\, g(\tilde{x}^*) \\ \left[L_{x|\tilde{x}^*} g\right](x) &= \int f_{x|x^*}(x|R(\tilde{x}^*))\, g(\tilde{x}^*)\, d\tilde{x}^*. \end{aligned}$$

It can be verified (by calculating $L_{x|\tilde{x}^*} L^{-1}_{x|\tilde{x}^*} g$) that $L^{-1}_{x|\tilde{x}^*}$ is given by

$$\left[L^{-1}_{x|\tilde{x}^*} g\right](\tilde{x}^*) = r(\tilde{x}^*) \left[L^{-1}_{x|x^*} g\right](R(\tilde{x}^*)).$$

where $r(\tilde{x}^*) = dR(\tilde{x}^*)/d\tilde{x}^*$ whenever this differential exists and $r(\tilde{x}^*) = 0$ otherwise.[17] We can then calculate

$$\begin{aligned} \left[L_{x|\tilde{x}^*} L_{y;\tilde{x}^*|\tilde{x}^*} L^{-1}_{x|\tilde{x}^*} g\right](x) &= \int f_{x|x^*}(x|R(\tilde{x}^*))\, f_{y|x^*}(y|R(\tilde{x}^*))\, r(\tilde{x}^*) \left[L^{-1}_{x|x^*} g\right](R(\tilde{x}^*))\, d\tilde{x}^* \\ &= \int f_{x|x^*}(x|R(\tilde{x}^*))\, f_{y|x^*}(y|R(\tilde{x}^*)) \left[L^{-1}_{x|x^*} g\right](R(\tilde{x}^*))\, dR(\tilde{x}^*) \\ &= \int f_{x|x^*}(x|x^*)\, f_{y|x^*}(y|x^*) \left[L^{-1}_{x|x^*} g\right](x^*)\, dx^* \text{ (substituting } x^* = R(\tilde{x}^*)\text{)} \\ &= \left[L_{x|x^*} L_{y;x^*|x^*} L^{-1}_{x|x^*} g\right](x). \end{aligned}$$

It follows that indexing the eigenfunctions by $\tilde{x}^*$ or x^* produces observationally equivalent models but imply different joint densities of x and of the true regressor (x^* or $\tilde{x}^*$).

[17] Since $R(\tilde{x}^*)$ is piecewise differentiable, $dR(\tilde{x}^*)/d\tilde{x}^*$ exists almost everywhere and the points where it does not will not affect the value of the integral.

B Proofs

Proof of Lemma 2. First note that Assumptions 1-5 imply that the model is identified so that α_0 is uniquely defined. We prove the results by checking the conditions in Theorem 4.1 in Newey and Powell (2003). Their assumption1 on identification of the unknown parameter is assumed directly. We assume $k_n \to \infty$ and $k_n/n \to 0$ in assumption 12 so that the relevant part of their assumption 2 is satisfied. Note that we do not have any "plug-in" nonparametric part in the likelihood function. The first part of their condition 3 is assumed in our assumption 11(i). For the rest of their condition 3, we consider pathwise derivative

$$\begin{aligned} & \ln f_{yx|z}(D;\alpha_1) - \ln f_{yx|z}(D;\alpha_2) \\ = & \frac{d \ln f_{yx|z}(D,\overline{\alpha}_0)}{d\alpha}[\alpha_1 - \alpha_2] \\ = & \left. \frac{d}{dt} \ln f_{yx|z}(D;\overline{\alpha}_0 + t(\alpha_1-\alpha_2)) \right|_{t=0}, \end{aligned} \tag{45}$$

where $\overline{\alpha}_0 = (\overline{\theta}, \overline{f}_1, \overline{f}_2)^T$ is a mean value between α_1 and α_2. Let $\alpha_1 = (\theta_1, f_{11}, f_{21})^T$ and $\alpha_2 = (\theta_2, f_{12}, f_{22})^T$, we have

$$\begin{aligned} & \left. \frac{d}{dt} \ln f_{yx|z}(D;\overline{\alpha}_0 + t(\alpha_1-\alpha_2)) \right|_{t=0} \\ = & \frac{1}{f_{yx|z}(D,\overline{\alpha}_0)} \left\{ \int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})(\theta_1-\theta_2)\overline{f}_1(x|x^*)\overline{f}_2(x^*|z)dx^* + \right. \\ & + \int f_{y|x^*}(y|x^*;\overline{\theta})[f_{11}-f_{12}]\overline{f}_2(x^*|z)dx^* + \\ & \left. + \int f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)[f_{21}-f_{22}]dx^* \right\}. \end{aligned} \tag{46}$$

The bounds can be found as follows:

$$\begin{aligned} & \left| \frac{d}{dt} \ln f_{yx|z}(D;\overline{\alpha}_0 + t(\alpha_1-\alpha_2)) \right|_{t=0} \\ \leq & \frac{1}{\left|f_{yx|z}(D,\overline{\alpha}_0)\right|} \left\{ \int \left| \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(\xi)\overline{f}_1(x|x^*)\overline{f}_2(x^*|z) \right| dx^* \left\|\theta_1-\theta_2\right\|_s + \right. \\ & + \int \left| f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(x,x^*)\overline{f}_2(x^*|z) \right| dx^* \left\|f_{11}-f_{12}\right\|_s + \\ & \left. + \int \left| f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)\omega^{-1}(x^*,z) \right| dx^* \left\|f_{21}-f_{22}\right\|_s \right\} \end{aligned} \tag{47}$$

$$
\begin{aligned}
&\leq \frac{1}{\left|f_{yx|z}(D,\overline{\alpha}_0)\right|}\left\{\int\left|\frac{d}{d\theta}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(\xi)\,\overline{f}_1(x|x^*)\overline{f}_2(x^*|z)\right|dx^*+\right.\\
&\quad+\int\left|f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(x,x^*)\,\overline{f}_2(x^*|z)\right|dx^*+\\
&\quad\left.+\int\left|f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)\omega^{-1}(x^*,z)\right|dx^*\right\}\|\alpha-\alpha_0\|_s\\
&\equiv \left|\frac{f^{|1|}_{yx|z}(D,\overline{\alpha}_0,\bar{\omega})}{f_{yx|z}(D,\overline{\alpha}_0)}\right|\|\alpha-\alpha_0\|_s\,,
\end{aligned}
$$

where $f^{|1|}_{yx|z}(D,\overline{\alpha}_0,\bar{\omega})$ is defined as $\frac{d}{dt}f_{yx|z}(D;\overline{\alpha}_0+t\bar{\omega})\big|_{t=0}$ with each linear term, i.e., $\frac{d}{d\theta}f_{y|x^*}$, $\overline{f}_1$, and $\overline{f}_2$, replaced by its absolute value. The function $\bar{\omega}$ is defined as

$$
\overline{\omega}(\xi,x,x^*,z)=\left[1,\omega^{-1}(\xi),\omega^{-1}\left((x,x^*)^T\right),\omega^{-1}\left((x^*,z)^T\right)\right]^T
$$

with $\xi\in\mathcal{U}$. Therefore, our assumption 11(ii), i.e., $E\left(\frac{f^{|1|}_{yx|z}(D,\overline{\alpha}_0,\bar{\omega})}{f_{yx|z}(D,\overline{\alpha}_0)}\right)^2\leq E\left(h_1(D)\right)^2<\infty$, implies that $\ln f_{yx|z}(D,\alpha)$ is Hölder continuous in α. Therefore, their condition 3 holds. Assumption 10 guarantees that $\mathcal{A}$ is compact under the norm $\|\cdot\|_s$, which is their condition 4. From Chen, Hansen, and Scheinkman (1997), for any $\alpha\in\mathcal{A}$

$$
\begin{aligned}
\|\alpha-\Pi_n\alpha\|_s &\leq \|\eta-\Pi_n\eta\|_s+\|f_1-\Pi_nf_1\|_s+\|f_2-\Pi_nf_2\|_s \qquad (48)\\
&= O\left(k_n^{-\gamma_1/d_1}\right)
\end{aligned}
$$

with $d_1=2$. Therefore, their condition 5 is satisfied with our assumption 12. A similar proof can also be found in that of Lemma 3.1 and Proposition 3.1 in Ai and Chen (2003). ■

