ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 76, No. 1 (January, 2008), 195–216

INSTRUMENTAL VARIABLE TREATMENT OF NONCLASSICAL MEASUREMENT ERROR MODELS

YINGYAO HU
Johns Hopkins University, Baltimore, MD 21218, U.S.A.

SUSANNE M. SCHENNACH
University of Chicago, Chicago, IL 60637, U.S.A.

Econometrica, Vol. 76, No. 1 (January, 2008), 195–216

INSTRUMENTAL VARIABLE TREATMENT OF NONCLASSICAL MEASUREMENT ERROR MODELS

BY YINGYAO HU AND SUSANNE M. SCHENNACH[1]

While the literature on nonclassical measurement error traditionally relies on the availability of an auxiliary data set 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 "index" 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.

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

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, 1998, 2001), Lewbel (1996, 1998), Hausman (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 of 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), Newey (2001), Schennach (2007)), repeated measurements (Hausman, Newey, Ichimura, and Powell (1991), Li (2002), Schennach (2004a, 2004b)), or validation data (Hu and Ridder (2004)) are available.

However, there 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

[1]S. M. Schennach acknowledges support from the National Science Foundation via Grant SES-0452089. The authors would like to thank Lars Hansen, James Heckman, Marine Carrasco, Maxwell Stinchcombe, and Xiaohong Chen, as well as seminar audiences at various universities, at the Cemmap/ESRC Econometric Study Group Workshop on Semiparametric Methods, and at the Econometric Society 2006 Winter Meetings for helpful comments.

variable estimators that are robust to the presence of such "nonclassical" measurement error have been developed for binary regressors (Mahajan (2006), Lewbel (2007)) and general discrete regressors (Hu (2007)). 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-posed 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 data set containing correctly measured observations (Chen, Hong, and Tamer (2005), Chen, Hong, and Tarozzi (2008)). 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 data sets.

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).[2] 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 data set, 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 (2008)).

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. Intuitively, since there is only one way to report the truth, while there are an infinite number of alternative ways to misreport, respondents would literally have to collude on misreporting in a similar way to violate the mode assumption. In addition,

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

data truncation usually preserves the mode, but not the mean, provided the truncation is not so severe that the mode itself is deleted.

Second, if respondents are equally likely to over- or underreport, 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 underreported and, in that case, the topmost quantile of the observed consumption conditional on the truth would plausibly equal true consumption.

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) that characterizes 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.

2. IDENTIFICATION

The "true" model is defined by the joint distribution of the dependent variable y and the true regressor 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 the joint

distribution of x^* and y is identified from knowledge of the distribution of all observed variables. 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.

Let $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{X}^*$, and $\mathcal{Z}$ denote the supports of the distributions of the random variables y, x, x^*, and z, respectively. We consider x, x^*, and z to be jointly continuously distributed ($\mathcal{X}$, $\mathcal{X}^* \subset \mathbb{R}^{n_x}$ and $\mathcal{Z} \subset \mathbb{R}^{n_z}$ with $n_z \geq n_x$), while y can be either continuous or discrete. Accordingly, we assume the following.

ASSUMPTION 1: *The joint density of y and x, x^*, z admits a bounded density with respect to the product measure of some dominating measure μ (defined on $\mathcal{Y}$) and the Lebesgue measure on $\mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$. All marginal and conditional densities are also bounded.*

We use the notation $f_a(a)$ and $f_{a|b}(a|b)$ to denote the density of variable a and the density of a conditional on b, respectively. Implicitly, these densities are relative to the relevant dominating measure, as described above. For simplicity, our notation does not distinguish between a random variable and a specific value it may take. The joint support of all the variables need not be rectangular, since we allow for vanishing densities.

To state our identification result, we start by making natural assumptions regarding the conditional densities of all the variables of the model.

ASSUMPTION 2: (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* $(x, x^*, z) \in \mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$.

Assumption 2(i) indicates that x and z do not provide any more information about y than x^* already provides, while Assumption 2(ii) specifies that z does not provide any more information about x than x^* already provides. These assumptions can be interpreted as standard exclusion restrictions. Conditional independence restrictions have been widely used in the recent econometrics literature (e.g., Hoderlein and Mammen (2007), Heckman and Vytlacil (2005), Altonji and Matzkin (2005)).

REMARK: 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 2. 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 suggested by 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.

To facilitate the statement of our next assumption, is it useful to note that a function of two variables can be associated with an integral operator.

DEFINITION 1: Let a and b denote random variables with respective supports $\mathcal{A}$ and $\mathcal{B}$. Given two corresponding spaces $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$ of functions with domains $\mathcal{A}$ and $\mathcal{B}$, respectively, let $L_{b|a}$ denote the operator mapping $g \in \mathcal{G}(\mathcal{A})$ to $L_{b|a}g \in \mathcal{G}(\mathcal{B})$ defined by

$$[L_{b|a}g](b) \equiv \int_{\mathcal{A}} f_{b|a}(b|a)g(a)\,da,$$

where $f_{b|a}(b|a)$ denotes the conditional density of b given a.

For the density $f_{b|a}(b|a)$ to be uniquely determined by the operator $L_{b|a}$, the space $\mathcal{G}(\mathcal{A})$ upon which the operator acts must be sufficiently large so that $f_{b|a}(b|a)$ is "sampled" everywhere. For an integral operator, it is sufficient to consider $\mathcal{G}(\mathcal{A})$ to be $\mathcal{L}^1(\mathcal{A})$, the set of all absolutely integrable functions with domain $\mathcal{A}$ (endowed with the norm $\|g\|_1 = \int_{\mathcal{A}} |g(a)|\,da$). It is even sufficient to limit $\mathcal{G}(\mathcal{A})$ to the set of functions in $\mathcal{L}^1(\mathcal{A})$ that are also bounded ($\sup_{a\in\mathcal{A}} |g(a)| < \infty$), denoted $\mathcal{L}^1_{\text{bnd}}(\mathcal{A})$.[3] In our subsequent treatment, we will consider the cases where $\mathcal{G} = \mathcal{L}^1$ or where $\mathcal{G} = \mathcal{L}^1_{\text{bnd}}$. We can then state our next assumption.

ASSUMPTION 3: *The operators $L_{x|x^*}$ and $L_{z|x}$ are injective* (*for either $\mathcal{G} = \mathcal{L}^1$ or $\mathcal{G} = \mathcal{L}^1_{\text{bnd}}$*).

An operator $L_{b|a}$ is said to be *injective* if its inverse $L^{-1}_{b|a}$ is defined over the range of the operator $L_{b|a}$ (see Section 3.1 in Carrasco, Florens, and Renault (2005)). The qualification on the range is needed to account for the fact that inverses are often defined only over a restricted domain in infinite-dimensional spaces. Assumption 3 could also be stated in terms of the injectivity of $L_{z|x^*}$ and $L_{x|x^*}$, since it can be shown that injectivity of $L_{z|x^*}$ and $L_{x|x^*}$ implies injectivity of $L_{z|x}$.

[3]This can be seen from the fact that

$$f_{b|a}(b|a_0) = \lim_{n\to\infty} [L_{b|a}g_{n,a_0}](b),$$

where $g_{n,a_0}(a) = n1(|a - a_0| \le n^{-1})$, a sequence of absolutely integrable and bounded functions (the limit of that sequence does not need to belong to $\mathcal{G}(\mathcal{A})$, since we are *not* calculating $L_{b|a}\lim_{n\to\infty} g_{n,a_0}$). The so-called kernel $f_{b|a}(b|a_0)$ of the integral operator $L_{b|a}$ is therefore uniquely determined by evaluating this limit for all values of $a_0 \in \mathcal{A}$. It is also straightforward to check that for a bounded $f_{b|a}(b|a)$, $g \in \mathcal{L}^1(\mathcal{A})$ implies $L_{b|a}g \in \mathcal{L}^1(\mathcal{B})$ and that $g \in \mathcal{L}^1_{\text{bnd}}(\mathcal{A})$ implies $L_{b|a}g \in \mathcal{L}^1_{\text{bnd}}(\mathcal{B})$. Indeed, $\|L_{b|a}g\|_1 \le \int\int f_{b|a}(b|a)\,db\,|g(a)|\,da = \int 1|g(a)|\,da = \|g\|_1$ and $\sup_{b\in\mathcal{B}} |[L_{b|a}g](b)| \le \int (\sup_{\tilde{b}\in\mathcal{B}} \sup_{\tilde{a}\in\mathcal{A}} |f_{b|a}(\tilde{b}|\tilde{a})|)|g(a)|\,da = (\sup_{\tilde{b}\in\mathcal{B}} \sup_{\tilde{a}\in\mathcal{A}} |f_{b|a}(\tilde{b}|\tilde{a})|)\|g\|_1$.

Intuitively, an operator $L_{b|a}$ will be injective if there is enough variation in the density of b for different values of a. For instance, a simple case where $L_{b|a}$ is not injective is when $f_{b|a}(b|a)$ is a uniform density on $\mathcal{B}$ for any $a \in \mathcal{A}$. In general, however, injectivity assumptions are quite weak and are commonly made in the literature on nonparametric instrumental variable methods. They are sometimes invoked by assuming that an operator $L_{b|a}$ admits a singular value decomposition with nonzero singular values (Darolles, Florens, and Renault (2002)) or by stating that an operator is nonsingular (Horowitz (2006), Hall and Horowitz (2005)).

Injectivity assumptions are often phrased in terms of *completeness* (or *bounded completeness*) of the family of distributions that play the role of the kernel of the integral operator considered (Newey and Powell (2003), Blundell, Chen, and Kristensen (2007), Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2007)). This characterization is worth explaining in more detail, as it leads to primitive sufficient conditions. Formally, a family of distribution $f_{a|b}(a|b)$ is complete if the only solution $\tilde{g}(a)$ to

$$\int_{\mathcal{A}} \tilde{g}(a) f_{a|b}(a|b)\, da = 0 \quad \text{for all } b \in \mathcal{B} \tag{1}$$

(among all $\tilde{g}(a)$ such that (1) is defined) is $\tilde{g}(a) = 0$. Under Assumption 1, this condition implies injectivity of $L_{b|a}$ (viewed as a mapping from $\mathcal{L}^1(\mathcal{A})$ to $\mathcal{L}^1(\mathcal{B})$). Indeed, $\int f_{a|b}(a|b)\tilde{g}(a)\, da = (f_b(b))^{-1} \int f_{b|a}(b|a) f_a(a)\tilde{g}(a)\, da$, and since $0 < f_a(a) < \infty$ and $0 < f_b(b) < \infty$ over the interior of their respective supports, having $\tilde{g}(a) = 0$ as the unique solution is equivalent to having $g(a) = 0$ as the unique solution to $\int f_{b|a}(b|a) g(a)\, da = 0$. If $g(a) = 0$ is the unique solution among all $g(a)$ such that the integral is defined, then it is also the unique solution in $\mathcal{L}^1(\mathcal{A})$, which implies that $L_{b|a}$ is injective. Bounded completeness is similarly defined by stating that the only solution to (1) among all *bounded* $\tilde{g}(a)$ is $\tilde{g}(a) = 0$. Analogously, this implies that $L_{b|a}$ is injective when viewed as a mapping from $\mathcal{L}^1_{\text{bnd}}(\mathcal{A})$ to $\mathcal{L}^1_{\text{bnd}}(\mathcal{B})$.

