ESTIMATION OF NONLINEAR MODELS WITH MISMEASURED REGRESSORS USING MARGINAL INFORMATION

Author(s): YINGYAO HU and GEERT RIDDER

Source: *Journal of Applied Econometrics*, April-May 2012, Vol. 27, No. 3 (April-May 2012), pp. 347-385

Published by: Wiley

Stable URL: https://www.jstor.org/stable/41433012

JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **27**: 347–385 (2012)
Published online 21 July 2010 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.1202

ESTIMATION OF NONLINEAR MODELS WITH MISMEASURED REGRESSORS USING MARGINAL INFORMATION

YINGYAO HU[a]* AND GEERT RIDDER[b]

[a] *Department of Economics, Johns Hopkins University, Baltimore, MD, USA*
[b] *Department of Economics, University of Southern California, Los Angeles, CA, USA*

SUMMARY

We consider the estimation of nonlinear models with mismeasured explanatory variables, when information on the marginal distribution of the true values of these variables is available. We derive a semi-parametric MLE that is shown to be $\sqrt{n}$ consistent and asymptotically normally distributed. In a simulation experiment we find that the finite sample distribution of the estimator is close to the asymptotic approximation. The semi-parametric MLE is applied to a duration model for AFDC welfare spells with misreported welfare benefits. The marginal distribution of the correctly measured welfare benefits is obtained from an administrative source. Copyright © 2010 John Wiley & Sons, Ltd.

1. INTRODUCTION

Many models that are routinely used in empirical research in microeconomics are nonlinear in the explanatory variables. Examples are nonlinear (in variables) regression models, models for limited-dependent variables (logit, probit, tobit etc.), and duration models. Often the parameters of such nonlinear models are estimated using economic data in which one or more independent variables are measured with error (Bound *et al.*, 2001). The identification and estimation of models that are nonlinear in mismeasured variables is a notoriously difficult problem (see Carroll *et al.*, 1995, for a survey).

There are three approaches to this problem: (i) the parametric approach; (ii) the instrumental variable method; and (iii) methods that use an additional sample, such as a validation sample. Throughout we assume that we have a parametric model for the relation between the dependent and independent variables, but that we want to make minimal assumptions on the measurement errors and the distribution of the explanatory variables.

The parametric approach makes strong and untestable distributional assumptions. In particular, it is assumed that the distribution of the measurement error is in some parametric class (Bickel and Ritov, 1987; Hsiao, 1989, 1991; Cheng and Van Ness, 1994; Murphy and Van Der Vaart, 1996; Wang, 1998; Kong and Gu, 1999; Hsiao and Wang, 2000; Augustin, 2004). With this assumption the estimation problem is complicated, but fully parametric. The second approach is the instrumental variable method. In an errors-in-variables model, a valid instrument is a variable that (a) can be excluded from the model, (b) is correlated with the latent true value, and (c) is independent of the measurement error (Amemiya and Fuller, 1988; Carroll and Stefanski, 1990; Hausman *et al.*, 1991, 1995; Li and Vuong, 1998; Newey, 2001). Schennach (2004a, 2007) and Hu and Schennach (2008) extend the IV estimator to general nonlinear models. The third approach is to use an additional sample, such as a validation sample (Bound *et al.*, 1989; Hsiao, 1989;

* Correspondence to: Yingyao Hu, Department of Economics, Johns Hopkins University, 3400 N Charles Street, Baltimore, MD 21 218, USA. E-mail: yhu@jhu.edu

Hausman *et al.*, 1991; Pepe and Fleming, 1991; Carroll and Wand, 1991; Lee and Sepanski, 1995; Chen *et al.*, 2005). A validation sample is a subsample of the original sample for which accurate measurements are available. The approach taken in this paper is along these lines.