Proof of Theorem 3. First note that Assumptions 1-5 imply that the model is identified so that α_0 is uniquely defined. We prove the results by checking the conditions in Theorem 3.1 in Ai and Chen (2003). Note that there are two different estimated criterion functions, i.e., $L_n(\alpha)$ and $\widehat{L}_n(\alpha)$ in their appendix B (page 1825). In our setup, we do not have that distinction and their proof still applies with $L_n(\alpha)=\frac{1}{n}\sum_{i=1}^n\ln f_{yx|z}(D_i,\alpha)$. From the proof of lemma 2, assumptions 11 and 13 imply their condition 3.5(iii), i.e., $\|\alpha-\Pi_n\alpha\|=o\left(n^{-1/4}\right)$. Assumption 3.6(iii), 3.7 and 3.8 in Chen and Shen (1998) are assumed directly in our assumptions 14, 17 and 18, respectively. According to its expression, $f_{yx|z}(D;\alpha)$ is pathwise differentiable at α_0 if $f_{y|x^*}(y|x^*;\theta)$ is pathwise differentiable at θ_0. Therefore, assumption 15 implies their condition 3.9(i). Condition 3.9(ii) in Ai and Chen (2003) is assumed directly in assumption 16. Thus, the results of consistency follow. ■

Proof of Theorem 4. First note that Assumptions 1-5 imply that the model is identified so that α_0 is uniquely defined. We prove the results by checking the conditions in theorem 1 in Shen (1997). We define the remainder term as follows:

$$r\left[\alpha-\alpha_0, D\right] \equiv \ln f_{yx|z}(D,\alpha)-\ln f_{yx|z}(D,\alpha_0)-\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]. \tag{49}$$

We also define $\mu_n(g)=\frac{1}{n}\sum_{i=1}^{n}\left[g(D,\alpha)-Eg(D,\alpha)\right]$ as the empirical process induced by g. We have the sieve estimator $\widehat{\alpha}_n$ for α_0 and a local alternative $\alpha^*\left(\widehat{\alpha}_n,\varepsilon_n\right)=\left(1-\varepsilon_n\right)\widehat{\alpha}_n+\varepsilon_n\left(v^*+\alpha_0\right)$ with $\varepsilon_n=o\left(n^{-1/2}\right)$. Let $\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)$ be the projection of $\alpha^*\left(\alpha,\varepsilon_n\right)$ to $\mathcal{A}_n$.

First of all, the Riesz representor v^* is finite because the matrix J is invertible and w^* is bounded. Second, equation (4.2) in Shen (1997), i.e.

$$\left|s(\alpha)-s(\alpha_0)-\frac{ds(\alpha)}{d\alpha}\left[\alpha-\alpha_0\right]\right|\leq c\left\|\alpha-\alpha_0\right\|^{\omega}, \tag{50}$$

as $\left\|\alpha-\alpha_0\right\|\to 0$, is required by theorem 1 in that paper, and holds trivially in our paper with $\omega=\infty$ because we have $s(\alpha)\equiv\lambda^T b$.

Third, condition A in Shen (1997) requires

$$\sup_{\alpha\in\mathcal{N}_{0n}}\mu_n\left(r\left[\alpha-\alpha_0,D\right]-r\left[\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)-\alpha_0,D\right]\right)=O_p\left(\varepsilon_n^2\right). \tag{51}$$

By the definition of $r\left[\alpha-\alpha_0,D\right]$, we have

$$\begin{aligned}
& \mu_n\left(r\left[\alpha-\alpha_0,D\right]-r\left[\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)-\alpha_0,D\right]\right) && (52)\\
= & \ \mu_n\left\{\left(\ln f_{yx|z}(D,\alpha)-\ln f_{yx|z}(D,\alpha_0)-\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)\right.\\
& \left.-\left(\ln f_{yx|z}(D,\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right))-\ln f_{yx|z}(D,\alpha_0)-\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)-\alpha_0\right]\right)\right\}\\
= & \ \mu_n\left(\ln f_{yx|z}(D,\alpha)-\ln f_{yx|z}(D,\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right))-\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right]\right).
\end{aligned}$$

The Taylor expansion gives

$$\begin{aligned}
& \ln f_{yx|z}(D,\alpha)-\ln f_{yx|z}(D,\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)) && (53)\\
= & \ \frac{d\ln f_{yx|z}(D,\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right))}{d\alpha}\left[\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right]+\\
& +\frac{1}{2}\frac{d^2\ln f_{yx|z}(D,\widetilde{\alpha}_1)}{d\alpha d\alpha^T}\left[\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right),\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right],
\end{aligned}$$

where $\widetilde{\alpha}_1$ is a mean value between α and $\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)$. Therefore, we have

$$\begin{aligned}
& \mu_n\left(r\left[\alpha-\alpha_0,D\right]-r\left[\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)-\alpha_0,D\right]\right) && (54)\\
= & \ \mu_n\left(\frac{d\ln f_{yx|z}(D,\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right))}{d\alpha}\left[\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right]-\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right]\right)+\\
& +\mu_n\left(\frac{1}{2}\frac{d^2\ln f_{yx|z}(D,\widetilde{\alpha}_1)}{d\alpha d\alpha^T}\left[\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right),\alpha-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right]\right).
\end{aligned}$$

Since

$$\alpha - \Pi_n \alpha^* (\alpha, \varepsilon_n) = \varepsilon_n \Pi_n (\alpha - \alpha_0 - v^*),$$

the right-hand side of equation 54 equals

$$\begin{aligned}
&= \mu_n \left(\frac{d^2 \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha d\alpha^T} \left[\alpha - \Pi_n \alpha^* (\alpha, \varepsilon_n), \Pi_n \alpha^* (\alpha, \varepsilon_n) - \alpha_0 \right] \right) + \\
&\quad + \mu_n \left(\frac{1}{2} \frac{d^2 \ln f_{yx|z}(D, \widetilde{\alpha}_1)}{d\alpha d\alpha^T} \left[\varepsilon_n \Pi_n (\alpha - \alpha_0 - v^*), \varepsilon_n \Pi_n (\alpha - \alpha_0 - v^*) \right] \right)
\end{aligned}$$

$$\begin{aligned}
&= \mu_n \left(\frac{d^2 \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha d\alpha^T} \left[\varepsilon_n \Pi_n (\alpha - \alpha_0 - v^*), \Pi_n \alpha^* (\alpha, \varepsilon_n) - \alpha_0 \right] \right) + \\
&\quad + \mu_n \left(\frac{1}{2} \frac{d^2 \ln f_{yx|z}(D, \widetilde{\alpha}_1)}{d\alpha d\alpha^T} \left[\varepsilon_n \Pi_n (\alpha - \alpha_0 - v^*), \varepsilon_n \Pi_n (\alpha - \alpha_0 - v^*) \right] \right)
\end{aligned}$$

$$\begin{aligned}
&= \varepsilon_n \mu_n \left(\frac{d^2 \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha d\alpha^T} \left[\Pi_n (\alpha - \alpha_0 - v^*), \varepsilon_n \Pi_n (v^* + \alpha_0 - \alpha) + (\alpha - \alpha_0) \right] \right) \\
&\quad + \varepsilon_n^2 \mu_n \left(\frac{1}{2} \frac{d^2 \ln f_{yx|z}(D, \widetilde{\alpha}_1)}{d\alpha d\alpha^T} \left[\Pi_n (\alpha - \alpha_0 - v^*), \Pi_n (\alpha - \alpha_0 - v^*) \right] \right)
\end{aligned}$$