A nice consequence of the connection between injectivity and (bounded) completeness is that primitive conditions for (bounded) completeness are readily available in the literature. For instance, some very general exponential families of distributions are known to be complete (as invoked in Newey and Powell (2003)). The weaker notion of bounded completeness can also be used to find even more general families of distributions leading to injective operators (as discussed in Blundell, Chen, and Kristensen (2007)). In particular, when $f_{a|b}(a|b)$ can be written in the form $f_\varepsilon(a - b)$, then $L_{b|a}$ is injective if and only if the Fourier transform of f_ε is everywhere nonvanishing (by Theorem 2.1 in Mattner (1993)), and similar results have also been obtained for more general families of distributions that cannot be written as $f_\varepsilon(a - b)$ (d'Haultfoeuille (2006)).

The assumption of injectivity of $L_{x|x^*}$ allows for x^* and x to be multivariate. Injectivity of $L_{z|x}$ in multivariate settings is also natural whenever the dimension of z is greater or equal to the dimension of x. 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.

While Assumption 3 places restrictions on the relationships between z, x, and x^*, the following assumption places restrictions on the relationship between y and x^*.

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

This assumption is even weaker than injectivity. It is automatically satisfied if $E[y|x^*]$ is strictly monotone (for univariate x^*), but also holds far more generally. The presence of conditional heteroskedasticity can be sufficient in the absence of monotonicity. Assumption 4 is only violated if the distribution of y conditional on x^* is identical at two values of x^*.

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 (or when $P[y = 0|x^*]$ is not monotone), it will be necessary to define y to be a vector that contains auxiliary variables in addition to the binary outcome 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.

We then characterize the nature of measurement error via an assumption that considerably generalizes the case of classical measurement error.

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

M is a very general functional that maps a 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, and the τ quantile, corresponding to the following definitions of M, respectively:

$$M[f] = \int_{\mathcal{X}} x f(x)\, dx, \tag{2}$$

$$M[f] = \arg\max_{x \in \mathcal{X}} f(x), \tag{3}$$

$$M[f] = \inf\left\{x^* \in \mathcal{X}^* : \int 1(x \le x^*) f(x)\, dx \ge \tau\right\}. \tag{4}$$

Case (2) above covers classical measurement error (in which $x = x^* + \varepsilon$, where $E[\varepsilon|x^*] = 0$), since $M[f_{x|x^*}(\cdot|x^*)] = E[x|x^*] = E[x^* + \varepsilon|x^*] = x^* + E[\varepsilon|x^*] = x^*$ in that case. The other two examples of M cover nonclassical measurement error of various forms. For multivariate x, (2) and (3) apply directly, while (4) could then take the form of a vector of univariate marginal quantiles, for instance.

It should be noted that Assumptions 1–5 are not mutually contradictory: Models that satisfy all of them can easily be constructed. For instance, one can set $f_{xyzx^*}(x, y, z, x^*) = f_{x|x^*}(x|x^*) f_{y|x^*}(y|x^*) f_{z|x^*}(z|x^*) f_{x^*}(x^*)$, where $f_{x^*}(x^*)$ is a normal, and where $f_{x|x^*}(x|x^*)$, $f_{y|x^*}(y|x^*)$, and $f_{z|x^*}(z|x^*)$ each are homoskedastic normals whose means depend linearly on x^* (with nonzero slope) and such that $E[x|x^*] = x^*$. We are now ready to state our main result.

THEOREM 1: *Under Assumptions 1–5, given the true observed density $f_{yx|z}$, the equation*

$$f_{yx|z}(y, x|z) = \int_{\mathcal{X}^*} f_{y|x^*}(y|x^*) f_{x|x^*}(x|x^*) f_{x^*|z}(x^*|z)\, dx^* \tag{5}$$
$$\text{for all } y \in \mathcal{Y}, x \in \mathcal{X}, z \in \mathcal{Z},$$

admits a unique solution[4] *$(f_{y|x^*}, f_{x|x^*}, f_{x^*|z})$. A similar result holds for*

$$f_{yxz}(y, x, z) = \int_{\mathcal{X}^*} f_{yx^*}(y, x^*) f_{x|x^*}(x|x^*) f_{z|x^*}(z|x^*)\, dx^* \tag{6}$$
$$\text{for all } y \in \mathcal{Y}, x \in \mathcal{X}, z \in \mathcal{Z}.$$

The proof can be found in the Appendix and can be outlined as follows. Assumption 2 lets us obtain the integral Equation (5) that relates the joint densities of the observable variables to the joint densities of the unobservable variables. This equation is then shown to define the operator equivalence relationship

$$L_{y;x|z} = L_{x|x^*} \Delta_{y;x^*} L_{x^*|z}, \tag{7}$$

where $L_{y;x|z}$ is defined analogously to $L_{x|z}$ with $f_{x|z}$ replaced by $f_{y,x|z}$ for a given $y \in \mathcal{Y}$ and where $\Delta_{y;x^*}$ is the "diagonal" operator mapping the function $g(x^*)$ to the function $f_{y|x^*}(y|x^*)g(x^*)$, for a given $y \in \mathcal{Y}$. Next, we note that the equivalence $L_{x|z} = L_{x|x^*} L_{x^*|z}$ also holds (by integration of (7) over all $y \in \mathcal{Y}$). Isolating $L_{x^*|z}$ to yield

$$L_{x^*|z} = L_{x|x^*}^{-1} L_{x|z}, \tag{8}$$

[4]More formally, if multiple solutions exist, they differ only on a set of zero probability.

substituting it into (7), and rearranging, we obtain

$$L_{y;x|z}L_{x|z}^{-1} = L_{x|x^*}\Delta_{y;x^*}L_{x|x^*}^{-1}, \tag{9}$$

where all inverses can be shown to exist over suitable domains by Assumption 3 and Lemma 1 in the Appendix. Equation (9) states that the operator $L_{y;x|z}L_{x|z}^{-1}$ admits a spectral decomposition (specifically, an eigenvalue–eigenfunction decomposition in this case). The operator to be diagonalized is defined in terms of observable densities, while the resulting eigenvalues $f_{y|x^*}(y|x^*)$ and eigenfunctions $f_{x|x^*}(\cdot|x^*)$ (both indexed by $x^* \in \mathcal{X}^*$) provide the unobserved densities of interest. To ensure uniqueness of this decomposition, we employ four techniques. First, a powerful result from spectral analysis (Theorem XV.4.5 in Dunford and Schwartz (1971)) ensures uniqueness up to some normalizations. Second, the a priori arbitrary scale of the eigenfunctions is fixed by the requirement that densities must integrate to 1. Third, to avoid any ambiguity in the definition of the eigenfunctions when degenerate eigenvalues are present, we use Assumption 4 and the fact that the eigenfunctions (which do not depend on y, unlike the eigenvalues $f_{y|x^*}(y|x^*)$) must be consistent across different values of the dependent variable y. Finally, to uniquely determine the ordering and indexing of the eigenvalues and eigenfunctions, we invoke Assumption 5: If one considers another variable $\tilde{x}^*$ related to x^* through $x^* = R(\tilde{x}^*)$, we have

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

which is only equal to $\tilde{x}^*$ if R is the identity function. These four steps ensure that the diagonalization operation uniquely specifies the unobserved densities $f_{y|x^*}(y|x^*)$ and $f_{x|x^*}(x|x^*)$ of interest. Next, Equation (8) implies that $f_{x^*|z}(x^*|z)$ is also identified. Since the identities (9) and (8) use and provide the same information as Equation (5), this establishes uniqueness of the solution to Equation (5). The second conclusion of the theorem (Equation (6)) follows by similar manipulations.

It is possible to replace $f_{y,x|z}(y, x|z)$ by $E[y|x, z]f_{x|z}(x|z)$ and $f_{y|x^*}(y|x^*)$ by $E[y|x^*]$ throughout to obtain an identification result for $E[y|x^*]$ directly, without fully identifying $f_{y|x^*}(y|x^*)$. This would slightly weaken Assumption 2(i) to $E[y|x, x^*, z] = E[y|x^*]$. However, under this approach, the analogues of Assumptions 1 and 4 would become somewhat restrictive for univariate y and x^*, requiring $E[y|x^*]$ to be strictly monotone in x^* and such that $\sup_{x^* \in \mathcal{X}^*} |E[y|x^*]| < \infty$. These restrictions are avoided if identification of $E[y|x^*]$ is secured through the identification of $f_{y|x^*}(y|x^*)$.

While Theorem 1 establishes identification, we can also show that the model is actually overidentified, thus permitting a test of the model. Equation (5) relates a function of three variables to a triplet of functions of two variables. Since the set of functions of three variables is much "larger" than the set of triplets of functions of two variables, there exist densities $f_{yx|z}(y, x|z)$ that cannot be generated by Equation (5), a telltale sign of an overidentifying restriction. The

availability of more than one valid instrument offers further opportunities to test the model's assumptions.

3. ESTIMATION USING SIEVE MAXIMUM LIKELIHOOD

As a starting point, we consider a model expressed in terms of the observed variable y and the unobserved mismeasured regressor x^*:

$$f_{y|x^*}(y|x^*;\theta). \tag{10}$$

It is often convenient to decompose the potentially infinite-dimensional parameter θ that we seek to estimate into two subvectors: b, a finite-dimensional parameter vector of interest, and η, a potentially infinite-dimensional nuisance parameter. Naturally, we assume that the parametrization (10) does not include redundant degrees of freedom, that is, $\theta \equiv (b,\eta)$ is identified if $f_{y|x^*}$ is identified.

This framework nests most commonly used models as subcases. First, setting $\theta \equiv b$ covers the parametric likelihood case (which will then become semiparametric once we account for measurement error). Second, models defined via conditional moment restrictions $E[m(y,x^*,b)|x^*]=0$ can be considered by defining a family of densities $f_{y|x^*}(y|x^*;b,\eta)$ such that $\int f_{y|x^*}(y|x^*;b,\eta)m(y,x^*,b)\,dy=0$ for all b and η, which is clearly equivalent to imposing a moment condition. For example, in a nonlinear regression model $y=g(x^*,b)+\varepsilon$ with $E(\varepsilon|x^*)=0$, we have $f_{y|x^*}(y|x^*;b,\eta)=f_{\varepsilon|x^*}(y-g(x^*,b)|x^*)$. The infinite-dimensional nuisance parameter η is the conditional density $f_{\varepsilon|x^*}(\cdot|\cdot)$, constrained to have zero mean. Another important example is the quantile regression case[5] (where the conditional density $f_{\varepsilon|x^*}(\cdot|\cdot)$ is constrained to have its conditional τ-quantile equal to 0). Quantile restrictions are useful, as they provide the fundamental concept that enables a natural treatment of nonseparable models (e.g., Chernozhukov, Imbens, and Newey (2007), Matzkin (2003), Chesher (2003)). More generally, our framework also covers most semiparametric setups. For instance, one could devise a family of densities $f_{y|x^*}(y|x^*;b,\eta)$ such that b sets the value of the average derivative $\int(dE[y|x^*]/dx^*)w(x^*)\,dx^*$ (for some weighting function $w(x^*)$), while η controls all remaining degrees of freedom that affect the shape of the density but that do not affect the value of the average derivative. More examples of a partition of θ into b and η can be found in Shen (1997).

Given a model expressed in terms of the true unobserved variables (10), Equation (5) in Theorem 1 suggests a corresponding measurement-error robust sieve maximum likelihood estimator (e.g., Grenander (1981), Shen (1997),

[5]The nonsmoothness of the moment conditions in this case does not pose special problems, because all quantities are effectively smoothed by the truncated series used to represent all densities.