In this paper we show that many of the benefits of a validation sample can be obtained if we have a random sample from the marginal distribution of the mismeasured variables, i.e. we need not observe the mismeasured and true value and the other independent variables for the same units. Information on the marginal distribution of the true value is available in administrative registers, as employer's records, tax returns, quality control samples, medical records, unemployment insurance and social security records, and financial institution records. In fact, most validation samples are constructed by matching survey data to administrative data. Creating such matched samples is very costly and sometimes impossible. Moreover, the owners of the administrative data may be reluctant to release the data because the matching raises privacy issues. Our approach only requires a random sample from the administrative register. Indeed the random sample and the survey need not have any unit in common. Of course, if available a validation sample is preferable over marginal information. With a validation sample the assumptions on the measurement error can be substantially weaker than with marginal information. Because we do not observe the mismeasured and accurate variables for the same units, marginal information cannot identify the correlation between the measurement error and the true value. For that reason we maintain the assumption of classical measurement error, i.e. the measurement error is independent of the true value and also independent of the other covariates in the model. The latter assumption can be relaxed if these covariates are common to the survey sample and the administrative data. Validation studies have found that the assumption of classical measurement errors may not hold in practice (see, for example, Bound *et al.*, 1989). The main advantage of a validation sample over marginal information is that it allows us to avoid this assumption. However, as with a validation sample, marginal information allows us to avoid assumptions on the distribution of the measurement error and the latent true value. Given the scarcity of validation samples relative to administrative datasets, the correction developed in this paper can be more widely applied, but researchers must be aware that the estimates are biased if the assumptions on the measurement error do not hold.[1]

In recent years many studies have used administrative data, because they are considered to be more accurate. For example, employer's records have been used to study annual earnings and hourly wages (Angrist and Krueger, 1999; Bound *et al.*, 1994), union coverage (Barron *et al.*, 1997), and unemployment spells (Mathiowetz and Duncan, 1988). Tax returns have been used in studies of wage and income (Code, 1992), unemployment benefits (Dibbs *et al.*, 1995), and asset ownership and interest income (Grondin and Michaud, 1994). Cohen and Carlson (1994) study health care expenditures using medical records, and Johnson and Sanchez (1993) use these records to study health outcomes. Transcript data have been used to study years of schooling (Kane *et al.*, 1999). Card *et al.* (2001) examine Medicaid coverage using Medicaid data. Bound *et al.* (2001) give a survey of studies that use administrative data. A problem with administrative records is that they usually contain only a small number of variables. We show that the marginal distribution of the latent true values from administrative records is sufficient to correct for measurement error in a survey sample. There have been earlier attempts to combine survey and administrative data to deal with the measurement error in survey data. In the 1970s statistical matching of surveys and administrative files without common units was used to create synthetic datasets that contained the accurate data. Ridder and Moffitt (2007) survey this literature. This paper can be considered as a better approach to the use of accurate data from a secondary source to deal with measurement error.

[1] It is possible to apply the estimator developed in this paper with a level of dependence between the true value and the measurement error. In that case prior knowledge must be used to set the degree of dependence.

DOI: 10.1002/jae

Our application indicates what type of data can be used. We consider a duration model for the relation between welfare benefits and the length of welfare spells. The survey data are from the Survey of Income and Program Participation (SIPP). The welfare benefits in the SIPP are self-reported and are likely to contain reporting errors. The federal government requires the states to report random samples from their welfare records to check whether the welfare benefits are calculated correctly. The random samples are publicly available as the AFDC Quality Control Survey (AFDC QC). For that reason they do not contain identifiers that could be used to match the AFDC QC to the SIPP, a task that would yield a small sample anyway because of the lack of overlap of the two samples. Besides the welfare benefits the AFDC QC contains only a few other variables.

This paper shows that the combination of a sample survey in which some of the independent variables are measured with error and a secondary dataset that contains a sample from the marginal distribution of the latent true values of the mismeasured variables identifies the conditional distribution of the latent true value given the reported value and the other independent variables. This distribution is used to integrate out the latent true values from the model. The resulting mixture model (with estimated mixing distribution) can then be estimated by maximum likelihood (ML) (or generalized method of moments (GMM)). Our semi-parametric MLE involves two deconvolutions. The use of deconvolution estimators in the first stage is potentially problematic (Taupin, 2001). We apply the results in Hu and Ridder (2010), who show that $\sqrt{n}$ consistency can be obtained if the distribution of the measurement error is range-restricted.[2] We derive its asymptotic variance that accounts for the fact that the mixing distribution is estimated. The semi-parametric MLE avoids any assumption on the distribution of the measurement error and/or the distribution of the latent true value.

The paper is organized as follows. Section 2 establishes non-parametric identification. Section 3 gives the estimator and its properties. Section 4 presents Monte Carlo evidence on the finite-sample performance of the estimator. An empirical application is given in Section 5. Section 6 contains extensions and conclusions. The proofs are in the Appendix.