$$\begin{aligned}
&= \varepsilon_n \mu_n \left(\frac{d^2 \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha d\alpha^T} \left[\Pi_n (\alpha - \alpha_0 - v^*), \alpha - \alpha_0 \right] \right) + \\
&\quad - \varepsilon_n^2 \mu_n \left(\frac{1}{2} \frac{d^2 \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha d\alpha^T} \left[\Pi_n (\alpha - \alpha_0 - v^*), \Pi_n (\alpha - \alpha_0 - v^*) \right] \right) + \\
&\quad + \varepsilon_n^2 \mu_n \left(\frac{1}{2} \frac{d^2 \ln f_{yx|z}(D, \widetilde{\alpha}_1)}{d\alpha d\alpha^T} \left[\Pi_n (\alpha - \alpha_0 - v^*), \Pi_n (\alpha - \alpha_0 - v^*) \right] \right) \\
&= A_1 + A_2 + A_3, \qquad (55)
\end{aligned}$$

where $\overline{\alpha}_1$ a mean value between α_0 and $\Pi_n \alpha^* (\alpha, \varepsilon_n)$. We consider the term A_1 as follows:

$$\sup_{\alpha \in \mathcal{N}_{0n}} A_1 = \varepsilon_n \sup_{\alpha \in \mathcal{N}_{0n}} \mu_n \left(\frac{d^2 \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha d\alpha^T} \left[\Pi_n (\alpha - \alpha_0 - v^*), \alpha - \alpha_0 \right] \right). \qquad (56)$$

Let $\overline{\alpha}_1 = \left(\overline{\theta}, \overline{f}_1, \overline{f}_2 \right)$ and $v_n = \Pi_n (\alpha - \alpha_0 - v^*) = \left([v_n]_\theta, [v_n]_{f_1}, [v_n]_{f_2} \right)$. We consider the term

$$\begin{aligned}
&\left|\sup_{\alpha\in\mathcal{N}_{0n}} \frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T}\left[v_n, \alpha-\alpha_0\right]\right| \qquad (57)\\
\leq\ & \sup_{\alpha\in\mathcal{N}_{0n}} \left| \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T}\left[v_n,(\alpha-\alpha_0)\right] + \right.\\
& \left. - \frac{d \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[v_n\right] \frac{d \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[\alpha-\alpha_0\right]\right|\\
\leq\ & \sup_{\alpha\in\mathcal{N}_{0n}} \left(\left| \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T}\left[v_n,(\alpha-\alpha_0)\right]\right| + \right.\\
& \left. + \left|\frac{d \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[v_n\right]\right| \left|\frac{d \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[\alpha-\alpha_0\right]\right|\right).
\end{aligned}$$

We need to find the bounds on three terms in the absolute value. We have

$$\begin{aligned}
& \frac{d \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[\alpha-\alpha_0\right] \qquad (58)\\
=\ & \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \left\{ \int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})\left(\theta-\theta_0\right) \overline{f}_1(x|x^*)\overline{f}_2(x^*|z)dx^* + \right.\\
& + \int f_{y|x^*}(y|x^*;\overline{\theta})\left[f_1 - f_{x|x^*}\right]\overline{f}_2(x^*|z)dx^* + \\
& \left. + \int f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)\left[f_2 - f_{x^*|z}\right]dx^* \right\}.
\end{aligned}$$

Therefore, the term $\left|\frac{d\ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[\alpha-\alpha_0\right]\right|$ can be bounded through

$$\begin{aligned}
& \left|\frac{d \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[\alpha-\alpha_0\right]\right| \qquad (59)\\
\leq\ & \frac{1}{\left|f_{yx|z}(D,\overline{\alpha}_1)\right|} \left\{ \int \left| \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(\xi\right) \overline{f}_1(x|x^*)\overline{f}_2(x^*|z)\right| dx^* \left\|\theta-\theta_0\right\|_s + \right.\\
& + \int \left| f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(x,x^*\right)\overline{f}_2(x^*|z)\right| dx^* \left\|f_1 - f_{x|x^*}\right\|_s + \\
& \left. + \int \left| f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)\omega^{-1}\left(x^*,z\right)\right| dx^* \left\|f_2 - f_{x^*|z}\right\|_s \right\}\\
\leq\ & \left| \frac{f^{|1|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)} \right| \left\|\alpha-\alpha_0\right\|_s,
\end{aligned}$$

where $f^{|1|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)$ is define in assumption 11 and equation 47. Similarly, we also have

$$\left|\frac{d \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha}\left[v_n\right]\right| \leq \left| \frac{f^{|1|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)} \right| \left\|v_n\right\|_s \qquad (60)$$

with

$$\begin{aligned} \left\| v_n \right\|_s &= \left\| \Pi_n \left(\alpha - \alpha_0 - v^* \right) \right\|_s \\ &\leq \left\| v_n^* \right\|_s + \left\| \Pi_n \left(\alpha - \alpha_0 \right) \right\|_s < \infty. \end{aligned} \tag{61}$$

We then consider the term $\frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T} \left[v_n, (\alpha - \alpha_0) \right]$ as follows:

$$\begin{aligned} & \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T} \left[v_n, (\alpha - \alpha_0) \right] \\ = \; & \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \left\{ \int \frac{d^2}{d\theta^2} f_{y|x^*}(y|x^*;\overline{\theta}) \left[v_n \right]_\theta (\theta - \theta_0) \, \overline{f}_1(x|x^*) \overline{f}_2(x^*|z) dx^* + \right. \\ & + \int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \left[v_n \right]_\theta \left[f_1 - f_{x|x^*} \right] \overline{f}_2(x^*|z) dx^* + \\ & + \int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \left[v_n \right]_\theta \overline{f}_1(x|x^*) \left[f_2 - f_{x^*|z} \right] dx^* \\ & \qquad + \int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \left(\theta - \theta_0 \right) \left[v_n \right]_{f_1} \overline{f}_2(x^*|z) dx^* + \\ & \qquad + \int f_{y|x^*}(y|x^*;\overline{\theta}) \left[v_n \right]_{f_1} \left[f_2 - f_{x^*|z} \right] dx^* + \\ & \qquad + \int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \left(\theta - \theta_0 \right) \overline{f}_1(x|x^*) \left[v_n \right]_{f_2} dx^* + \\ & \qquad \left. + \int f_{y|x^*}(y|x^*;\overline{\theta}) \left[f_1 - f_{x|x^*} \right] \left[v_n \right]_{f_2} dx^* \right\}. \end{aligned} \tag{62}$$