Chen and Shen (1998), Ai and Chen (2003)):

$$(\hat{\theta}, \hat{f}_1, \hat{f}_2) = \arg \max_{(\theta, f_1, f_2) \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^{n} \ln \int_{\mathcal{X}^*} f_{y|x^*}(y_i|x^*; \theta) f_1(x_i|x^*) f_2(x^*|z_i)\, dx^*. \tag{11}$$

Here, $(x_i, y_i, z_i)_{i=1}^n$ is an independent and identically distributed (i.i.d.) sample and $\mathcal{A}_n$ is a sequence of approximating sieve spaces that contain progressively more flexible parametric approximations to the densities (as sample size n increases). While Equation (11) enforces Assumption 2 by construction, functions in $\mathcal{A}_n$ are required to satisfy Assumption 5 as well as normalizations that ensure that the relevant conditional densities suitably integrate to 1. The remaining assumptions made in the identification theory are regularity conditions that the generating process is assumed to satisfy but that do not need to be imposed in the estimation procedure. Typically, the approximating spaces $\mathcal{A}_n$ are generated by the span of series approximations that are linear in the coefficients, such as polynomials, splines, and so forth. In this case, all restrictions on $\mathcal{A}_n$ imposed by the original semiparametric model or by Assumption 5 can be easily implemented, since they amount to imposing linear restrictions on the coefficients of the sieve that represent the unknown densities.

The supplementary material available on the *Econometrica* website (Hu and Schennach (2008)) fully develops the asymptotic theory of the proposed sieve maximum likelihood estimator. A nonparametric consistency result (in a weighted sup norm) is provided as well as a semiparametric root n consistency and asymptotic normality result for the estimated parametric component b of the parameter θ. Our treatment allows for the support of all variables y, x^*, x, z to be unbounded. For the purposes of simplicity and conciseness, our treatment provides primitive sufficient conditions for the independent and identically distributed case. However, 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 that include dependent and nonidentically distributed data. The regularity conditions invoked for the asymptotic theory fall into three general classes:

(i) Smoothness and boundedness restrictions that limit the "size" of the space of functions considered so as to obtain the compactness of the parameter space (where, here, the parameters include functions) that is traditionally invoked to show consistency.

(ii) Envelope conditions that limit how rapidly the objective function can change in value as the parameters change; this helps secure stochastic equicontinuity and uniform convergence results.

(iii) Sieve approximation rates (i.e., at what rate must the number of terms in the series increase to guarantee a given rate of decay of the approximation error?).

The practical implementation of the method requires the selection of the number of terms in the various approximating series. While a formal selection rule for these smoothing parameters (e.g., based on a higher-order asymptotic analysis) would be desirable, it is beyond the scope of the present paper. Some informal guidelines can nevertheless be given. In our semiparametric setting, selection of the smoothing parameters is somewhat facilitated (relative to fully nonparametric approaches), because semiparametric estimators are known to have the same asymptotic distribution for a wide range of smoothing parameter sequences. This observation suggests that a valid smoothing parameter can be obtained by scanning a range of values in search of a region where the estimates are not very sensitive to small variations in the smoothing parameter. Typically, for very short series, the smoothing bias dominates and the estimates will exhibit a marked trend as the number of terms is increased. At the other extreme, for very long series, the statistical noise dominates and, although the point estimates vary significantly as additional terms are added, no clear trend should be visible. In between those extremes should lie a region where any clear trend has leveled off and where the random noise in the estimates has not yet grown to an excessive level. The middle of that region points to a suitable value of the smoothing parameters.

A number of straightforward extensions of the above approach are possible. First, the model specified in (10) also could be conditional on any number of other, correctly measured, variables. The same identification proof and estimation method follow, after conditioning all densities on those variables.

The second conclusion of Theorem 1 also suggests an alternative expression for the observed density which proves useful if the model specifies $f_{yx^*}(y, x^*)$ instead of $f_{y|x^*}(y|x^*)$. Our sieve approach, now based on a likelihood expressed in terms of $f_{yxz}(y, x, z)$, covers this case as well. This also enables the treatment of models defined via unconditional moment restrictions (i.e., $E[m(y, x^*, b)] = 0$).

4. SIMULATIONS

This section investigates the performance of the proposed estimator with simulated data. We consider a simple parametric probit model

$$f_{y|x^*}(y|x^*) = [\Phi(a + bx^*)]^y[1 - \Phi(a + bx^*)]^{1-y} \quad \text{for} \quad y \in \mathcal{Y} = \{0, 1\},$$

where (a, b) is the unknown parameter vector and $\Phi(\cdot)$ is the standard normal cumulative distribution function (c.d.f.). In the simulations, we generate the instrumental variable and the latent 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 distributions of both z and η are truncated on $[0, 2]$ for simplicity in the implementation. To illustrate our method's ability to handle a variety of assumptions regarding

the measurement error, our examples of generating processes have the general form

$$f_{x|x^*}(x|x^*) = \frac{1}{\sigma(x^*)} f_\nu\left(\frac{x - x^*}{\sigma(x^*)}\right),$$

where f_ν is a density function that will be specified in each example below. We allow for considerable heteroskedasticity, setting $\sigma(x^*) = 0.5\exp(-x^*)$ in all examples. Sieves for functions of two variables are constructed through tensor product bases of univariate trigonometric series. We let i_n and j_n denote the number of terms taken from each of the two series. The smoothing parameters were determined following the guidelines given in the previous section, by locating the middle of a range of values of i_n and j_n over which the point estimates are relatively constant.

We consider three maximum likelihood estimators: (i) the (inconsistent) estimator obtained when ignoring measurement error, (ii) the (infeasible) estimator obtained using error-free data, and (iii) the proposed (consistent and feasible) sieve maximum likelihood estimator. We consider models where (i) the mode of f_ν is at zero, (ii) the median of f_ν is at zero, and (iii) the 100th percentile of f_ν is at zero. The supplementary material (Hu and Schennach (2008)) presents additional simulation examples.

The simulation results (see Table I) show that our proposed estimator performs well under a variety of identification conditions. The sieve estimator has a considerably smaller bias than the estimator ignoring the measurement error. As expected, the sieve estimator has a larger variance than the other two estimators, due to the estimation of nonparametric components. However, the sieve estimator still achieves a reduction in the overall root mean square error (RMSE), relative to the other feasible estimator.

5. CONCLUSION

This paper represents the first treatment of a wide class of nonclassical nonlinear errors-in-variables models with continuously distributed variables using instruments (or repeated measurements). 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. 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.

This nonparametric identification result suggests a natural sieve-based semiparametric maximum likelihood estimator that is relatively simple to implement. Our framework enables the construction of measurement-error-robust

TABLE I

SIMULATION RESULTS[a]

	Parameter (=True Value)					
	$a=-1$			$b=1$		
	Mean	Std. Dev.	RMSE	Mean	Std. Dev.	RMSE
Error distribution (zero mode): $f_\nu(\nu)=\exp[\nu-\exp(\nu)]$						
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						
Error distribution (zero median): $f_\nu(\nu)=\frac{1}{\pi}(1+[\frac{1}{2}+\frac{1}{2}\exp(\nu)-\exp(-\nu)]^2)^{-1}$						
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						
Error distribution (100th percentile at zero): $f_\nu(\nu)=\exp(\nu)$ for $\nu\in[-\infty,0]$						
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						

[a]For each estimator, we report the mean, the standard deviation (std. dev.), and the square root of the mean squared error (RMSE) of the estimators averaged over all 1,000 replications. The sample size is 2,000.

counterparts of parametric likelihood or moment conditions models, as well as numerous semiparametric models. Our semiparametric estimator is shown to be root n consistent and asymptotically normal.

Dept. of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, U.S.A.; yhu@jhu.edu

and

Dept. of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, U.S.A.; smschenn@uchicago.edu.

Manuscript received June, 2006; final revision received September, 2007.

APPENDIX: PROOFS

PROOF OF THEOREM 1: By the definition of conditional densities and Assumption 2,

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

$$= \int f_{y|x^*}(y|x^*) f_{xx^*|z}(x, x^*|z)\, dx^*$$

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

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

This establishes Equation (5) of the theorem. We now show the uniqueness of the solution.

Let the operators $L_{x|z}$, $L_{x|x^*}$, and $L_{x^*|z}$, be given by Definition 1, and let

$$L_{y;x|z}: \quad \mathcal{G}(\mathcal{Z}) \mapsto \mathcal{G}(\mathcal{X}) \quad \text{with} \quad L_{y;x|z} g \equiv \int f_{yx|z}(y, \cdot|z) g(z)\, dz,$$

$$\Delta_{y;x^*}: \quad \mathcal{G}(\mathcal{X}^*) \mapsto \mathcal{G}(\mathcal{X}^*) \quad \text{with} \quad \Delta_{y;x^*} g \equiv f_{y|x^*}(y|\cdot) g(\cdot).$$

The notation $L_{y;x|z}$ 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. The $\Delta_{y;x^*}$ operator is a "diagonal" operator since it is just a multiplication by a function (for a given y), that is, $[\Delta_{y;x^*} g](x^*) = f_{y|x^*}(y|x^*) g(x^*)$. By calculating $L_{y;x|z} g$ for an arbitrary $g \in \mathcal{G}(\mathcal{Z})$, we rewrite Equation (5) as an operator equivalence relationship,

$$[L_{y;x|z} g](x) = \int f_{yx|z}(y, x|z) g(z)\, dz \tag{12}$$

$$= \int \int f_{yx,x^*|z}(y, x, x^*|z)\, dx^* g(z)\, dz$$

$$= \int \int f_{x|x^*}(x|x^*) f_{y|x^*}(y|x^*) f_{x^*|z}(x^*|z)\, dx^* g(z)\, dz$$

$$= \int f_{x|x^*}(x|x^*) f_{y|x^*}(y|x^*) \int f_{x^*|z}(x^*|z) g(z)\, dz\, dx^*$$

$$= \int f_{x|x^*}(x|x^*) f_{y|x^*}(y|x^*) [L_{x^*|z} g](x^*)\, dx^*$$

$$= \int f_{x|x^*}(x|x^*) [\Delta_{y;x^*} L_{x^*|z} g](x^*)\, dx^*$$

$$= [L_{x|x^*} \Delta_{y;x^*} L_{x^*|z} g](x),$$

where we have used (i) Equation (5), (ii) an interchange of the order of integration (justified by Fubini's theorem), (iii) the definition of $L_{x^*|z}$, (iv) the definition of $\Delta_{y;x^*}$ operating on the function $[L_{x^*|z} g]$, and (v) the definition of $L_{x|x^*}$ operating on the function $[\Delta_{y;x^*} L_{x^*|z} g]$.

Equation (12) thus implies the operator equivalence (which holds over the domain $\mathcal{G}(\mathcal{Z})$)

$$L_{y;x|z} = L_{x|x^*}\Delta_{y;x^*}L_{x^*|z}. \tag{13}$$

By integration over y and noting that $\int_{\mathcal{Y}} L_{y;x|z}\mu(dy) = L_{x|z}$ and $\int_{\mathcal{Y}} \Delta_{y;x^*}\mu(dy) = I$, the identity operator, we similarly get

$$L_{x|z} = L_{x|x^*}L_{x^*|z}. \tag{14}$$

Since $L_{x|x^*}$ is injective (by Assumption 3), Equation (14) can be written as

$$L_{x^*|z} = L_{x|x^*}^{-1}L_{x|z}. \tag{15}$$

The domain of the inverse is guaranteed to be dense in the range of $L_{x|z}$ because the results of the inversion $L_{x|x^*}^{-1}L_{x|z}$ yield a well-defined integral operator $L_{x^*|z}$. Moreover, the operator equivalence (15) holds for the same domain space $\mathcal{G}(\mathcal{Z})$ as in (14) because the inverse operator was applied from the left side of Equation (14). The expression (15) for $L_{x^*|z}$ can be substituted into Equation (13) to yield

$$L_{y;x|z} = L_{x|x^*}\Delta_{y;x^*}L_{x|x^*}^{-1}L_{x|z}. \tag{16}$$

As shown in Lemma 1 below, the fact that $L_{z|x}$ is injective (by Assumption 3) implies that the inverse $L_{x|z}^{-1}$ can be applied "from the right" on each side of Equation (16) to yield

$$L_{y;x|z}L_{x|z}^{-1} = L_{x|x^*}\Delta_{y;x^*}L_{x|x^*}^{-1}, \tag{17}$$

where the operator equivalence holds over a dense subset of the domain space $\mathcal{G}(\mathcal{X})$. The equivalence can then be extended to the whole domain space $\mathcal{G}(\mathcal{X})$ by the standard extension procedure for linear operators.