2. IDENTIFICATION USING MARGINAL INFORMATION

A parametric model for the relation between a dependent variable y, a latent true variable x^* and other covariates w can be expressed as a conditional density of y given x^*, w, $f^*(y|x^*, w; \theta)$. The relation between the observed x and the latent x^* is

$$x = x^* + \varepsilon \tag{1}$$

with the classical measurement error assumption $\varepsilon \perp x^*, w, y$ where $\perp$ indicates stochastic independence. In the linear regression model the independence of the measurement error and y given x^*, w, which is implied by this assumption, is equivalent to the independence of the measurement error and the regression error. The variable x^* (and hence x) is continuous. The independent variables in w can be either discrete or continuous. To keep the notation simple, the theory will be developed for the case that w is scalar.

The data are a random sample y_i, x_i, w_i, $i = 1, \ldots, n$ from the joint distribution of y, x, w, the survey data, and a random sample x_i^*, $i = 1, \ldots, n_1$ from the marginal distribution of x^*, the secondary sample that in most cases is a random sample from an administrative file. In asymptotic

[2] The 'range-restricted' condition does exclude some interesting distributions, such as the normal distribution. We clarify this limitation in the estimation section.

DOI: 10.1002/jae

arguments we assume that both n, n_1 become large and that their ratio converges to a positive and finite number.

Efficient inference for the parameters θ is based on the likelihood function. The individual contribution to the likelihood is the conditional density of y given x, w, $f(y|x, w; \theta)$. The relation between this density and that of the parametric model is

$$f(y|x, w; \theta) = \int_{\mathcal{X}^*} f^*(y|x^*, w; \theta) g(x^*|x, w) \, \mathrm{d}x^* \tag{2}$$

The conditional density $g(x^*|x, w)$ does not depend on θ, because x^*, w is assumed to be ancillary for θ, and the measurement error is independent of y given x^*, w.

The key problem with the use of the conditional density (2) in likelihood inference is that it requires knowledge of the density $g(x^*|x, w)$. This density can be expressed as

$$g(x^*|x, w) = \frac{g(x|x^*, w) g_2(x^*, w)}{g(x, w)} \tag{3}$$

For likelihood inference we must identify the densities $g(x|x^*, w)$ and $g_2(x^*, w)$, while the density in the denominator does not affect the inference. We could choose a parametric density for $g(x^*|x, w)$ and estimate its parameters jointly with θ. There are at least two problems with that approach. First, it is not clear whether the parameters in the density are identified, and if so, whether the identification is by the arbitrary distributional assumptions and/or the functional form of the parametric model. If there is (parametric) identification, misspecification of $g(x^*|x, w)$ will bias the MLE of θ. Second, empirical researchers are reluctant to make distributional assumptions on the independent variables in conditional models. For that reason we consider non-parametric identification and estimation of the density of x^* given x, w.

We have to show that the densities in the numerator are non-parametrically identified. First, the assumption that the measurement error ε is independent of x^*, w implies that

$$g(x|x^*, w) = g_1(x - x^*) \tag{4}$$

with g_1 the density of ε. Because the observed x is the convolution, i.e. sum, of the latent true value and the measurement error, it is convenient to work with the characteristic function of the random variables, instead of their density or distribution functions. Of course, there is a one-to-one correspondence between characteristic functions and distributions. Let $\phi_x(t) = \mathrm{E}(\exp(itx))$ be the characteristic function of the random variable x. From (1) and the assumption that x^* and ε are independent we have $\phi_x(t) = \phi_{x^*}(t)\phi_\varepsilon(t)$. Hence, if the marginal distribution of x^* is known, we can solve for the characteristic function of the measurement error distribution

$$\phi_\varepsilon(t) = \frac{\phi_x(t)}{\phi_{x^*}(t)} \tag{5}$$

Because of the one-to-one correspondence between characteristic functions and distributions, this identifies $g(x|x^*, w)$. By the law of total probability the density $g_2(x^*, w)$ is related to the density $g(x, w)$ as

$$g(x, w) = \int_{\mathcal{X}^*} g(x|x^*, w) g_2(x^*, w) \, \mathrm{d}x^* = \int_{\mathcal{X}^*} g_1(x - x^*) g_2(x^*, w) \, \mathrm{d}x^* \tag{6}$$

DOI: 10.1002/jae

This implies that the joint characteristic function $\phi_{xw}(r,s) = \mathrm{E}(\exp(irx + isw))$ of the distribution of x, w, is equal to (with a change of variables to $\varepsilon = x - x^*$)

$$\phi_{xw}(r,s) = \int_W \int_{X^*} \int_X e^{ir(x-x^*)} g_1(x-x^*)\,\mathrm{d}x e^{irx^*+isw} g_2(x^*,w)\,\mathrm{d}x^*\,\mathrm{d}w = \phi_\varepsilon(r)\phi_{x^*w}(r,s) \tag{7}$$

so that

$$\phi_{x^*,w}(r,s) = \frac{\phi_{x,w}(r,s)}{\phi_\varepsilon(r)} = \frac{\phi_{x,w}(r,s)\phi_{x^*}(r)}{\phi_x(r)} \tag{8}$$

If the data consist of a primary sample from the joint distribution of y, x, w and a secondary sample from the marginal distribution of x^*, then the right-hand side of (8) contains only characteristic functions of distributions that can be observed in either sample.