Therefore, the term $\left| \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T} \left[v_n, (\alpha - \alpha_0) \right] \right|$ can be bounded through

$$\begin{aligned} & \left| \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)} \frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T} \left[v_n, (\alpha - \alpha_0) \right] \right| \\ \leq \; & \frac{1}{\left| f_{yx|z}(D,\overline{\alpha}_1) \right|} \left\{ \int \left| \frac{d^2}{d\theta^2} f_{y|x^*}(y|x^*;\overline{\theta}) \omega^{-1}(\xi) \, \omega^{-1}(\xi) \, \overline{f}_1(x|x^*) \overline{f}_2(x^*|z) \right| dx^* \left\| \left[v_n \right]_\theta \right\|_s \left\| \theta - \theta_0 \right\|_s + \right. \\ & + \int \left| \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \omega^{-1}(\xi) \, \omega^{-1}(x,x^*) \, \overline{f}_2(x^*|z) \right| dx^* \left\| \left[v_n \right]_\theta \right\|_s \left\| f_1 - f_{x|x^*} \right\|_s + \\ & + \int \left| \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \omega^{-1}(\xi) \, \overline{f}_1(x|x^*) \omega^{-1}(x^*,z) \right| dx^* \left\| \left[v_n \right]_\theta \right\|_s \left\| f_2 - f_{x^*|z} \right\|_s \\ & \qquad + \int \left| \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \omega^{-1}(\xi) \, \omega^{-1}(x,x^*) \, \overline{f}_2(x^*|z) \right| dx^* \left\| \theta - \theta_0 \right\|_s \left\| \left[v_n \right]_{f_1} \right\|_s + \\ & \qquad + \int \left| f_{y|x^*}(y|x^*;\overline{\theta}) \omega^{-1}(x,x^*) \, \omega^{-1}(x^*,z) \right| dx^* \left\| \left[v_n \right]_{f_1} \right\|_s \left\| f_2 - f_{x^*|z} \right\|_s + \\ & \qquad + \int \left| \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta}) \omega^{-1}(\xi) \, \overline{f}_1(x|x^*) \omega^{-1}(x^*,z) \right| dx^* \left\| \theta - \theta_0 \right\|_s \left\| \left[v_n \right]_{f_2} \right\|_s + \\ & \qquad \left. + \int \left| f_{y|x^*}(y|x^*;\overline{\theta}) \omega^{-1}(x,x^*) \, \omega^{-1}(x^*,z) \right| dx^* \left\| f_1 - f_{x|x^*} \right\|_s \left\| \left[v_n \right]_{f_2} \right\|_s \right\} \end{aligned} \tag{63}$$

$$
\begin{aligned}
\leq\ & \frac{1}{\left|f_{yx|z}(D,\overline{\alpha}_1)\right|}\left\{\int\left|\frac{d^2}{d\theta^2}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(\xi\right)\omega^{-1}\left(\xi\right)\overline{f}_1(x|x^*)\overline{f}_2(x^*|z)\right|dx^*+\right.\\
&+\int\left|\frac{d}{d\theta}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(\xi\right)\omega^{-1}\left(x,x^*\right)\overline{f}_2(x^*|z)\right|dx^*+\\
&+\int\left|\frac{d}{d\theta}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(\xi\right)\overline{f}_1(x|x^*)\omega^{-1}\left(x^*,z\right)\right|dx^*
\end{aligned}
$$

$$
\begin{aligned}
&+\int\left|\frac{d}{d\theta}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(\xi\right)\omega^{-1}\left(x,x^*\right)\overline{f}_2(x^*|z)\right|dx^*+\\
&+\int\left|f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(x,x^*\right)\omega^{-1}\left(x^*,z\right)\right|dx^*+\\
&+\int\left|\frac{d}{d\theta}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(\xi\right)\overline{f}_1(x|x^*)\omega^{-1}\left(x^*,z\right)\right|dx^*+\\
&\left.+\int\left|f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}\left(x,x^*\right)\omega^{-1}\left(x^*,z\right)\right|dx^*\right\}\left\|\alpha-\alpha_0\right\|_s\left\|v_n\right\|_s\\
\equiv\ & \left|\frac{f^{|2|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)}\right|\left\|\alpha-\alpha_0\right\|_s\left\|v_n\right\|_s,
\end{aligned}
$$

where $f^{|2|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)$ is defined in assumption 20. Plug-in the bounds in equations 59, 60, and 63 back to equation 57, we have

$$
\begin{aligned}
&\left|\sup_{\alpha\in\mathcal{N}_{0n}}\frac{d^2\ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T}\left[v_n,(\alpha-\alpha_0)\right]\right| \qquad (64)\\
\leq\ & \sup_{\overline{\alpha}_1\in\mathcal{N}_{0n}}\left[\left|\frac{f^{|1|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)}\right|^2+\left|\frac{f^{|2|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)}\right|\right]\left\|\alpha-\alpha_0\right\|_s\left\|v_n\right\|_s\\
\leq\ & h_2\left(D\right)\left\|\alpha-\alpha_0\right\|_s\left\|v_n\right\|_s.
\end{aligned}
$$

By the envelope condition in assumption 20, equation 56 becomes

$$
\begin{aligned}
&\sup_{\alpha\in\mathcal{N}_{0n}}A_1 \qquad (65)\\
=\ & \varepsilon_n O_p\left(n^{-1/2}\right)\sqrt{E\left(\sup_{\alpha\in\mathcal{N}_{0n}}\frac{d^2\ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T}\left[\Pi_n\left(\alpha-\alpha_0-v^*\right),(\alpha-\alpha_0)\right]\right)^2}\\
\leq\ & \varepsilon_n O_p\left(n^{-1/2}\right)\sqrt{E\left(h_2\left(D\right)\right)^2}\left\|\alpha-\alpha_0\right\|_s\left\|v_n\right\|_s\\
=\ & O_p\left(\varepsilon_n^2\right)
\end{aligned}
$$

with $\left\|\alpha-\alpha_0\right\|_s = o(1)$. The last two terms A_2 and A_3 in equation 55 are bounded as follows:

$$\begin{aligned}
& \left|\sup_{\alpha\in\mathcal{N}_{0n}} A_2\right| && (66)\\
= \; & \varepsilon_n^2 \left|\sup_{\alpha\in\mathcal{N}_{0n}} \mu_n \left(\frac{1}{2}\frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha d\alpha^T}\left[\Pi_n\left(\alpha-\alpha_0-v^*\right),\Pi_n\left(\alpha-\alpha_0-v^*\right)\right]\right)\right| \\
\leq \; & \varepsilon_n^2 \frac{1}{2}\mu_n \left(\left|\frac{f^{|1|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)}\right|^2 + \left|\frac{f^{|2|}_{yx|z}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)}\right|\right) \left\|\Pi_n\left(\alpha-\alpha_0-v^*\right)\right\|_s^2 \\
\leq \; & \varepsilon_n^2 \frac{1}{2} O_p\left(E\left|h_2(D)\right|\right)\left\|\Pi_n\left(\alpha-\alpha_0-v^*\right)\right\|_s^2 \\
= \; & O_p\left(\varepsilon_n^2\right)
\end{aligned}$$

The same result holds for $\left|\sup_{\alpha\in\mathcal{N}_{0n}} A_3\right|$, and therefore, condition A in Shen (1997) holds.

Fourth, condition B requires

$$\begin{aligned}
\sup_{\alpha\in\mathcal{N}_{0n}} \Bigg[& E\left(\ln \frac{f_{yx|z}(D,\alpha_0)}{f_{yx|z}(D,\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right))}\right) - E\left(\ln\frac{f_{yx|z}(D,\alpha_0)}{f_{yx|z}(D,\alpha)}\right) + && (67)\\
& -\frac{1}{2}\left(\left\|\alpha^*\left(\alpha,\varepsilon_n\right)-\alpha_0\right\|^2 - \left\|\alpha-\alpha_0\right\|^2\right)\Bigg] = O\left(\varepsilon_n^2\right).
\end{aligned}$$

As corollary 2 in Shen (1997) points out that condition B can be replaced by condition B' as follows:

$$E\left(\ln\frac{f_{yx|z}(D,\alpha_0)}{f_{yx|z}(D,\alpha)}\right) = \frac{1}{2}\left\|\alpha-\alpha_0\right\|^2\left(1+o(h_n)\right). \tag{68}$$

with some positive sequence $\{h_n\}\to 0$ as $n\to\infty$. We consider the Taylor expansion

$$\begin{aligned}
& E\left[\ln f_{yx|z}(D,\alpha)) - \ln f_{yx|z}(D,\alpha_0)\right] && (69)\\
= \; & E\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right) + \frac{1}{2}E\left(\frac{d^2\ln f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right) + \\
& +\frac{1}{6}\; E\frac{d^3}{dt^3}\ln f_{yx|z}(D;\alpha_0+t\left(\alpha-\alpha_0\right))\Big|_{t=0} + \\
& +\frac{1}{24}\; E\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t\left(\alpha-\alpha_0\right))\Big|_{t=0},
\end{aligned}$$

where $\overline{\alpha}$ is a mean value between α and α_0.