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 (17) states that the known operator $L_{y;x|z}L_{x|z}^{-1}$ admits a spectral decomposition that takes the form of an eigenvalue–eigenfunction decomposition.[6] The eigenvalues of the $L_{y;x|z}L_{x|z}^{-1}$ operator are given by the "diagonal elements" of the $\Delta_{y;x^*}$ operator (i.e., $\{f_{y|x^*}(y|x^*)\}$ 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 $L_{x|x^*}$, that is,

[6]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 $f_{x|x^*}(\cdot|x^*)$ belong to $\mathcal{G}(\mathcal{X})$ and 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).

$\{f_{x|x^*}(\cdot|x^*)\}$ for all x^*. To establish identification of the unobserved functions of interest $f_{y|x^*}(y|x^*)$ and $f_{x|x^*}(\cdot|x^*)$, we need to show that the decomposition (17) is unique.

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. In particular, if a bounded operator T can be written as $T = A + N$, where A is an operator of the form

$$A = \int_{\sigma} \lambda P(d\lambda), \tag{18}$$

where P is a projection-valued measure[7] 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 $f_{y|x^*}(y|x^*)$, that is, $\{f_{y|x^*}(y|x^*) : x^* \in \mathcal{X}^*\}$. Since the largest element of the spectrum is bounded (by Assumption 1), the operator T is indeed bounded in the sense required by Dunford and Schwartz's result.[8]

In our situation, the projection-valued measure P assigned to any subset Λ of $\mathbb{R}$ is

$$P(\Lambda) = L_{x|x^*} I_{\Lambda} L_{x|x^*}^{-1},$$

where the operator I_{Λ} is defined via

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

An equivalent way to define $P(\Lambda)$ is by introducing the subspace

$$\mathcal{S}(\Lambda) = \text{span}\{f_{x|x^*}(\cdot|x^*) : x^* \text{ such that } f_{y|x^*}(y|x^*) \in \Lambda\} \tag{19}$$

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

The fact that $\int_{\sigma} \lambda P(d\lambda) = L_{x|x^*} \Delta_{y;x^*} L_{x|x^*}^{-1}$, thus connecting Equation (17) with Equation (18), can be shown by noting that

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

[7] 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*, that is, $QQ = Q$.

[8] As explained in Section XV.4 of Dunford and Schwartz (1971).

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

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

where we have used the fact that the generalized differential of a step function $1(\lambda \leq 0)$ is a Dirac delta[9] $\delta(\lambda)$, as defined by the property that $\int \delta(\lambda - \lambda_0) h(\lambda)\, d\lambda = h(\lambda_0)$ for any function $h(\lambda)$ continuous at $\lambda = \lambda_0$ and, in particular, for $h(\lambda) = \lambda$. Hence, we can indeed conclude that $\int_\sigma \lambda P(d\lambda) = L_{x|x^*} \Delta_{y;x^*} L_{x|x^*}^{-1}$.

Having established uniqueness of the decomposition (18) does not yet imply that the representation (17) is unique. The situation is analogous to standard matrix diagonalization:

(i) Each eigenvalue λ is associated with a unique subspace $\mathcal{S}(\{\lambda\})$ for $\mathcal{S}(\cdot)$ as defined in Equation (19). 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.

(ii) 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. 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 illustrated in the supplementary material (Hu and Schennach (2008)).

Issue (i)(a) is avoided because the requirement that $\int f_{x|x^*}(x|x^*)\, dx = 1$ sets the scale of the eigenfunctions.

Issue (i)(b) above, is handled via Assumption 4. The idea is that the operator $L_{x|x^*}$ that defines the eigenfunctions does not depend on y, while the eigenvalues given by $f_{y|x^*}(y|x^*)$ do. Hence, if there is an eigenvalue degeneracy that involves 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

[9]This derivation can alternatively be written in terms of Lebesgue–Stieltjes integrals, which avoids the need to explicitly introduce delta functions, but this is notationally cumbersome.

ambiguity. Formally, this can be shown as follows. Consider a given eigenfunction $f_{x|x^*}(\cdot|x^*)$ and let $D(y, x^*) = \{\tilde{x}^* : f_{y|x^*}(y|\tilde{x}^*) = f_{y|x^*}(y|x^*)\}$, the set of other values of x^* that index 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 $v(x^*) \equiv \bigcap_{y\in\mathcal{Y}} \operatorname{span}(\{f_{x|x^*}(\cdot|\tilde{x}^*)\}_{\tilde{x}^*\in D(y,x^*)})$ 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 $v(x^*)$ is not 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 $\bigcap_{y\in\mathcal{Y}} D(y, x^*)$ 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 $y \in \mathcal{Y}$, thus violating Assumption 4. (The qualification that the set on which the densities differ must have positive probability merely accounts for the fact that densities that differ on a set of zero probability actually represent the same density.)

Next, Assumption 5 resolves the ordering/indexing ambiguity (issue (ii) above), because if one considers another variable $\tilde{x}^*$ related to x^* through $x^* = R(\tilde{x}^*)$, we have

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

which is only equal to $\tilde{x}^*$ if R is the identity function. Having shown that $f_{y|x^*}(y|x^*)$ and $f_{x|x^*}(x|x^*)$ are uniquely determined, we can then show that $f_{x^*|z}(x^*|z)$ is uniquely determined, since $L_{x^*|z} = L_{x|x^*}^{-1}L_{x|z}$, where $L_{x|x^*}$ is now known and where $L_{x|z}$ is also known because its kernel is an observed density.

The second conclusion of the theorem is obtained by noting that

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

and showing that $f_{x|x^*}$, f_{yx^*}, and $f_{z|x^*}$ are uniquely determined from f_{yxz}. First, we have already shown that $f_{x|x^*}(x|x^*)$ is identified from $f_{yx|z}(y, x|z)$ (and therefore from $f_{yxz}(y, x, z)$). By Equation (15), $f_{x^*|z}(x^*|z)$ is also identified. Next, $f_{x^*}(x^*) = \int f_{x^*|z}(x^*|z)f_z(z)\,dz$, where $f_z(z)$ is observed. Then $f_{z|x^*}(z|x^*) = f_{x^*|z}(x^*|z)f_z(z)/f_{x^*}(x^*)$ and, finally, $f_{y,x^*}(y, x^*) = f_{y|x^*}(y|x^*) \times f_{x^*}(x^*)$. Hence the solution to Equation (6) is unique. *Q.E.D.*

LEMMA 1: *Under Assumption 1, if $L_{z|x}$ is injective, then $L_{x|z}^{-1}$ exists and is densely defined over $\mathcal{G}(\mathcal{X})$ (for $\mathcal{G} = \mathcal{L}^1, \mathcal{L}^1_{\text{bnd}}$).*

PROOF: Under Assumption 1, injectivity of $L_{z|x}$ implies injectivity of $L_{x|z}^\dagger$, the adjoint of $L_{x|z}$. This follows from arguments similar to those given after Equation (1) and the fact that $g(\cdot)/f_x(\cdot) \in \mathcal{G}^*(\mathcal{X})$, where $\mathcal{G}^*(\mathcal{X})$ denotes the dual space of $\mathcal{G}(\mathcal{X})$, implies that $g \in \mathcal{G}(\mathcal{X})$.

Next, $L_{x|z}$ can be shown to be injective when viewed as a mapping of $\overline{\mathcal{R}(L_{x|z}^\dagger)}$ into $\mathcal{G}(\mathcal{X})$, where $\overline{\mathcal{R}(L_{x|z}^\dagger)}$ denotes the closure in $\mathcal{G}(\mathcal{Z})$ of the range of $L_{x|z}^\dagger$. Indeed, by Lemma VI.2.8 in Dunford and Schwartz (1971), $\overline{\mathcal{R}(L_{x|z}^\dagger)}$ is the orthogonal complement of the null space of $L_{x|z}$, denoted $\mathcal{N}(L_{x|z})$. It follows that $L_{x|z}^{-1}$ exists.

By Lemma VI.2.8 in Dunford and Schwartz (1971) again, $\overline{\mathcal{R}(L_{x|z})}$ is the orthogonal complement of $\mathcal{N}(L_{x|z}^\dagger)$, but since $L_{x|z}^\dagger$ is injective, $\mathcal{N}(L_{x|z}^\dagger) = \{0\}$. Hence, $\overline{\mathcal{R}(L_{x|z})} = \mathcal{G}(\mathcal{X})$ and $L_{x|z}^{-1}$ is therefore defined on a dense subset of $\mathcal{G}(\mathcal{X})$. *Q.E.D.*

REFERENCES

AI, C., AND X. CHEN (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843. [205]

ALTONJI, J. G., AND R. L. MATZKIN (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1102. [198]

BLUNDELL, R., X. CHEN, AND D. KRISTENSEN (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669. [200]

BOLLINGER, C. R. (1998): "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 16, 576–594. [196]

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. [196]

BOUND, J., C. BROWN, AND N. MATHIOWETZ (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics*, Vol. 5, ed. by J. J. Heckman and E. Leamer. Amsterdam: Elsevier Science, 3705–3843. [195]

CARRASCO, M., AND J.-P. FLORENS (2005): "Spectral Method for Deconvolving a Density," Working Paper, University of Rochester. [195]

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. 6. Amsterdam: Elsevier Science, forthcoming. [196,199]

CHALAK, K., AND H. WHITE (2006): "An Extended Class of Instrumental Variables for the Estimation of Causal Effects," Working Paper, UCSD. [198,201]

CHEN, X., H. HONG, AND E. TAMER (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343–366. [196]

CHEN, X., H. HONG, AND A. TAROZZI (2008): "Semiparametric Efficiency in GMM Models with Nonclassical Measurement Error," *The Annals of Statistics*, forthcoming. [196]

CHEN, X., AND X. SHEN (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314. [205]

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. [200]

CHERNOZHUKOV, V., G. W. IMBENS, AND W. K. NEWEY (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4–14. [200,204]

CHESHER, A. (1991): "The Effect of Measurement Error," *Biometrika*, 78, 451. [195]

——— (1998): "Polynomial Regression with Normal Covariate Measurement Error," Discussion Paper 98/448, University of Bristol. [195]

——— (2001): "Parameter Approximations for Quantile Regressions with Measurement Error," Working Paper CWP02/01, Department of Economics, University College London. [195]

——— (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. [204]

CHESHER, A., M. DUMANGANE, AND R. J. SMITH (2002): "Duration Response Measurement Error," *Journal of Econometrics*, 111, 169–194. [195]

DAROLLES, S., J.-P. FLORENS, AND E. RENAULT (2002): "Nonparametric Instrumental Regression," Working Paper 05-2002, Centre de Recherche et Développement en Économique. [196,200]

D'HAULTFOEUILLE, X. (2006): "On the Completeness Condition in Nonparametric Instrumental Problems," Working Paper, ENSAE, CREST-INSEE, and Université de Paris I. [200]

DUNFORD, N., AND J. T. SCHWARTZ (1971): *Linear Operators*. New York: Wiley. [203,211,214]

GRENANDER, U. (1981): *Abstract Inference*. New York: Wiley. [204]