The conditional density in (2) is a mixture with a mixing distribution that can be identified from the joint distribution of x, w and the marginal distribution of x^*. We still must establish that θ can be identified from this mixture. The parametric model for the relation between y and x^*, w, specifies the conditional density of y given x^*, w, $f^*(y|x^*, w; \theta)$. The parameters in this model are identified, if for all $\theta \neq \theta_0$ with θ_0 the population value of the parameter vector, there is a set $A(\theta)$ with positive measure, such that for $(y, x^*, w) \in A(\theta)$, $f^*(y|x^*, w; \theta) \neq f^*(y|x^*, w; \theta_0)$. If the parameters are identified, then the expected (with respect to the population distribution of y, x^*, w) log likelihood has a unique and well-separated maximum in θ_0 (Van Der Vaart, 1998, Lemma 5.35).

Under weak assumptions on the distribution of the measurement error, identification of θ in $f^*(y|x^*, w; \theta)$ implies identification of θ in $f(y|x, w; \theta)$.

Theorem 1. If (i) θ_0 is identified if we observe y, x^*, w, (ii) the characteristic function of ε has a countable number of zeros, and (iii) the density of x^*, w and $f^*(y|x^*, w; \theta)$ have two absolutely integrable derivatives with respect to x^*, then θ_0 is identified if we observe y, x, w.

Proof. See Appendix.

The fact that the density of x^* given x, w is non-parametrically identified makes it possible to study, for example, non-parametric regression of y on x^*, w using data from the joint distribution of y, w and the marginal distribution of x^*. This is beyond the scope of the present paper, which considers only parametric models. However, it must be stressed that the conditional density of y given x^*, w is non-parametrically identified, so that we do not rely on functional form or distributional assumptions in the identification of θ.

3. ESTIMATION WITH MARGINAL INFORMATION

3.1. Non-parametric Fourier Inversion Estimators

Our estimator is a two-step semi-parametric estimator. The first step in the estimation is to obtain a non-parametric estimator of $g_1(x - x^*)g_2(x^*, w)$. The density g_1 of the measurement error ε has characteristic function, abbreviated as cf, $\phi_\varepsilon(t) = \frac{\phi_x(t)}{\phi_{x^*}(t)}$. The operation by which the cf of one of the random variables in a convolution is obtained from the cf of the sum and the cf of the other component is called *deconvolution*. By Fourier inversion we have, if ϕ_ε is absolutely integrable (see below),

$$g_1(x-x^*) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it(x-x^*)}\frac{\phi_x(t)}{\phi_{x^*}(t)}\,\mathrm{d}t \tag{9}$$

DOI: 10.1002/jae

The joint characteristic function of x^*, w is $\phi_{x^*w}(r,s)=\dfrac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)}$. Again Fourier inversion gives, if ϕ_{x^*w} is absolutely integrable, the joint density of x^*, w as

$$g_2(x^*,w)=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-irx^*-isw}\frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)}\,\mathrm{d}r\,\mathrm{d}s \tag{10}$$

The Fourier inversion formulas become non-parametric estimators, if we replace the cf by empirical characteristic functions (ecf). If we have a random sample x_i, $i=1,\ldots,n$ from the distribution of x, then the ecf is defined as