As for the leading terms on the right-hand side, we have η satisfying $\int_{\mathcal{Y}}\frac{\partial}{\partial\eta}f_{y|x^*}(y|x^*;\theta)dy = 0$, $\int_{\mathcal{Y}}\frac{\partial^2}{\partial\eta^2}f_{y|x^*}(y|x^*;\theta)dy = 0$, and $\int_{\mathcal{Y}}\frac{\partial^3}{\partial\eta^3}f_{y|x^*}(y|x^*;\theta)dy = 0$ for all $\theta\in\Theta$, and f_1, f_2 satisfying

$\int_{\mathcal{X}} f_1(x|x^*)dx = 1$ and $\int_{\mathcal{X}^*} f_2(x^*|z)dx = 1$. It is then tedious but straightforward to show [18]

$$\begin{aligned} E\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right) &= 0, \qquad (70)\\ E\left(\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^2 f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right) &= 0,\end{aligned}$$

$$E\left[\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^3 f_{yx|z}(D,\alpha_0)}{d\alpha^3}\left[\alpha-\alpha_0,\alpha-\alpha_0,\alpha-\alpha_0\right]\right]=0.$$

Therefore,

$$\begin{aligned} & E\left(\frac{d^2\ln f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right) \qquad (71)\\ = \ & E\left[\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^2 f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha,\alpha\right]-\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)\right]\\ = \ & -E\left[\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)\right]\\ = \ & -\left\|\alpha-\alpha_0\right\|^2. \end{aligned}$$

Therefore, equation 69 becomes

$$\begin{aligned} & E\left[\ln f_{yx|z}(D,\alpha))-\ln f_{yx|z}(D,\alpha_0)\right] \qquad (72)\\ = \ & -\frac{1}{2}\left\|\alpha-\alpha_0\right\|^2+\frac{1}{6}\ E\frac{d^3}{dt^3}\ln f_{yx|z}(D;\alpha_0+t\left(\alpha-\alpha_0\right))\bigg|_{t=0}+\\ & +\frac{1}{24}\ E\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t\left(\alpha-\alpha_0\right))\bigg|_{t=0}. \end{aligned}$$

For the second term on the right-hand side, we have

$$\begin{aligned} & \frac{d^3}{dt^3}\ln f_{yx|z}(D;\alpha_0+t\left(\alpha-\alpha_0\right))\bigg|_{t=0} \qquad (73)\\ = \ & E\left[\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^3 f_{yx|z}(D,\alpha_0)}{d\alpha^3}\left[\alpha-\alpha_0,\alpha-\alpha_0,\alpha-\alpha_0\right]\right]+\\ & -3E\left[\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^2 f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right]+\\ & +2E\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)^3\\ = \ & B_1+B_2+B_3. \end{aligned}$$

[18] We abuse the notation $\frac{d^3\ln f_{yx|z}}{d\alpha^3}$ to stand for the third order derivative with respect to a vector α.

Again, it is straightforward to show $B_1 = 0$. The term B_2 is bounded as follows:

$$\begin{aligned} & E\left[\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^2 f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right] \qquad (74)\\ \leq\ & E\left[\left|\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right|\left|\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^2 f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right|\right] \end{aligned}$$

$$\begin{aligned} \leq\ & \left[E\left|\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^2 f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right|^2\right]^{1/2}\left[E\left|\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right|^2\right]^{1/2}\\ =\ & \left[E\left|\frac{1}{f_{yx|z}(D,\alpha_0)}\frac{d^2 f_{yx|z}(D,\alpha_0)}{d\alpha d\alpha^T}\left[\alpha-\alpha_0,\alpha-\alpha_0\right]\right|^2\right]^{1/2}\left\|\alpha-\alpha_0\right\|\\ \leq\ & \left[E\left|\frac{f^{|2|}_{yx|z}\left(D,\alpha_0,\bar{\omega}\right)}{f_{yx|z}(D,\alpha_0)}\right|^2\right]^{1/2}\left\|\alpha-\alpha_0\right\|_s^2\left\|\alpha-\alpha_0\right\|\\ \leq\ & \left[E\left|h_2(D)\right|^2\right]^{1/2}\left\|\alpha-\alpha_0\right\|_s^2\left\|\alpha-\alpha_0\right\|. \end{aligned}$$

For the term B_3, we have

$$\begin{aligned} B_3\ \leq\ & E\left|\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right|^3 \qquad (75)\\ \leq\ & \left[E\left|\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right|^4\right]^{1/2}\left[E\left|\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right|^2\right]^{1/2}\\ =\ & \left[E\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)^4\right]^{1/2}\left\|\alpha-\alpha_0\right\|\\ \leq\ & \left[E\left|\frac{f^{|1|}_{yx|z}\left(D,\alpha_0,\bar{\omega}\right)}{f_{yx|z}(D,\alpha_0)}\right|^4\right]^{1/2}\left\|\alpha-\alpha_0\right\|_s^2\left\|\alpha-\alpha_0\right\|\\ \leq\ & \left[E\left|h_1(D)\right|^4\right]^{1/2}\left\|\alpha-\alpha_0\right\|_s^2\left\|\alpha-\alpha_0\right\|. \end{aligned}$$

Note that $E\left|h_2(D)\right|^2 < \infty$ implies $E\left|h_1(D)\right|^4 < \infty$. Therefore, equation 72 becomes

$$\begin{aligned} & E\left[\ln f_{yx|z}(D,\alpha)) - \ln f_{yx|z}(D,\alpha_0)\right] \qquad (76)\\ =\ & -\frac{1}{2}\left\|\alpha-\alpha_0\right\|^2 + O\left(\left\|\alpha-\alpha_0\right\|_s^2\left\|\alpha-\alpha_0\right\|\right) +\\ & +\frac{1}{24}\,E\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t\left(\alpha-\alpha_0\right))\Big|_{t=0}. \end{aligned}$$

By assumption 23, we have

$$\begin{aligned} & E\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t\left(\alpha-\alpha_0\right))\bigg|_{t=0} & (77) \\ \leq\ & E\left|\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t\left(\alpha-\alpha_0\right))\right|_{t=0} \\ \leq\ & E\left|h_4(D)\right|\left\|\alpha-\alpha_0\right\|_s^4 \\ =\ & O\left(\left\|\alpha-\alpha_0\right\|_s^4\right), & (78) \end{aligned}$$

and therefore,

$$E\left[\ln f_{yx|z}(D,\alpha_0))-\ln f_{yx|z}(D,\alpha)\right]=\frac{1}{2}\left\|\alpha-\alpha_0\right\|^2\left(1+O\left(h_n\right)\right), \tag{79}$$

with

$$h_n=\frac{\left\|\alpha-\alpha_0\right\|_s^2}{\left\|\alpha-\alpha_0\right\|}+\frac{\left\|\alpha-\alpha_0\right\|_s^4}{\left\|\alpha-\alpha_0\right\|^2}.$$