HALL, P., AND J. L. HOROWITZ (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *The Annals of Statistics*, 33, 2904–2929. [200]

HAUSMAN, J. (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15, 57–67. [195]

HAUSMAN, J., W. NEWEY, H. ICHIMURA, AND J. POWELL (1991): "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 273–295. [195]

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. [198]

HODERLEIN, S., AND E. MAMMEN (2007): "Identification of Marginal Effects in Nonseparable Models Without Monotonicity," *Econometrica*, 75, 1513–1518. [198]

HONG, H., AND E. TAMER (2003): "A Simple Estimator for Nonlinear Error in Variable Models," *Journal of Econometrics*, 117, 1–19. [195]

HOROWITZ, J. L. (2006): "Testing a Parametric Model Against a Nonparametric Alternative with Identification Through Instrumental Variables," *Econometrica*, 74, 521–538. [200]

HU, Y. (2007): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," Working Paper, Johns Hopkins University. [196]

HU, Y., AND G. RIDDER (2004): "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," Working Paper, Department of Economics, University of Southern California. [195]

HU, Y., AND S. M. SCHENNACH (2008): "Supplement to 'Instrumental Variable Treatment of Nonclassical Measurement Error Models'," *Econometrica Supplementary Material*, 76, http://www.econometricsociety.org/ecta/Supmat/6545_proofs.pdf. [205,207,212]

LEWBEL, A. (1996): "Demand Estimation with Expenditure Measurement Errors on the Left and Right Hand Side," *The Review of Economics and Statistics*, 78, 718–725. [195]

——— (1998): "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121. [195]

——— (2007): "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 75, 537–551. [196]

LI, T. (2002): "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 110, 1–26. [195]

MAHAJAN, A. (2006): "Identification and Estimation of Single Index Models with Misclassified Regressor," *Econometrica*, 74, 631–665. [196]

MATTNER, L. (1993): "Some Incomplete but Boundedly Complete Location Families," *The Annals of Statistics*, 21, 2158–2162. [200]

MATZKIN, R. L. (2003): "Nonparametric Estimation of Nonparametric Nonadditive Random Functions," *Econometrica*, 71, 1339–1375. [204]

NEWEY, W. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627. [195]

NEWEY, W. K., AND J. L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578. [196,200]

SCHENNACH, S. M. (2004a): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33–75. [195]

——— (2004b): "Nonparametric Estimation in the Presence of Measurement Error," *Econometric Theory*, 20, 1046–1093. [195]

——— (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75, 201–239. [195]

SHEN, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591. [204,205]

SUPPLEMENT TO "INSTRUMENTAL VARIABLE TREATMENT OF NONCLASSICAL MEASUREMENT ERROR MODELS"

(*Econometrica*, Vol. 76, No. 1, January 2008, 195–216)

BY YINGYAO HU AND SUSANNE M. SCHENNACH[1]

This supplementary material contains some of the more technical details omitted from the main paper. First, the asymptotic theory of the proposed sieve maximum likelihood estimator is fully developed, providing suitable regularity conditions, a nonparametric consistency result, and a semiparametric asymptotic normality and root n consistency result. Second, we provide an example that shows the necessity, for identification purposes, of our location constraint assumption regarding the measurement error. Third, a detailed example that illustrates the implementation of this location constraint with linear sieves is given. Finally, additional simulation results are reported.

S1. ASYMPTOTICS

LET US FIRST RECALL the assumptions needed for identification.

ASSUMPTION 1: *The joint density of y and x, x^*, z admits a bounded density with respect to the product measure of some dominating measure μ (defined on $\mathcal{Y}$) and the Lebesgue measure on $\mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$. All marginal and conditional densities are also bounded.*

ASSUMPTION 2: (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* $(x, x^*, z) \in \mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$.

ASSUMPTION 3: *The operators $L_{x|x^*}$ and $L_{z|x}$ are injective (for either $\mathcal{G} = \mathcal{L}^1$ or $\mathcal{G} = \mathcal{L}^1_{\text{bnd}}$).*

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

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

Our sieve estimator is based on the following expression for the observed density (following Theorem 1 in the main text):

$$f_{yx|z}(y, x|z; \alpha_0) = \int_{\mathcal{X}^*} f_{y|x^*}(y|x^*; \theta_0) f_{x|x^*}(x|x^*) f_{x^*|z}(x^*|z)\, dx^*. \tag{S1}$$

[1]Susanne M. Schennach acknowledges support from the National Science Foundation via grant SES-0452089. The authors would like to thank Lars Hansen, James Heckman, Marine Carrasco, Maxwell Stinchcombe, and Xiaohong Chen, as well as seminar audiences at various universities, at the Cemmap/ESRC Econometric Study Group Workshop on Semiparametric Methods, and at the Econometric Society 2006 Winter Meetings for helpful comments.

The unknown α_0 in the density function $f_{yx|z}$ includes θ_0 and density functions $f_{x|x^*}$ and $f_{x^*|z}$, that is, $\alpha_0 = (\theta_0, f_{x|x^*}, f_{x^*|z})^T$. The estimation procedure basically consists of replacing $f_{x|x^*}$ and $f_{x^*|z}$ (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

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

denote the $(a_1 + \cdots + a_d)$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 $\underline{\gamma}$ derivative is bounded, and the $\underline{\gamma}$th derivative are Hölder continuous with the exponent $\gamma - \underline{\gamma} \in (0, 1]$, that is,

$$\max_{a_1+\cdots+a_d=\underline{\gamma}} |\nabla^a g(\xi) - \nabla^a g(\xi')| \le c(\|\xi - \xi'\|_E)^{\gamma-\underline{\gamma}}$$

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

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

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 form $\omega(\xi) = (1 + \|\xi\|_E^2)^{-s/2}$, $s > \gamma > 0$, 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 $\Lambda_c^{\gamma,\omega}(\mathcal{V}) \equiv \{g \in \Lambda^{\gamma,\omega}(\mathcal{V}) : \|g\|_{\Lambda^{\gamma,\omega}} \le c < \infty\}$.

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 $\gamma_1 > 1$.*[2]

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 $x^* \in \mathcal{X}^*$.*

[2]If η is a density function, certain restrictions should be added to Assumption 6 analogous to those in Assumptions 8 and 7.

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 $z \in \mathcal{Z}$.*

$$\mathcal{M} = \{\eta(\cdot,\cdot) : \textit{Assumption 6 holds}\},$$
$$\mathcal{F}_1 = \{f_1(\cdot|\cdot) : \textit{Assumptions 3, 5, and 7 hold}\},$$
$$\mathcal{F}_2 = \{f_2(\cdot|\cdot) : \textit{Assumptions 3 and 8 hold}\}.$$

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

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

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, and so forth. To approximate functions of two variables, we use a tensor–product linear sieve basis, denoted by $p^{k_n}(\cdot,\cdot) = (p_1^{k_n}(\cdot,\cdot), p_2^{k_n}(\cdot,\cdot), \ldots, p_{k_n}^{k_n}(\cdot,\cdot))^T$. 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[3]

$$\begin{aligned}
\mathcal{M}_n = \{&\eta(\xi_1,\xi_2) = p^{k_n}(\xi_1,\xi_2)^T\delta \text{ for all } \delta\\
&\text{s.t. Assumption 6 holds}\},\\
\mathcal{F}_{1n} = \{&f(x|x^*) = p^{k_n}(x,x^*)^T\beta \text{ for all } \beta\\
&\text{s.t. Assumptions 3, 5, and 7 hold}\},\\
\mathcal{F}_{2n} = \{&f(x^*|z) = p^{k_n}(x^*,z)^T\gamma \text{ for all } \gamma\\
&\text{s.t. Assumptions 3 and 8 hold}\}.
\end{aligned}$$

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

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

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 through either one of a number of numerical

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

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 purposes of simplicity and conciseness, this section provides specific sufficient regularity conditions for the independent and identically distributed (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

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

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

where $p_i(\cdot)$ 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,

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

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 $\sum_{j=0}^{j_n} (\sum_{i=0}^{i_n} \beta_{ij} \times \int_{\mathcal{E}} p_i(\varepsilon)\, d\varepsilon) p_j(x^*) = 1$ for all x^*, where $\varepsilon = x - x^*$. Suppose $p_0(\cdot)$ is the only constant in $p_j(\cdot)$. 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 $j = 1, 2, \ldots, j_n$. Similar restrictions can be found

for $\int_{\mathcal{X}^*} f_2(x^*|z)\,dx^* = 1$. Moreover, the identification assumption, 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 $\frac{\partial}{\partial x} f_{x|x^*}(x|x^*)|_{x=x^*} = 0$ in the definition of $\mathcal{F}_{1n}$ implies $\sum_{j=0}^{j_n}(\sum_{i=0}^{i_n} \beta_{ij}(\partial p_i(0))/\partial x) q_j(x^*) = 0$. 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 $j = 1, 2, \ldots, j_n$. Similar restrictions can be found for the zero mean and the zero median cases. In all three cases, Assumption 5 can be expressed as linear restrictions on β, which are easy to implement. See Section S4 for an explicit expression for the restrictions in the case where Fourier series are used in the sieve approximation.

S1.1. *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 $\omega(\xi) = (1 + \|\xi\|_E^2)^{-\varsigma/2}$, $\varsigma > \gamma_1 > 0$. 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) $E[(\ln f_{yx|z}(D))^2]$ *is bounded*. (ii) *There exists a measurable function* $h_1(D)$ *with* $E\{(h_1(D))^2\} < \infty$ *such that, for any* $\overline{\alpha} = (\overline{\theta}, \overline{f}_1, \overline{f}_2)^T \in \mathcal{A}$,

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

where $f_{yx|z}^{|1|}(D, \overline{\alpha}, \bar{\omega})$ *is defined as* $\frac{d}{dt} f_{yx|z}(D; \overline{\alpha} + t\bar{\omega})|_{t=0}$ *with each linear term, that is,* $\frac{d}{d\theta} f_{y|x^*}$, $\overline{f}_1$, *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* $f_{yx|z}^{|1|}(D, \overline{\alpha}, \bar{\omega})$ *can be found in Equation* (S6) *in the proof of Lemma 2.*)

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, to control the bias, while ensuring that the number of terms in the sieve grows slower than the sample size, to control the variance. We show consistency in the following lemma.

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

See Section S2 for the proof.

Consistency under the norm $\|\cdot\|_s$ is the first step needed to obtain the asymptotic properties of the estimator. To proceed toward our main semiparametric asymptotic normality and root n consistency result, we then need to establish convergence at the rate $o_p(n^{-1/4})$ in a suitable norm. 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 $o_p(n^{-1/4})$ 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

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

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, that is, 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,

$$\begin{aligned}&\frac{d \ln f_{yx|z}(D, \alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \\ &\quad\equiv \frac{d \ln f_{yx|z}(D, \alpha_0)}{d\alpha}[\alpha_1 - \alpha_0] - \frac{d \ln f_{yx|z}(D, \alpha_0)}{d\alpha}[\alpha_2 - \alpha_0].\end{aligned}$$

In our setting, the pathwise derivative at α_0 is (from Equation (S1))

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

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

$$\|\alpha_1-\alpha_2\|\equiv\sqrt{E\left\{\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[\alpha_1-\alpha_2]\right)^2\right\}} \tag{S2}$$

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

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

ASSUMPTION 14: (i) *There exists a measurable function $c(D)$ with $E\{c(D)^4\}<\infty$ such that $|\ln f_{yx|z}(D;\alpha)|\le c(D)$ 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 $c_1,c_2>0$,*

$$c_1E\left(\ln\frac{f_{yx|z}(D;\alpha_0)}{f_{yx|z}(D;\alpha)}\right)\le\|\alpha-\alpha_0\|^2\le c_2E\left(\ln\frac{f_{yx|z}(D;\alpha_0)}{f_{yx|z}(D;\alpha)}\right)$$

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

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

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

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, and so forth (see Newey (1995, 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 $K(\alpha, \alpha_0) = E(\ln(f_{yx|z}(D;\alpha_0))/(f_{yx|z}(D;\alpha)))$, 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 2: *Under Assumptions* 1–5 *and* 9–18, *we have* $\|\widehat{\alpha}_n - \alpha_0\| = o_p(n^{-1/4})$.