$$\hat{\phi}_x(t)=\frac{1}{n}\sum_{i=1}^{n}e^{itx_i} \tag{11}$$

However, the estimators that we obtain if we substitute the ecf of x and x^* in (9) and the ecf of x, w, x^* and x in (10) are not well defined. In particular, sampling variations cause the integrals not to converge. Moreover, to prove consistency of the estimators we need results on the uniform convergence of the empirical cf (as a function of t). Uniform convergence for $-\infty>t>\infty$ cannot be established. For these reasons we introduce integration limits in the definition of the non-parametric density estimators by multiplying the integrand by a weight function $K_n^*(t)$ that is 0 for $|t|>T_n$. For reasons that will become clear, we choose $K_n^*(t)=K^*\left(\frac{t}{T_n}\right)$ with K^* the Fourier transform of the function K, i.e. $K^*(t)=\int_{-\infty}^{\infty}e^{-itz}K(z)\,\mathrm{d}z$. The function K is a kernel that satisfies: (i) $K(z)=K(-z)$ and K^2 is integrable; (ii) $K^*(t)=0$ for $|t|>1$; (iii) $\int_{-\infty}^{\infty}K(z)\,\mathrm{d}z=1$, $\int_{-\infty}^{\infty}z^jK(z)\,\mathrm{d}z=0$ for $j=1,2,\ldots,q-1$, and $\int_{-\infty}^{\infty}|z|^qK(z)\,\mathrm{d}z>\infty$, i.e. K is a kernel of order q. In the nonparametric density estimator of x^*, w we multiply by a bivariate weight function $K_n^*(r,s)=K^*\left(\frac{r}{R_n},\frac{s}{S_n}\right)$ with $K^*(r,s)$ the bivariate Fourier transform of the kernel $K(v,z)$ that satisfies (i)–(ii), and (iii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K(v,z)\,\mathrm{d}z\,\mathrm{d}v=1$, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}v^kz^lK(v,z)\,\mathrm{d}z\,\mathrm{d}v=0$ if $k+1>q$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|v|^k|z|^lK(v,z)\,\mathrm{d}z\,\mathrm{d}v>\infty$ if $k+l=q$.

With these weight functions the nonparametric density estimators are

$$\hat{g}_1(x-x^*)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-it(x-x^*)}\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t)}K_n^*(t)\,\mathrm{d}t \tag{12}$$

$$\hat{g}_2(x^*,w)=\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-irx^*-isw}\frac{\hat{\phi}_{xw}(r,s)\hat{\phi}_{x^*}(r)}{\hat{\phi}_x(r)}K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r \tag{13}$$

The implicit integration limits T_n, R_n, and S_n diverge at an appropriate rate, to be defined below. Although we integrate a complex-valued function the integrals are real.[3] However, because we

[3] Because $e^{itx_j}=\cos(tx_j)+i\sin(tx_j)$ the ecf has a real part that is an even function of t and an imaginary part that is an odd function of t. Let $E_k(t)$ for $k=1,2,3,4$ be real even functions in t, i.e. $E_k(t)=E_k(-t)$, where t may be a vector. Let $O_k(t)$ for $k=1,2,3,4$ be real odd functions in t, i.e. $O_k(-t)=-O_k(t)$. For any $E_1(t)$, $E_2(t)$, $O_1(t)$, $O_2(t)$, we have that

$$[E_1(t)+iO_1(t)][E_2(t)+iO_2(t)]=E_3(t)+iO_3(t)$$

and

$$\frac{E_1(t)+iO_1(t)}{E_2(t)+iO_2(t)}=E_4(t)+iO_4(t)$$

Let the even functions and the odd functions be the real and the imaginary part of the ecf. Then the multiplication/division of ecf results in functions with an imaginary part that is an odd function of t. This implies that the imaginary part of the integrand is an odd function of t so that its integral is 0.

truncate the range of integration, the estimated densities need not be positive. Diggle and Hall (1993) suggest that this phenomenon may be due to the sharp boundary cut-off of the domain of $K_n^*(t)$. Figure 1 illustrates this for our application with $K^*(t) = (1 - |t|)I(|t| \leq 1)$.

Demonstration of consistency of the non-parametric estimators in (12) and (13) requires some restrictions on the distributions of x^* and ε. A relatively weak restriction is that the cf of ε and that of x^*, w must be absolutely integrable, i.e. $\int_{-\infty}^{\infty} |\phi_\varepsilon(t)|\mathrm{d}t > \infty$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\phi_{x^*w}(r, s)|\mathrm{d}r\ \mathrm{d}s > \infty$. A sufficient condition is that, for example, $\int_{-\infty}^{\infty} |g_1(\varepsilon)''|\mathrm{d}\varepsilon > \infty$ with g_1'' the second derivative of the pdf of ε, which is a weak smoothness condition (and an analogous condition on the joint density of x, w).

A second restriction derives from the fact that deconvolution involves division by an (empirical) characteristic function. For this reason a common assumption in the deconvolution literature is that the characteristic function in the denominator is never equal to 0. For instance, the characteristic function of the normal distribution with mean 0 has this property. This assumption is not necessary to ensure the consistency of the semiparametric MLE. However, we have been unable to prove $\sqrt{n}$ consistency of the semi-parametric MLE without it. The assumption is not innocuous, because it excludes, for example, the symmetrically truncated normal distribution (with mean 0). To ensure $\sqrt{n}$ consistency we restrict the distributions whose cf appears in the denominator to the class of range-restricted distributions. Because these distributions are asymmetrically truncated, their cf's do not have (real) zeros (see Figure 2 for a counterexample). The nonzero cf assumption is a peculiarity of the deconvolution approach to the solution of linear integral equations. Its resolution may require a different solution method for the linear integral equations that determine the densities of ε and of x^*, w. This is beyond the scope of the present paper.