Next, we show that $\frac{\|\alpha-\alpha_0\|_s^2}{\|\alpha-\alpha_0\|}\to 0$ as $n\to\infty$. We will need the convergence rate of the sieve coefficients. Therefore, we define for $\alpha\in\mathcal{N}_{0n}$

$$\begin{aligned} \alpha &= \left(\begin{array}{cccc} b^T, & \Pi_n\eta, & \Pi_n f_1, & \Pi_n f_2 \end{array}\right)^T & (80) \\ &= \left(\begin{array}{cccc} b^T, & p^{k_n}\left(\xi_1,\xi_2\right)^T\delta, & p^{k_n}\left(x,x^*\right)^T\beta, & p^{k_n}\left(x^*,z\right)^T\gamma \end{array}\right)^T, \end{aligned}$$

$$\begin{aligned} \Pi_n\alpha_0 &= \left(\begin{array}{cccc} b_0^T, & \Pi_n\eta_0, & \Pi_n f_{x|x^*}, & \Pi_n f_{x^*|z} \end{array}\right)^T \\ &= \left(\begin{array}{cccc} b_0^T, & p^{k_n}\left(\xi_1,\xi_2\right)^T\delta_0, & p^{k_n}\left(x,x^*\right)^T\beta_0, & p^{k_n}\left(x^*,z\right)^T\gamma_0 \end{array}\right)^T, \end{aligned}$$

where p^{k_n}'s are k_n-by-1 vectors i.e., $p^{k_n}\left(\cdot,\cdot\right)=\left(\begin{array}{cccc} p_1^{k_n}\left(\cdot,\cdot\right), & p_2^{k_n}\left(\cdot,\cdot\right), & ..., & p_{k_n}^{k_n}\left(\cdot,\cdot\right) \end{array}\right)^T$. Note that all the vectors are column vectors. We also define the vector of the sieve coefficients as

$$\begin{aligned} \alpha^c &= \left(\begin{array}{cccc} b^T, & \delta^T, & \beta^T, & \gamma^T \end{array}\right)^T, & (81) \\ \alpha_0^c &= \left(\begin{array}{cccc} b_0^T, & \delta_0^T, & \beta_0^T, & \gamma_0^T \end{array}\right)^T. \end{aligned}$$

We then have

$$\begin{aligned} & \alpha-\alpha_0 & (82) \\ =\ & \alpha-\Pi_n\alpha_0+\Pi_n\alpha_0-\alpha_0 \\ =\ & \left(\begin{array}{cccc} \left(b^T-b_0^T\right), & p^{k_n}\left(\xi_1,\xi_2\right)^T\left(\delta-\delta_0\right), & p^{k_n}(x,x^*)^T\left(\beta-\beta_0\right), & p^{k_n}(x^*,z)^T\left(\gamma-\gamma_0\right) \end{array}\right) \\ & +\Pi_n\alpha_0-\alpha_0. \end{aligned}$$

Suppose that

$$\left\|\alpha-\alpha_0\right\|=O\left(n^{-1/4-\varsigma_0}\right)$$

with some small $\varsigma_0 > 0$. By assumption 13 and equation 48, we let

$$\left\|\Pi_n\alpha_0 - \alpha_0\right\|_s = O\left(k_n^{-\gamma_1/d_1}\right) = O\left(n^{-1/4-\varsigma}\right) \tag{83}$$

for some small $\varsigma > \varsigma_0$.

We then show $\left\|\alpha^c - \alpha_0^c\right\|_E = O\left(n^{-1/4-\varsigma_0}\right)$ from $\left\|\alpha - \alpha_0\right\| = O\left(n^{-1/4-\varsigma_0}\right)$. For any $\alpha \in \mathcal{N}_{0n}$, we have

$$\left|\left\|\alpha - \alpha_0\right\| - \left\|\Pi_n\alpha_0 - \alpha_0\right\|\right| \leq \left\|\alpha - \Pi_n\alpha_0\right\| \leq \left\|\alpha - \alpha_0\right\| + \left\|\Pi_n\alpha_0 - \alpha_0\right\|. \tag{84}$$

We have shown that assumption 11 implies $E\left|\frac{f_{yx|z}^{|1|}(D,\overline{\alpha}_1,\bar{\omega})}{f_{yx|z}(D,\overline{\alpha}_1)}\right|^2 \leq E\left|h_1(D)\right|^2 < \infty$. We then have

$$\begin{aligned} & \left\|\Pi_n\alpha_0 - \alpha_0\right\| && (85)\\ \leq\ & \sqrt{E\left(\frac{f_{yx|z}^{|1|}\left(D,\overline{\alpha}_1,\bar{\omega}\right)}{f_{yx|z}(D,\overline{\alpha}_1)}\right)^2}\left\|\Pi_n\alpha_0 - \alpha_0\right\|_s \\ =\ & O\left(\left\|\Pi_n\alpha_0 - \alpha_0\right\|_s\right) \\ \leq\ & O\left(k_n^{-\gamma_1/d_1}\right) \\ =\ & O\left(n^{-1/4-\varsigma}\right), \end{aligned}$$

and therefore, for some constants $0 < C_1, C_2 < \infty$

$$C_1\left\|\alpha - \alpha_0\right\| \leq \left\|\alpha - \Pi_n\alpha_0\right\| \leq C_2\left\|\alpha - \alpha_0\right\|. \tag{86}$$

Moreover, we define

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{db} = \left(\ \frac{d\ln f_{yx|z}(D,\alpha_0)}{db_1},\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{db_2},\quad ...,\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{db_{d_b}}\ \right)^T, \tag{87}$$

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}\left[p^{k_n}\right] = \left(\ \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}\left[p_1^{k_n}\right],\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}\left[p_2^{k_n}\right],\quad ...,\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}\left[p_{k_n}^{k_n}\right]\ \right)^T,$$

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[p^{k_n}\right] = \left(\ \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[p_1^{k_n}\right],\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[p_2^{k_n}\right],\quad ...,\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[p_{k_n}^{k_n}\right]\ \right)^T,$$

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[p^{k_n}\right] = \left(\ \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[p_1^{k_n}\right],\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[p_2^{k_n}\right],\quad ...,\quad \frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[p_{k_n}^{k_n}\right]\ \right)^T,$$

$$\begin{aligned} & \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[p^{k_n}\right] \\ =\ & \left[\ \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{db}\right)^T,\quad \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}\left[p^{k_n}\right]\right)^T,\quad \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[p^{k_n}\right]\right)^T,\quad \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[p^{k_n}\right]\right)^T\ \right]^T. \end{aligned}$$

With the notations above, we have

$$\begin{aligned}
& \frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\Pi_n\alpha_0\right] && (88)\\
= \; & \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{db}\right)^T (b-b_0)+\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}\left[p^{k_n}\right]\right)^T(\delta-\delta_0)\\
& +\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}\left[p^{k_n}\right]\right)^T(\beta-\beta_0)+\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}\left[p^{k_n}\right]\right)^T(\gamma-\gamma_0)\\
= \; & \left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[p^{k_n}\right]\right)^T(\alpha^c-\alpha_0^c),
\end{aligned}$$

and

$$\begin{aligned}
& \left\|\alpha-\Pi_n\alpha_0\right\|^2 && (89)\\
= \; & E\left\{\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\Pi_n\alpha_0\right]\right)^2\right\}\\
= \; & (\alpha^c-\alpha_0^c)^T E\left\{\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[p^{k_n}\right]\right)\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[p^{k_n}\right]\right)^T\right\}(\alpha^c-\alpha_0^c)\\
\equiv \; & (\alpha^c-\alpha_0^c)^T\,\Omega_{k_n}(\alpha^c-\alpha_0^c).
\end{aligned}$$

The matrix Ω_{k_n} is positive definite with its smallest eigenvalue bounded away from zero uniformly in k_n according to assumption 21. Since $\|\alpha-\Pi_n\alpha_0\|$ is always finite, the largest eigenvalue of Ω_{k_n} is finite. Thus, we have for some constants $0<C_1,C_2<\infty$

$$C_1\left\|\alpha^c-\alpha_0^c\right\|_E\le\left\|\alpha-\Pi_n\alpha_0\right\|\le C_2\left\|\alpha^c-\alpha_0^c\right\|_E. \tag{90}$$

Note that C_1 and C_2 are general constants that may take different values at each appearance.