The proof is given in Section S2.

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 $o_p(n^{-1/4})$ convergence under rather weak conditions.

S1.2. *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

$$\langle v_1, v_2\rangle = E\left\{\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[v_1]\right)\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[v_2]\right)\right\}. \tag{S3}$$

Obviously, the weak norm $\|\cdot\|$ defined in Equation (S2) 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 $\overline{\mathcal{W}} \equiv \overline{\mathcal{M}\times\mathcal{F}_1\times\mathcal{F}_2} - \{(\eta_0, f_{x|x^*}, f_{x^*|z})^T\}$) and define the Hilbert space $(\overline{\mathbf{V}}, \langle\cdot,\cdot\rangle)$ with its inner product defined in Equation (S3).

As shown above, we have

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

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

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

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

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

and the row vector

$$G_{w^*}(D) = \frac{d \ln f_{yx|z}(D, \alpha_0)}{db^T} - \frac{d \ln f_{yx|z}(D, \alpha_0)}{df}[w^*]. \tag{S4}$$

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 $E[G_{w^*}(D)^T G_{w^*}(D)]$ 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

$$s(\alpha) - s(\alpha_0) \equiv \lambda^T (b - b_0) = \langle v^*, \alpha - \alpha_0 \rangle \tag{S5}$$

for all $\alpha \in \mathcal{A}$. Here $v^* \equiv \binom{v_b^*}{v_f^*}$, $v_b^* = J^{-1}\lambda$, and $v_f^* = -w^* v_b^*$ with $J = E[G_{w^*}(D)^T \times G_{w^*}(D)]$. 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 $(\eta_0, f_{x|x^*}, f_{x^*|z})$ for the problem of estimating b_0. In the language of Stein (1956), v_f^* yields the most difficult one-dimensional subproblem. Equation (S5) implies that it is sufficient to find the asymptotic distribution of $\langle v^*, \widehat{\alpha}_n - \alpha_0 \rangle$ to obtain that of $\lambda^T(\widehat{b}_n - b_0)$ under suitable conditions. We denote

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

For a sieve MLE, we have that

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

Note that $((d \ln f_{yx|z}(D, \alpha))/d\alpha[v^*]) = G_{w^*}(D) J^{-1} \lambda$. Thus, by the classical central limit theorem, the asymptotic distribution of $\sqrt{n}(\widehat{b}_n - b_0)$ is $N(0, J^{-1})$. 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

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

with $\upsilon_n = o(1)$ and define $\mathcal{N}_0$ 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 $\varepsilon_n = o(n^{-1/2})$. Let $\Pi_n \alpha^*(\alpha, \varepsilon_n)$ be the projection of $\alpha^*(\alpha, \varepsilon_n)$ onto $\mathcal{A}_n$.

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

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

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

where $f^{|2|}_{yx|z}(D, \overline{\alpha}, \bar{\omega})$ *is defined as* $(d^2/dt^2) f_{yx|z}(D; \overline{\alpha} + t\bar{\omega})|_{t=0}$ *with each linear term, that is,* $\frac{d}{d\theta} f_{y|x^*}$, $\frac{d^2}{d\theta^2} f_{y|x^*}$, $\overline{f}_1$, *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* (S17) *in the proof of Theorem* 3.)

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

$$\begin{aligned}
&\frac{d \ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}[p^{k_n}] \\
&\quad = \left(\frac{d \ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}[p_1^{k_n}], \frac{d \ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}[p_2^{k_n}],\right. \\
&\qquad \left. \ldots, \frac{d \ln f_{yx|z}(D, \alpha_0)}{d\widetilde{f}}[p_{k_n}^{k_n}]\right)^T, \\
&\frac{d \ln f_{yx|z}(D, \alpha_0)}{db} \\
&\quad = \left(\frac{d \ln f_{yx|z}(D, \alpha_0)}{db_1}, \frac{d \ln f_{yx|z}(D, \alpha_0)}{db_2}, \ldots, \frac{d \ln f_{yx|z}(D, \alpha_0)}{db_{d_b}}\right)^T, \\
&\frac{d \ln f_{yx|z}(D, \alpha_0)}{d\alpha}[p^{k_n}] \\
&\quad = \left(\left(\frac{d \ln f_{yx|z}(D, \alpha_0)}{db}\right)^T, \left(\frac{d \ln f_{yx|z}(D, \alpha_0)}{d\eta}[p^{k_n}]\right)^T,\right. \\
&\qquad \left.\left(\frac{d \ln f_{yx|z}(D, \alpha_0)}{df_1}[p^{k_n}]\right)^T, \left(\frac{d \ln f_{yx|z}(D, \alpha_0)}{df_2}[p^{k_n}]\right)^T\right)^T,
\end{aligned}$$

and

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

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

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

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

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

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, 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, that is, Condition B in Shen (1997), and can be simplified to: for all $\alpha \in \mathcal{N}_{0n}$,

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

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 3: *Under Assumptions* 1–5, 9–16, *and* 19–23, $\sqrt{n}(\widehat{b}_n - b_0) \xrightarrow{d} N(0, J^{-1})$, *where* $J = E[G_{w^*}(D)^T G_{w^*}(D)]$ *for* $G_{w^*}(D)$ *given in Equation* (S4).

See Section S2 for the proof.

Achieving the level of generality provided by Theorem 3 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 (2007), 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-variables 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 (2004)). 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, that is, it is 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, that is, when nonlinear remainder terms in the estimator's stochastic expansion are not negligible, then any one of Assumption 20, 21, or 23 will not hold. Intuitively, this represents a case where the local quadratic behavior of the objective function is lost.

S2. 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 Assumption 1 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

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

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

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

$$+\int f_{y|x^*}(y|x^*;\overline{\theta})[f_{11}-f_{12}]\overline{f}_2(x^*|z)\,dx^*$$
$$+\int f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)[f_{21}-f_{22}]\,dx^*\Bigg\}.$$

The bounds can be found as

$$\text{(S6)}\qquad \left|\frac{d}{dt}\ln f_{yx|z}(D;\overline{\alpha}_0+t(\alpha_1-\alpha_2))\right|_{t=0}$$
$$\le \frac{1}{|f_{yx|z}(D,\overline{\alpha}_0)|}\left\{\int\left|\frac{d}{d\theta}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(\xi)\overline{f}_1(x|x^*)\overline{f}_2(x^*|z)\right|dx^*\right.$$
$$\times\|\theta_1-\theta_2\|_s$$
$$+\int|f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(x,x^*)\overline{f}_2(x^*|z)|\,dx^*\,\|f_{11}-f_{12}\|_s$$
$$\left.+\int|f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)\omega^{-1}(x^*,z)|\,dx^*\,\|f_{21}-f_{22}\|_s\right\}$$
$$\le \frac{1}{|f_{yx|z}(D,\overline{\alpha}_0)|}\left\{\int\left|\frac{d}{d\theta}f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(\xi)\overline{f}_1(x|x^*)\overline{f}_2(x^*|z)\right|dx^*\right.$$
$$+\int|f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(x,x^*)\overline{f}_2(x^*|z)|\,dx^*$$
$$\left.+\int|f_{y|x^*}(y|x^*;\overline{\theta})\overline{f}_1(x|x^*)\omega^{-1}(x^*,z)|\,dx^*\right\}\|\alpha-\alpha_0\|_s$$
$$\equiv\left|\frac{f^{|1|}_{yx|z}(D,\overline{\alpha}_0,\bar{\omega})}{f_{yx|z}(D,\overline{\alpha}_0)}\right|\|\alpha-\alpha_0\|_s,$$

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

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

with $\xi\in\mathcal{U}$. Therefore, our Assumption 11(ii), that is, $E((f^{|1|}_{yx|z}(D,\overline{\alpha}_0,\bar{\omega}))/(f_{yx|z}(D,\overline{\alpha}_0)))^2\le E(h_1(D))^2<\infty$, 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 $\alpha\in\mathcal{A}$

$$\text{(S7)}\qquad \|\alpha-\Pi_n\alpha\|_s\le\|\eta-\Pi_n\eta\|_s+\|f_1-\Pi_n f_1\|_s+\|f_2-\Pi_n f_2\|_s$$
$$=O(k_n^{-\gamma_1/d_1})$$

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). *Q.E.D.*

PROOF OF THEOREM 2: First note that Assumptions 2–5 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, that is, $L_n(\alpha)$ and $\widehat{L}_n(\alpha)$ in their Appendix B (Ai and Chen (2003, p. 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), that is, $\|\alpha - \Pi_n\alpha\| = o(n^{-1/4})$. Assumptions 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. *Q.E.D.*

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 1 in Shen (1997). We define the remainder term as

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

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

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

as $\|\alpha - \alpha_0\| \to 0$, 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

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

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

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

The Taylor expansion gives

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

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

$$\begin{aligned}
&\mu_n\big(r[\alpha-\alpha_0,D]-r[\Pi_n\alpha^*(\alpha,\varepsilon_n)-\alpha_0,D]\big)\\
&\quad=\mu_n\Bigg(\frac{d\ln f_{yx|z}(D,\Pi_n\alpha^*(\alpha,\varepsilon_n))}{d\alpha}[\alpha-\Pi_n\alpha^*(\alpha,\varepsilon_n)]\\
&\qquad-\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[\alpha-\Pi_n\alpha^*(\alpha,\varepsilon_n)]\Bigg)\\
&\qquad+\mu_n\Bigg(\frac{1}{2}\frac{d^2\ln f_{yx|z}(D,\widetilde{\alpha}_1)}{d\alpha\,d\alpha^T}[\alpha-\Pi_n\alpha^*(\alpha,\varepsilon_n),\alpha-\Pi_n\alpha^*(\alpha,\varepsilon_n)]\Bigg).
\end{aligned}\tag{S8}$$

Since

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

the right-hand side of Equation (S8) equals

$$
\begin{aligned}
(S9)\qquad &= \mu_n\left(\frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[\alpha - \Pi_n\alpha^*(\alpha,\varepsilon_n), \Pi_n\alpha^*(\alpha,\varepsilon_n) - \alpha_0]\right) \\
&\quad + \mu_n\left(\frac{1}{2}\frac{d^2 \ln f_{yx|z}(D,\widetilde{\alpha}_1)}{d\alpha\, d\alpha^T}[\varepsilon_n\Pi_n(\alpha - \alpha_0 - v^*), \varepsilon_n\Pi_n(\alpha - \alpha_0 - v^*)]\right) \\
&= \mu_n\left(\frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[\varepsilon_n\Pi_n(\alpha - \alpha_0 - v^*), \Pi_n\alpha^*(\alpha,\varepsilon_n) - \alpha_0]\right) \\
&\quad + \mu_n\left(\frac{1}{2}\frac{d^2 \ln f_{yx|z}(D,\widetilde{\alpha}_1)}{d\alpha\, d\alpha^T}[\varepsilon_n\Pi_n(\alpha - \alpha_0 - v^*), \varepsilon_n\Pi_n(\alpha - \alpha_0 - v^*)]\right) \\
&= \varepsilon_n\mu_n\left(\frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}\right. \\
&\quad \left.\times [\Pi_n(\alpha - \alpha_0 - v^*), \varepsilon_n\Pi_n(v^* + \alpha_0 - \alpha) + (\alpha - \alpha_0)]\right) \\
&\quad + \varepsilon_n^2\mu_n\left(\frac{1}{2}\frac{d^2 \ln f_{yx|z}(D,\widetilde{\alpha}_1)}{d\alpha\, d\alpha^T}[\Pi_n(\alpha - \alpha_0 - v^*), \Pi_n(\alpha - \alpha_0 - v^*)]\right) \\
&= \varepsilon_n\mu_n\left(\frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[\Pi_n(\alpha - \alpha_0 - v^*), \alpha - \alpha_0]\right) \\
&\quad - \varepsilon_n^2\mu_n\left(\frac{1}{2}\frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[\Pi_n(\alpha - \alpha_0 - v^*), \Pi_n(\alpha - \alpha_0 - v^*)]\right) \\
&\quad + \varepsilon_n^2\mu_n\left(\frac{1}{2}\frac{d^2 \ln f_{yx|z}(D,\widetilde{\alpha}_1)}{d\alpha\, d\alpha^T}[\Pi_n(\alpha - \alpha_0 - v^*), \Pi_n(\alpha - \alpha_0 - v^*)]\right) \\
&= A_1 + A_2 + A_3,
\end{aligned}
$$