Finally, to obtain a rate of convergence of the first-stage nonparametric density estimators that is fast enough to ensure $\sqrt{n}$ consistency of the semi-parametric MLE, a sufficient condition is that the characteristic functions of x^* and ε are ordinarily smooth (Fan, 1991), i.e. for large t the

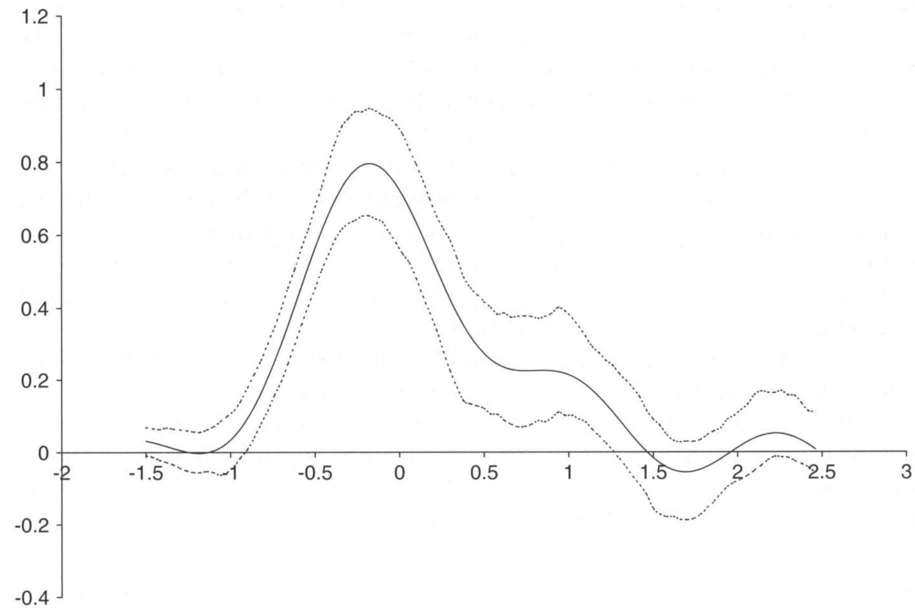

Figure 1. Estimate and 95% CI of the measurement error density ($T_n = 1$; 200 repetitions)

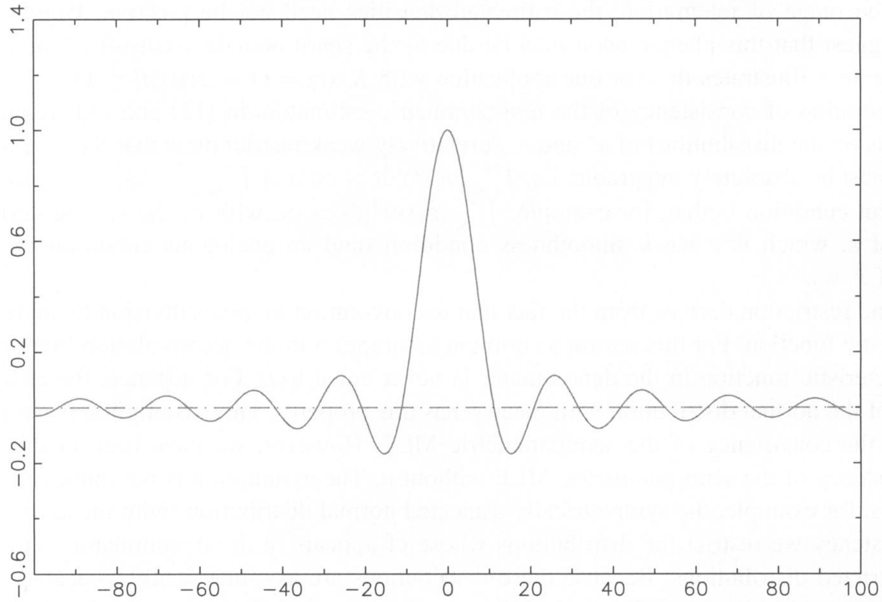

Figure 2. Characteristic function of symmetrically truncated (at −3 and 3) Laplace distribution. This figure is available in color online at wileyonlinelibrary.com/journal/jae