We then consider the ratio $\frac{\|\alpha-\alpha_0\|_s^2}{\|\alpha-\alpha_0\|}$. From equations 86 and 90, we have

$$\left\|\alpha-\alpha_0\right\|\ge C_1\left\|\alpha^c-\alpha_0^c\right\|_E \tag{91}$$

and $\left\|\alpha^c-\alpha_0^c\right\|_E=O\left(n^{-1/4-\varsigma_0}\right)$. Assumption 21 implies $\left\|\alpha-\Pi_n\alpha_0\right\|_s^2\le C_2\left\|\alpha^c-\alpha_0^c\right\|_1^2$, where $\|\cdot\|_1$ is the L_1 vector norm. Thus, we have

$$\begin{aligned}
\left\|\alpha-\alpha_0\right\|_s^2 &\le \left\|\alpha-\Pi_n\alpha_0\right\|_s^2+\left\|\Pi_n\alpha_0-\alpha_0\right\|_s^2 && (92)\\
&\le C_2\left\|\alpha^c-\alpha_0^c\right\|_1^2+O\left(k_n^{-2\gamma_1/d_1}\right)\\
&\le C_2k_n\left\|\alpha^c-\alpha_0^c\right\|_E^2+O\left(n^{2(-1/4-\varsigma)}\right).
\end{aligned}$$

Since $\left\|\alpha^c-\alpha_0^c\right\|_E=O\left(n^{-1/4-\varsigma_0}\right)$ and $\varsigma>\varsigma_0$, we have

$$\left\|\alpha-\alpha_0\right\|_s^2\le C_2k_n\left\|\alpha^c-\alpha_0^c\right\|_E^2. \tag{93}$$

By equations 91 and 93, we have

$$\begin{aligned} \frac{\left\|\alpha-\alpha_0\right\|_s^2}{\left\|\alpha-\alpha_0\right\|} &\leq \frac{C_2 k_n \left\|\alpha^c-\alpha_0^c\right\|_E^2}{C_1 \left\|\alpha^c-\alpha_0^c\right\|_E} \\ &\leq O\left(k_n \left\|\alpha^c-\alpha_0^c\right\|_E\right). \end{aligned} \tag{94}$$

Assumption 13 requires $k_n^{-\gamma_1/d_1} = O\left(n^{-1/4-\varsigma}\right)$, i.e., $k_n = n^{\left(\frac{1}{4}+\varsigma\right)\frac{1}{\gamma_1/d_1}}$. We then have

$$\begin{aligned} k_n \left\|\alpha^c-\alpha_0^c\right\|_E &= O\left(n^{-\frac{1}{4}\left(1-\frac{1}{\gamma_1/d_1}\right)+\varsigma\frac{1}{\gamma_1/d_1}-\varsigma_0}\right) \\ &= o(1) \end{aligned} \tag{95}$$

for $\varsigma < \frac{1}{4}\left(\frac{\gamma_1}{d_1}-1\right)+\frac{\gamma_1}{d_1}\varsigma_0$ with $\gamma_1/d_1 > 1$ in assumption 13. Therefore, equation 79 holds with the positive sequence $\{h_n\} \to 0$ as $n \to \infty$. That means that condition B' in Shen (1997) holds.

Fifth, Condition C in Shen (1997) requires

$$\sup_{\alpha\in\mathcal{N}_{0n}} \left\|\alpha^*\left(\alpha,\varepsilon_n\right)-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right\| = O\left(n^{-1/4}\varepsilon_n\right). \tag{96}$$

By definition, we have $\alpha^*\left(\alpha,\varepsilon_n\right) = \left(1-\varepsilon_n\right)\alpha+\varepsilon_n\left(v^*+\alpha_0\right)$ with $\alpha\in\mathcal{N}_{0n}$. Therefore,

$$\begin{aligned} & \left\|\alpha^*\left(\alpha,\varepsilon_n\right)-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right\| \\ = \; & \varepsilon_n \left\|v^*+\alpha_0-\Pi_n\left(v^*+\alpha_0\right)\right\| \\ \leq \; & \varepsilon_n \left\|v^*-\Pi_n v^*\right\|+\varepsilon_n\left\|\alpha_0-\Pi_n\alpha_0\right\| \\ = \; & O\left(n^{-1/4}\varepsilon_n\right). \end{aligned} \tag{97}$$

The last step is due to assumption 22. Condition C also requires

$$\sup_{\alpha\in\mathcal{N}_{0n}} \mu_n\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha^*\left(\alpha,\varepsilon_n\right)-\Pi_n\alpha^*\left(\alpha,\varepsilon_n\right)\right]\right) = O_p\left(\varepsilon_n^2\right). \tag{98}$$

The left-hand side equals

$$\varepsilon_n\mu_n\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[v^*-v_n^*\right]\right)+\varepsilon_n\mu_n\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha_0-\Pi_n\alpha_0\right]\right). \tag{99}$$

By the envelope condition in assumption 11, the first term corresponding to v^* is

$$\begin{aligned} & \left|\mu_n\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[v^*-v_n^*\right]\right)\right| \\ = \; & \sqrt{E\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[v^*-v_n^*\right]\right)^2} O_p\left(n^{-1/2}\right) \\ = \; & \left\|v^*-v_n^*\right\| O_p\left(n^{-1/2}\right) \\ = \; & o_p\left(n^{-1/2}\right), \end{aligned} \tag{100}$$

and the second term corresponding to α_0 is

$$\begin{aligned}
& \left|\mu_n\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha_0-\Pi_n\alpha_0\right]\right)\right| \qquad (101)\\
= & \sqrt{E\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha_0-\Pi_n\alpha_0\right]\right)^2}O_p\left(n^{-1/2}\right)\\
= & \left\|\alpha_0-\Pi_n\alpha_0\right\|O_p\left(n^{-1/2}\right)\\
= & o_p\left(n^{-1/2}\right).
\end{aligned}$$

The last step is due to $\left\|\alpha_0-\Pi_n\alpha_0\right\|=o\left(n^{-1/4}\right)$. Therefore, condition C in theorem 1 in Shen (1997) holds. Note that condition C' in corollary 2 is also satisfied, i.e., $\|v_n^*-v^*\|=o(n^{-1/4})$ and $o\left(h_n\right)\left\|\alpha_0-\Pi_n\alpha_0\right\|^2=o_p\left(n^{-1/2}\right)$.