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

$$
(S10)\qquad \sup_{\alpha\in\mathcal{N}_{0n}} A_1 = \varepsilon_n \sup_{\alpha\in\mathcal{N}_{0n}} \mu_n\left(\frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[\Pi_n(\alpha - \alpha_0 - v^*), \alpha - \alpha_0]\right).
$$

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

$$
\begin{aligned}
(S11)\qquad &\left|\sup_{\alpha\in\mathcal{N}_{0n}} \frac{d^2 \ln f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[v_n, \alpha - \alpha_0]\right| \\
&\quad \le \sup_{\alpha\in\mathcal{N}_{0n}} \left|\frac{1}{f_{yx|z}(D,\overline{\alpha}_1)}\frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[v_n, (\alpha - \alpha_0)]\right.
\end{aligned}
$$

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

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

$$
\begin{aligned}
&\frac{d \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha}[\alpha - \alpha_0] \\
&= \frac{1}{f_{yx|z}(D, \overline{\alpha}_1)} \left\{ \int \frac{d}{d\theta} f_{y|x^*}(y|x^*; \overline{\theta})(\theta - \theta_0) \overline{f}_1(x|x^*) \overline{f}_2(x^*|z) \, dx^* \right. \\
&\quad + \int f_{y|x^*}(y|x^*; \overline{\theta})[f_1 - f_{x|x^*}] \overline{f}_2(x^*|z) \, dx^* \\
&\quad \left. + \int f_{y|x^*}(y|x^*; \overline{\theta}) \overline{f}_1(x|x^*)[f_2 - f_{x^*|z}] \, dx^* \right\}.
\end{aligned} \tag{S12}
$$

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

$$
\begin{aligned}
&\left| \frac{d \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha}[\alpha - \alpha_0] \right| \\
&\leq \frac{1}{|f_{yx|z}(D, \overline{\alpha}_1)|} \left\{ \int \left| \frac{d}{d\theta} f_{y|x^*}(y|x^*; \overline{\theta}) \omega^{-1}(\xi) \overline{f}_1(x|x^*) \overline{f}_2(x^*|z) \right| dx^* \right. \\
&\quad \times \|\theta - \theta_0\|_s \\
&\quad + \int |f_{y|x^*}(y|x^*; \overline{\theta}) \omega^{-1}(x, x^*) \overline{f}_2(x^*|z)| \, dx^* \, \|f_1 - f_{x|x^*}\|_s \\
&\quad \left. + \int |f_{y|x^*}(y|x^*; \overline{\theta}) \overline{f}_1(x|x^*) \omega^{-1}(x^*, z)| \, dx^* \, \|f_2 - f_{x^*|z}\|_s \right\} \\
&\leq \left| \frac{f^{|1|}_{yx|z}(D, \overline{\alpha}_1, \bar{\omega})}{f_{yx|z}(D, \overline{\alpha}_1)} \right| \|\alpha - \alpha_0\|_s,
\end{aligned} \tag{S13}
$$

where $f^{|1|}_{yx|z}(D, \overline{\alpha}_1, \bar{\omega})$ is defined in Assumption 11 and Equation (S6). Similarly, we also have

$$
\left| \frac{d \ln f_{yx|z}(D, \overline{\alpha}_1)}{d\alpha}[v_n] \right| \leq \left| \frac{f^{|1|}_{yx|z}(D, \overline{\alpha}_1, \bar{\omega})}{f_{yx|z}(D, \overline{\alpha}_1)} \right| \|v_n\|_s \tag{S14}
$$

with

$$\|v_n\|_s = \|\Pi_n(\alpha - \alpha_0 - v^*)\|_s \le \|v_n^*\|_s + \|\Pi_n(\alpha - \alpha_0)\|_s < \infty. \tag{S15}$$

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

$$\begin{aligned}
&\frac{1}{f_{yx|z}(D,\overline{\alpha}_1)}\frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[v_n,(\alpha-\alpha_0)] \\
&\quad= \frac{1}{f_{yx|z}(D,\overline{\alpha}_1)}\left\{\int \frac{d^2}{d\theta^2} f_{y|x^*}(y|x^*;\overline{\theta})[v_n]_\theta(\theta-\theta_0)\right. \\
&\qquad\times \overline{f}_1(x|x^*)\overline{f}_2(x^*|z)\,dx^* \\
&\qquad+\int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})[v_n]_\theta[f_1 - f_{x|x^*}]\overline{f}_2(x^*|z)\,dx^* \\
&\qquad+\int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})[v_n]_\theta\overline{f}_1(x|x^*)[f_2 - f_{x^*|z}]\,dx^* \\
&\qquad+\int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})(\theta-\theta_0)[v_n]_{f_1}\overline{f}_2(x^*|z)\,dx^* \\
&\qquad+\int f_{y|x^*}(y|x^*;\overline{\theta})[v_n]_{f_1}[f_2 - f_{x^*|z}]\,dx^* \\
&\qquad+\int \frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})(\theta-\theta_0)\overline{f}_1(x|x^*)[v_n]_{f_2}\,dx^* \\
&\qquad\left.+\int f_{y|x^*}(y|x^*;\overline{\theta})[f_1 - f_{x|x^*}][v_n]_{f_2}\,dx^*\right\}.
\end{aligned} \tag{S16}$$

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

$$\begin{aligned}
&\left|\frac{1}{f_{yx|z}(D,\overline{\alpha}_1)}\frac{d^2 f_{yx|z}(D,\overline{\alpha}_1)}{d\alpha\, d\alpha^T}[v_n,(\alpha-\alpha_0)]\right| \\
&\quad\le \frac{1}{|f_{yx|z}(D,\overline{\alpha}_1)|}\left\{\int\left|\frac{d^2}{d\theta^2} f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(\xi)\omega^{-1}(\xi)\right.\right. \\
&\qquad\left.\times \overline{f}_1(x|x^*)\overline{f}_2(x^*|z)\right| dx^*\,\|[v_n]_\theta\|_s\|\theta-\theta_0\|_s \\
&\qquad+\int\left|\frac{d}{d\theta} f_{y|x^*}(y|x^*;\overline{\theta})\omega^{-1}(\xi)\omega^{-1}(x,x^*)\overline{f}_2(x^*|z)\right| dx^* \\
&\qquad\times\|[v_n]_\theta\|_s\|f_1 - f_{x|x^*}\|_s
\end{aligned} \tag{S17}$$

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

where $f^{|2|}_{yx|z}(D, \overline{\alpha}_1, \bar{\omega})$ is defined in Assumption 20. Plugging the bounds in Equations (S13), (S14), and (S17) back in to Equation (S11), we have

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

By the envelope condition in Assumption 20, Equation (S10) becomes

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

with $\|\alpha - \alpha_0\|_s = o(1)$. The last two terms, A_2 and A_3 in Equation (S9), are bounded as

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

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

Fourth, Condition B requires

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

As Corollary 2 in Shen (1997) points out, Condition B can be replaced by Condition B′ as

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

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

$$\begin{aligned}(\text{S18})\qquad &E[\ln f_{yx|z}(D,\alpha)-\ln f_{yx|z}(D,\alpha_0)]\\ &\quad=E\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[\alpha-\alpha_0]\right)\\ &\qquad+\frac{1}{2}E\left(\frac{d^2\ln f_{yx|z}(D,\alpha_0)}{d\alpha\,d\alpha^T}[\alpha-\alpha_0,\alpha-\alpha_0]\right)\\ &\qquad+\frac{1}{6}E\frac{d^3}{dt^3}\ln f_{yx|z}(D;\alpha_0+t(\alpha-\alpha_0))\Big|_{t=0}\\ &\qquad+\frac{1}{24}E\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t(\alpha-\alpha_0))\Big|_{t=0},\end{aligned}$$

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 $\int_{\mathcal{Y}}(\partial/\partial\eta)f_{y|x^*}(y|x^*;\theta)\,dy=0$, $\int_{\mathcal{Y}}(\partial^2/\partial\eta^2)f_{y|x^*}(y|x^*;\theta)\,dy=0$, and $\int_{\mathcal{Y}}(\partial^3/\partial\eta^3)f_{y|x^*}(y|x^*;\theta)\,dy=0$ for all $\theta\in\Theta$, and we have f_1 and 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 [5]

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

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

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

Therefore,

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

Therefore, Equation (S18) becomes

$$\begin{aligned}
\text{(S19)}\qquad &E[\ln f_{yx|z}(D,\alpha)-\ln f_{yx|z}(D,\alpha_0)]\\
&\quad=-\frac{1}{2}\|\alpha-\alpha_0\|^2+\frac{1}{6}E\frac{d^3}{dt^3}\ln f_{yx|z}(D;\alpha_0+t(\alpha-\alpha_0))\Big|_{t=0}\\
&\qquad+\frac{1}{24}E\frac{d^4}{dt^4}\ln f_{yx|z}(D;\overline{\alpha}+t(\alpha-\alpha_0))\Big|_{t=0}.
\end{aligned}$$

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

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

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

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

For the term B_3, we have

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

Note that $E|h_2(D)|^2 < \infty$ implies $E|h_1(D)|^4 < \infty$. Therefore, Equation (S19) becomes

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

By Assumption 23, we have

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

and, therefore,

$$E[\ln f_{yx|z}(D,\alpha_0) - \ln f_{yx|z}(D,\alpha)] = \frac{1}{2}\|\alpha-\alpha_0\|^2(1+O(h_n)) \tag{S20}$$

with

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

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

$$\begin{aligned}
\alpha &= (b^T, \Pi_n\eta, \Pi_n f_1, \Pi_n f_2)^T \\
&= (b^T, p^{k_n}(\xi_1,\xi_2)^T\delta, p^{k_n}(x,x^*)^T\beta, p^{k_n}(x^*,z)^T\gamma)^T, \\
\Pi_n\alpha_0 &= (b_0^T, \Pi_n\eta_0, \Pi_n f_{x|x^*}, \Pi_n f_{x^*|z})^T \\
&= (b_0^T, p^{k_n}(\xi_1,\xi_2)^T\delta_0, p^{k_n}(x,x^*)^T\beta_0, p^{k_n}(x^*,z)^T\gamma_0)^T,
\end{aligned}$$

where p^{k_n}'s are k_n-by-1 vectors, that is, $p^{k_n}(\cdot,\cdot) = (p_1^{k_n}(\cdot,\cdot), p_2^{k_n}(\cdot,\cdot), \ldots, p_{k_n}^{k_n}(\cdot,\cdot))^T$. Note that all the vectors are column vectors. We also define the vector of the sieve coefficients as

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

We then have

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

Suppose that

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

with some small $\varsigma_0>0$. By Assumption 13 and Equation (S7), we let

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

for some small $\varsigma>\varsigma_0$.