characteristic functions must be such that for some C_0, C_1, $k > 0$

$$C_0 t^{-(k+1)} \leq |\phi_v(t)| \leq C_1 t^{-(k+1)} \tag{14}$$

The integer k is the index of smoothness. In the deconvolution literature, assumptions on the tail behavior of characteristic functions are common. Hu and Ridder (2010) relate these assumptions to the underlying distributions. They show that a sufficient condition is that distributions of the latent true value x^* and the measurement error ε are *range-restricted* (Hu and Ridder, 2010).The distribution of a random variable v is *range restricted of order* k with $k = 0, 1, 2, \ldots$ if: (i) its density f_v has support $[L, U]$ with either L or U finite; (ii) the density f_v has $k+2$ absolutely integrable derivatives $f_v^{(j)}$; (iii) $f_v^{(j)}(U) = f_v^{(j)}(L) = 0$ for $j = 0, \ldots, k-1$ and $|f_v^{(k)}(U)| \neq |f_v^{(k)}(L)|$. This is a sufficient but not necessary condition for ordinary smoothness.[4]

If $k = 0$, then a sufficient condition for range restriction is that the density is not equal at the upper and lower truncation points. This is obviously satisfied if the truncation is one-sided, e.g. if the distribution is half normal. Furthermore, a range restricted distribution may also be obtained by truncating a distribution with unbounded support, where the bounds L and U may diverge to $-\infty$ and ∞ with the sample size going to infinity.

Because we observe the marginal distribution of x and that of x^*, one might wonder whether the assumption that the distributions of x^* and the measurement error are both range restricted together with the measurement error model has testable implications. For instance, if both x^* and the measurement error are non-negative, then x is also non-negative. If both are bounded, then x is also bounded with a support that is larger than that of x^*, if the measurement error has a support that includes both negative and positive values. If the support of ε is bounded from below

[4] 'Range-restricted' distributions do not include distributions that are 'supersmooth' like the normal distribution. In that case, Schennach (2004b) shows that fast nonparametric rates (i.e. $o_p(n^{-1/4})$) of convergence are still possible when both the error distribution and that of the latent true values are supersmooth. We do not explore that possibility in this paper.

DOI: 10.1002/jae

by a positive number, the lower bound on the support of x will be larger than that of the support of x^*. The only case that is excluded is a support of x that is a strict subset of that of x^*. Note that our assumption is compatible with the classical measurement error assumption, because we do not impose restrictions on the support of x.

The properties, and in particular the rate of uniform convergence, of the first-stage nonparametric density estimators are given by the following lemma.

Lemma 1.

(i) Let ϕ_ε be absolutely integrable and let the density of ε be q times differentiable with a q-th derivative that is bounded on its support. Suppose $|\phi_{x^*}(t)| > 0$ for all $t \in \Re$ and that the distribution of x^* is range restricted of order k_{x^*}. Let $T_n = O\left(\left(\frac{n}{\log n}\right)^\gamma\right)$ for $0 > \gamma > \frac{1}{2}$. Then a.s. if $\frac{n_1}{n} \to \lambda$ with $0 > \lambda > \infty$ for $n \to \infty$

$$\sup_{x\in\mathcal{X}, x^*\in\mathcal{X}^*} |\hat{g}_1(x - x^*) - g_1(x - x^*)| = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2}-(k_{x^*}+3)\gamma-\eta}\right) + O\left(\left(\frac{\log n}{n}\right)^{q\gamma}\right) \quad (15)$$

for $\eta > 0$ and q the order of the kernel in the density estimator.

(ii) Let $\phi_{x^*w}(t, s)$ be absolutely integrable and let the density of x^*, w be q times differentiable with all qth derivatives bounded. Suppose $|\phi_x(t)| > 0$, $|\phi_{x^*}(t)| > 0$ for all $t \in \Re$, the distribution of x^* is range restricted of order k_{x^*} and the distribution of ε is range restricted of order k_ε. Let $S_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$ and $R_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$ with $0 > \gamma' > \frac{1}{2}$. Then a.s. if $\frac{n_1}{n} \to \lambda$ with $0 > \lambda > \infty$ for $n \to \infty$

$$\sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}} |\hat{g}_2(x^*, w) - g_2(x^*, w)| = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2}-(k_{x^*}+k_\varepsilon+5)\gamma'-\eta}\right) + O\left(\left(\frac{\log n}{n}\right)^{q\gamma'}\right) \quad (16)$$

for $\eta > 0$ and q the order of the kernel in the density estimator.