Finally, condition D in Shen (1997), i.e.,

$$\sup_{\alpha\in\mathcal{N}_{0n}}\mu_n\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)=o_p\left(n^{-1/2}\right), \qquad (102)$$

can be verified as follows: We first have

$$\begin{aligned}
& \sup_{\alpha\in\mathcal{N}_{0n}}\left|\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right| \qquad (103)\\
\leq & \left|\frac{1}{f_{yx|z}(D,\alpha_0)}\int\frac{d}{d\theta}f_{y|x^*}(y|x^*;\theta_0)\omega^{-1}\left(\xi\right)f_{x|x^*}(x|x^*)f_{x^*|z}(x^*|z)dx^*\right|\left\|\theta-\theta_0\right\|_s+\\
& +\left|\frac{1}{f_{yx|z}(D,\alpha_0)}\int f_{y|x^*}(y|x^*;\theta_0)\omega^{-1}\left(x,x^*\right)f_{x^*|z}(x^*|z)dx^*\right|\left\|f_1-f_{x|x^*}\right\|_s+\\
& +\left|\frac{1}{f_{yx|z}(D,\alpha_0)}\int f_{y|x^*}(y|x^*;\theta_0)f_{x|x^*}(x|x^*)\omega^{-1}\left(x^*,z\right)dx^*\right|\left\|f_2-f_{x^*|z}\right\|_s\\
\leq & \left|\frac{f_{yx|z}^{|1|}\left(D,\alpha_0,\bar{\omega}\right)}{f_{yx|z}(D,\alpha_0)}\right|\left\|\alpha-\alpha_0\right\|_s\\
\leq & \left|h_1(D)\right|\left\|\alpha-\alpha_0\right\|_s
\end{aligned}$$

with $E\left|h_1(D)\right|^2<\infty$ by the envelope condition in assumption 11. We then have

$$\begin{aligned}
& \sup_{\alpha\in\mathcal{N}_{0n}}\mu_n\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right) \qquad (104)\\
= & \sqrt{E\left(\sup_{\alpha\in\mathcal{N}_{0n}}\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right]\right)^2}O_p\left(n^{-1/2}\right)\\
\leq & \sqrt{E\left|h_1(D)\right|^2}\left\|\alpha-\alpha_0\right\|_sO_p\left(n^{-1/2}\right)\\
= & o_p\left(n^{-1/2}\right).
\end{aligned}$$

Thus, condition D in theorem 1 in Shen (1997) holds. Since all the conditions in theorem 1 in Shen (1997) hold, the results of asymptotic normality follow. ■

C Restrictions with Fourier series

As shown above, the sieve estimators are as follows:

$$f_1(x|x^*) = \sum_{i=0}^{i_n}\sum_{j=0}^{j_n} \beta_{ij} p_i(x-x^*) q_j(x^*), \quad f_2(x^*|z) = \sum_{i=0}^{i_n}\sum_{j=0}^{j_n} \gamma_{ij} p_i(x^*-z) q_j(z). \tag{105}$$

Let $z, x^* \in [0, l_x]$ and $(x - x^*) \in [-l_e, l_e]$. We use the Fourier series:

$$\begin{aligned} p_k(x-x^*) &= \cos\frac{k\pi}{l_e}(x-x^*) \text{ or } \sin\frac{k\pi}{l_e}(x-x^*) \\ p_k(x^*-z) &= \cos\frac{k\pi}{l_x}(x^*-z) \text{ or } \sin\frac{k\pi}{l_x}(x^*-z) \end{aligned} \tag{106}$$

and $q_k(x) = \cos\frac{k\pi}{l_x}x$. For simplicity, we consider the case where $i_n = 3$ and $j_n = 2$. Longer series can be handled similarly. We have

$$\begin{aligned} f_1(x|x^*) &= \left(a_{00} + a_{01}\cos\frac{\pi}{l_x}x^* + a_{02}\cos\frac{2\pi}{l_x}x^*\right) \\ &\quad + \sum_{k=1}^{3}\left(a_{k0} + a_{k1}\cos\frac{\pi}{l_x}x^* + a_{k2}\cos\frac{2\pi}{l_x}x^*\right)\cos\frac{k\pi}{l_e}(x-x^*) \\ &\quad + \sum_{k=1}^{3}\left(b_{k0} + b_{k1}\cos\frac{\pi}{l_x}x^* + b_{k2}\cos\frac{2\pi}{l_x}x^*\right)\sin\frac{k\pi}{l_e}(x-x^*) \end{aligned} \tag{107}$$

Consider the restriction $\int_{\mathcal{X}} f_1(x|x^*)dx = 1$. We can show that

$$\int_{\mathcal{X}} f_1(x|x^*)dx = 2l_e\left(a_{00} + a_{01}\cos\frac{\pi}{l_x}x^* + a_{02}\cos\frac{2\pi}{l_x}x^*\right) \tag{108}$$

for all x^*. Therefore, $a_{00} = \frac{1}{2l_e}$and $a_{01} = a_{02} = 0$. We can similarly find the sieve expression of the function $f_2(x^*|z)$ satisfying $\int_{\mathcal{X}^*} f_2(x^*|z)dx^* = 1$.

Next, we consider the identification restrictions on $f_1(x|x^*)$. First, in the zero mode case, we have $\frac{\partial}{\partial x} f_1(x|x^*)\big|_{x=x^*} = 0$ for all x^* with

$$\left.\frac{\partial}{\partial x} f_1(x|x^*)\right|_{x=x^*} = \sum_{k=1}^{3}\frac{k\pi}{l_e}\left(b_{k0} + b_{k1}\cos\frac{\pi}{l_x}x^* + b_{k2}\cos\frac{2\pi}{l_x}x^*\right) \tag{109}$$

Thus, the restrictions on the coefficients are

$$\sum_{k=1}^{3} kb_{k0} = \sum_{k=1}^{3} kb_{k1} = \sum_{k=1}^{3} kb_{k2} = 0. \tag{110}$$

Second, if we make the zero mean assumption instead of the zero mode one, we have $\int_{\mathcal{X}}(x-x^*)f_1(x|x^*)dx=0$ for all x^* with

$$\int_{\mathcal{X}}(x-x^*)f_1(x|x^*)dx=\sum_{k=1}^{3}\left(b_{k0}+b_{k1}\cos\frac{\pi}{l_x}x^*+b_{k2}\cos\frac{2\pi}{l_x}x^*\right)\left(-\frac{2l_e^2}{k\pi}\left(-1\right)^k\right) \tag{111}$$

We have

$$\sum_{k=1}^{3}\frac{\left(-1\right)^k}{k}b_{k0}=\sum_{k=1}^{3}\frac{\left(-1\right)^k}{k}b_{k1}=\sum_{k=1}^{3}\frac{\left(-1\right)^k}{k}b_{k2}=0. \tag{112}$$

Third, if we make the zero median assumption, we have $\int_{\mathcal{X}\cap\{x<x^*\}}f_{x|x^*}(x|x^*)dx=\frac{1}{2}$ for all x^* with

$$\int_{\mathcal{X}\cap\{x<x^*\}}f_1(x|x^*)dx=\frac{1}{2}+\sum_{k=1}^{3}\left(b_{k0}+b_{k1}\cos\frac{\pi}{l_x}x^*+b_{k2}\cos\frac{2\pi}{l_x}x^*\right)l_e\frac{\left(-1\right)^k-1}{k\pi} \tag{113}$$

Therefore,

$$\sum_{k=1}^{3}\frac{\left(-1\right)^k-1}{k}b_{k0}=\sum_{k=1}^{3}\frac{\left(-1\right)^k-1}{k}b_{k1}=\sum_{k=1}^{3}\frac{\left(-1\right)^k-1}{k}b_{k2}=0 \tag{114}$$

Fourth, if x^* is the 100th percentile of $f_{x|x^*}$, we assume $(x-x^*)\in[-l_e,0]$. The sieve estimator of $f_1(x|x^*)$ is as follows:

$$\begin{aligned} f_1(x|x^*) &= \left(a_{00}+a_{01}\cos\frac{\pi}{l_x}x^*+a_{02}\cos\frac{2\pi}{l_x}x^*\right) \\ &\quad +\sum_{k=1}^{3}\left(a_{k0}+a_{k1}\cos\frac{\pi}{l_x}x^*+a_{k2}\cos\frac{2\pi}{l_x}x^*\right)\cos\frac{k\pi}{l_e}(x-x^*) \end{aligned} \tag{115}$$

The restriction $\int_{\mathcal{X}\cap\{x<x^*\}}f_{x|x^*}(x|x^*)dx=1$ for all x^* is equivalent to the restrictions $a_{00}=\frac{1}{l_e}$and $a_{01}=a_{02}=0$.