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

$$\begin{aligned}&\big|\|\alpha-\alpha_0\|-\|\Pi_n\alpha_0-\alpha_0\|\big|\\ &\quad\le \|\alpha-\Pi_n\alpha_0\|\le\|\alpha-\alpha_0\|+\|\Pi_n\alpha_0-\alpha_0\|.\end{aligned}$$

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

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

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

$$C_1\|\alpha-\alpha_0\|\le\|\alpha-\Pi_n\alpha_0\|\le C_2\|\alpha-\alpha_0\|. \tag{S21}$$

Moreover, we define

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

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[p^{k_n}]$$
$$=\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[p_1^{k_n}],\right.$$
$$\left.\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[p_2^{k_n}],\ldots,\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[p_{k_n}^{k_n}]\right)^T,$$
$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}[p^{k_n}]$$
$$=\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}[p_1^{k_n}],\right.$$
$$\left.\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}[p_2^{k_n}],\ldots,\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}[p_{k_n}^{k_n}]\right)^T,$$
$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}[p^{k_n}]$$
$$=\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}[p_1^{k_n}],\right.$$
$$\left.\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}[p_2^{k_n}],\ldots,\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}[p_{k_n}^{k_n}]\right)^T,$$
$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[p^{k_n}]$$
$$=\left[\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{db}\right)^T,\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[p^{k_n}]\right)^T,\right.$$
$$\left.\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}[p^{k_n}]\right)^T,\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}[p^{k_n}]\right)^T\right]^T.$$

With the notations above, we have

$$\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[\alpha-\Pi_n\alpha_0]$$
$$=\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{db}\right)^T(b-b_0)+\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\eta}[p^{k_n}]\right)^T(\delta-\delta_0)$$
$$+\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_1}[p^{k_n}]\right)^T(\beta-\beta_0)$$

$$+\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{df_2}[p^{k_n}]\right)^T(\gamma-\gamma_0)$$
$$=\left(\frac{d\ln f_{yx|z}(D,\alpha_0)}{d\alpha}[p^{k_n}]\right)^T(\alpha^c-\alpha_0^c)$$

and

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

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 $0<C_1,C_2<\infty$,

$$C_1\|\alpha^c-\alpha_0^c\|_E\le\|\alpha-\Pi_n\alpha_0\|\le C_2\|\alpha^c-\alpha_0^c\|_E. \tag{S22}$$

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

We then consider the ratio $\|\alpha-\alpha_0\|_s^2/\|\alpha-\alpha_0\|$. From Equations (S21) and (S22), we have

$$\|\alpha-\alpha_0\|\ge C_1\|\alpha^c-\alpha_0^c\|_E \tag{S23}$$

and $\|\alpha^c-\alpha_0^c\|_E=O(n^{-1/4-\mathfrak{s}_0})$. Assumption 21 implies $\|\alpha-\Pi_n\alpha_0\|_s^2\le C_2\|\alpha^c-\alpha_0^c\|_1^2$, where $\|\cdot\|_1$ is the L_1 vector norm. Thus, we have

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

Since $\|\alpha^c-\alpha_0^c\|_E=O(n^{-1/4-\mathfrak{s}_0})$ and $\mathfrak{s}>\mathfrak{s}_0$, we have

$$\|\alpha-\alpha_0\|_s^2\le C_2k_n\|\alpha^c-\alpha_0^c\|_E^2. \tag{S24}$$

By Equations (S23) and (S24), we have

$$\frac{\|\alpha-\alpha_0\|_s^2}{\|\alpha-\alpha_0\|}\le\frac{C_2k_n\|\alpha^c-\alpha_0^c\|_E^2}{C_1\|\alpha^c-\alpha_0^c\|_E}\le O(k_n\|\alpha^c-\alpha_0^c\|_E).$$

Assumption 13 requires $k_n^{-\gamma_1/d_1} = O(n^{-1/4-\varsigma})$, that is, $k_n = n^{(1/4+\varsigma)1/(\gamma_1/d_1)}$. We then have

$$k_n\|\alpha^c - \alpha_0^c\|_E = O\big(n^{-1/4(1-1/(\gamma_1/d_1))+\varsigma(1/(\gamma_1/d_1))-\varsigma_0}\big) = o(1)$$

for $\varsigma < \frac{1}{4}(\gamma_1/d_1 - 1) + (\gamma_1/d_1)\varsigma_0$ with $\gamma_1/d_1 > 1$ in Assumption 13. Therefore, Equation (S20) 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

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

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

$$\begin{aligned}
&\|\alpha^*(\alpha, \varepsilon_n) - \Pi_n\alpha^*(\alpha, \varepsilon_n)\| \\
&\quad = \varepsilon_n\|v^* + \alpha_0 - \Pi_n(v^* + \alpha_0)\| \\
&\quad \le \varepsilon_n\|v^* - \Pi_n v^*\| + \varepsilon_n\|\alpha_0 - \Pi_n\alpha_0\| \\
&\quad = O(n^{-1/4}\varepsilon_n).
\end{aligned}$$

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

$$\sup_{\alpha\in\mathcal{N}_{0n}} \mu_n\left(\frac{d\ln f_{yx|z}(D, \alpha_0)}{d\alpha}[\alpha^*(\alpha, \varepsilon_n) - \Pi_n\alpha^*(\alpha, \varepsilon_n)]\right) = O_p(\varepsilon_n^2). \tag{S25}$$

The left-hand side equals

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

By the envelope condition in Assumption 11, the first term (corresponding to v^*) is

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

and the second term (corresponding to α_0) is

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

The last step is due to $\|\alpha_0-\Pi_n\alpha_0\|=o(n^{-1/4})$. Therefore, Condition C in Theorem 1 in Shen (1997) holds. Note that Condition C′ in Corollary 2 is also satisfied, that is, $\|v_n^*-v^*\|=o(n^{-1/4})$ and $o(h_n)\|\alpha_0-\Pi_n\alpha_0\|^2=o_p(n^{-1/2})$.

Finally, Condition D in Shen (1997), that is,

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

can be verified as follows: We first have

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

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

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

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. *Q.E.D.*

S3. 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 $\tilde{x}^*$ is the unobserved true regressor are observationally equivalent, because

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

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

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

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

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

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

$$[L_{x|\tilde{x}^*}^{-1}g](\tilde{x}^*) = r(\tilde{x}^*)[L_{x|x^*}^{-1}g](R(\tilde{x}^*)),$$

where $r(\tilde{x}^*) = dR(\tilde{x}^*)/d\tilde{x}^*$ whenever this differential exists and $r(\tilde{x}^*) = 0$ otherwise.[6] We can then calculate

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

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

S4. RESTRICTIONS WITH FOURIER SERIES

As shown above, the sieve estimators are

$$
f_1(x|x^*) = \sum_{i=0}^{i_n}\sum_{j=0}^{j_n}\beta_{ij}p_i(x-x^*)q_j(x^*),
$$

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

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

$$
p_k(x-x^*) = \cos\frac{k\pi}{l_e}(x-x^*) \quad \text{or} \quad \sin\frac{k\pi}{l_e}(x-x^*),
$$

$$
p_k(x^*-z) = \cos\frac{k\pi}{l_x}(x^*-z) \quad \text{or} \quad \sin\frac{k\pi}{l_x}(x^*-z),
$$

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

$$
f_1(x|x^*) = \left(a_{00} + a_{01}\cos\frac{\pi}{l_x}x^* + a_{02}\cos\frac{2\pi}{l_x}x^*\right)
$$

[6]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.

$$+\sum_{k=1}^{3}\left(a_{k0}+a_{k1}\cos\frac{\pi}{l_x}x^*+a_{k2}\cos\frac{2\pi}{l_x}x^*\right)\cos\frac{k\pi}{l_e}(x-x^*)$$

$$+\sum_{k=1}^{3}\left(b_{k0}+b_{k1}\cos\frac{\pi}{l_x}x^*+b_{k2}\cos\frac{2\pi}{l_x}x^*\right)\sin\frac{k\pi}{l_e}(x-x^*).$$

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

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

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

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

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

Thus, the restrictions on the coefficients are

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

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

$$\int_{\mathcal{X}}(x-x^*)f_1(x|x^*)\,dx$$

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

We have

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

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

$$\int_{\mathcal{X}\cap\{x<x^*\}} f_1(x|x^*)\,dx$$

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

Therefore,

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

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

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

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

S5. ADDITIONAL SIMULATIONS

EXAMPLE IV—Heteroskedastic Error with Zero Mean: Consider a measurement error

$$x = x^* + \sigma(x^*)\nu \tag{S26}$$

with $x^* \perp \nu$, $E(\nu) = 0$, and $\sigma(\cdot) > 0$ being an unknown nonstochastic function. These assumptions can also be written as $E(x - x^*|x^*) = 0$, that is, the measurement error is the conditional mean independent of the true value. The identification condition is also satisfied because it can be 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 SI.

EXAMPLE V—Nonadditive Error with Zero Mode: An error equation like (S26) is usually set up for convenience. The additive structure of (S26) with $x^* \perp \nu$ may not always be appropriate 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

$$f_{x|x^*}(x|x^*) = \frac{g(x, x^*)}{\int_{-\infty}^{\infty} g(x, x^*)\, dx},$$

TABLE SI

SIMULATION RESULTS

	$a=-1$			$b=1$		
	Mean	Std. Dev.	RMSE	Mean	Std. Dev.	RMSE
Example I						
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						
Example II						
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 III						
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						

$$g(x,x^*)=\exp\left\{h(x^*)\left[\left(\frac{x-x^*}{\sigma(x^*)}\right)-\exp\left(\frac{x-x^*}{\sigma(x^*)}\right)\right]\right\}.$$

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 (S26) 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 (S26) 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 SI.

EXAMPLE VI—Nonadditive Error with Zero Median: We let the cumulative distribution function that corresponds to $f_{x|x^*}$ be

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

with $h(x^*)>0$. Note that $F_{x|x^*}(x^*|x^*)=\frac{1}{2}$ 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 (S26). In the simulation, we use $\sigma(x^*)=0.5\exp(-x^*)$ and $h(x^*)=\exp(-0.1x^*)$. The simulation results are in Table SI.

Dept. of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, U.S.A.; yhu@jhu.edu

and

Dept. of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, U.S.A.; smschenn@uchicago.edu.

REFERENCES

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

BEGUN, J., W. HALL, W. HUANG, AND J. WELLNER (1983): "Information and Asymptotic Efficiency in Parametric–Nonparametric Models," *The Annals of Statistics*, 11, 432–452.

BLUNDELL, R., X. CHEN, AND D. KRISTENSEN (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.

CHEN, X., AND X. SHEN (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

CHEN, X., L. HANSEN, AND J. SCHEINKMAN (1997): "Shape-Preserving Estimation of Diffusions," Working Paper, University of Chicago.

CHEN, X., H. HONG, AND E. TAMER (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343–366.

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

GALLANT, A., AND D. NYCHKA (1987): "Semi-Nonparametric Maximum Likelihood Estimators," *Econometrica*, 55, 363–390.

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. Cambridge, MA: Blackwell, 254–275.

——— (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. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33–75.

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

STEIN, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium on Mathematics, Statistics and Probability*, 1. Berkeley, CA: University of California, 187–195.

TAUPIN, M.-L. (1998): "Estimation in the Nonlinear Errors-in-Variables Model," *Comptes Rendus de l'Académie des Science, Série I, Mathématiques*, 326, 885–890.