Proof. See Appendix.

Note that in the bounds the first term is the variance and the second the bias term. As usual, the bias term can be made arbitrarily small by choosing a higher-order kernel. To obtain a rate of convergence of $n^{-\frac{1}{4}}$ or faster (Newey, 1994) we require that

$$\frac{1}{4q} > \gamma > \frac{1}{4(k_{x^*} + 3)} \quad (17)$$

and

$$\frac{1}{4q} > \gamma' > \frac{1}{4(k_{x^*} + k_\varepsilon + 5)} \quad (18)$$

which requires that we choose the order of the kernel q to be greater than $k_{x^*} + k_\varepsilon + 5$.

DOI: 10.1002/jae

3.2. The Semi-parametric MLE

The data consist of a random sample y_i, x_i, w_i, $i = 1, \ldots, n$ and an independent random sample x_i^*, $i = 1, \ldots, n_1$. The population density of the observations in the first sample is

$$f(y|x, w; \theta_0) = \int_{\mathcal{X}^*} f^*(y|x^*, w; \theta_0) \frac{g_1(x - x^*) g_2(x^*, w)}{g(x, w)} \, dx^* \tag{19}$$

in which $f^*(y|x^*, w; \theta)$ is the parametric model for the conditional distribution of y given w and the latent x^*. The densities f_x, f_{x^*}, $f_{w|x}$ have support $\mathcal{X}$, $\mathcal{X}^*$, $\mathcal{W}$, respectively. These supports may be bounded.

The semi-parametric MLE is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln \hat{f}(y_i|x_i, w_i; \theta) \tag{20}$$

with $\hat{f}(y_i|x_i, w_i; \theta)$ the conditional density in which we replace g_1, g_2 by their non-parametric Fourier inversion estimators. The parameter vector θ is of dimension d. The semi-parametric MLE satisfies the moment condition

$$\sum_{i=1}^{n} m(y_i, x_i, w_i, \hat{\theta}, \hat{g}_1, \hat{g}_1) = 0 \tag{21}$$

where the moment function $m(y, x, w, \theta, g_1, g_2)$ is the score of the integrated likelihood

$$m(y, x, w, \theta, g_1, g_2) = \frac{\displaystyle\int_{\mathcal{X}^*} \frac{\partial f^*(y|x^*, w; \theta)}{\partial \theta} g_1(x - x^*) g_2(x^*, w) \, dx^*}{\displaystyle\int_{\mathcal{X}^*} f^*(y|x^*, w; \theta) g_1(x - x^*) g_2(x^*, w) \, dx^*} \tag{22}$$

The next two theorems give conditions under which the semi-parametric MLE is consistent and asymptotically normal.

Theorem 2. If

(A1) The parametric model $f^*(y|x^*, w; \theta)$ is such that there are constants $0 > m_0 > m_1 > \infty$ such that for all $(y, x^*, w) \in \mathcal{Y} \times \mathcal{X}^* \times \mathcal{W}$ and $\theta \in \Theta$

$$m_0 \leq f^*(y|x^*, w; \theta) \leq m_1 \tag{23}$$

$$\left| \frac{\partial f^*(y|x^*, w; \theta)}{\partial \theta_k} \right| \leq m_1 \tag{24}$$

and that for all $(y, w) \in \mathcal{Y} \times \mathcal{W}$ and $\theta \in \Theta$

$$\int_{\mathcal{X}^*} f^*(y|x^*, w; \theta) \, dx^* > \infty \tag{25}$$

$$\left| \int_{\mathcal{X}^*} \frac{\partial f^*(y|x^*, w; \theta)}{\partial \theta_k} \, dx^* \right| > \infty \tag{26}$$

with $k = 1, \ldots, d$. The density of x, w is bounded from 0 on its support $\mathcal{X} \times \mathcal{W}$.

DOI: 10.1002/jae

(A2) The characteristic functions of ε and x^*, w are absolutely integrable and their densities q times differentiable with qth derivatives that are bounded on their support. The cf of ε and x^* do not have (real) zeros and are range-restricted of order k_ε and k_{x^*}, respectively.

(A3) $T_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma}\right)$ with $0 > \gamma > \frac{1}{2(k_{x^*}+3)}$, and $S_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$, $R_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$ with $0 > \gamma' > \frac{1}{2(k_{x^*}+k_\varepsilon+5)}$, and $\lim_{n\to\infty}\frac{n_1}{n} = \lambda$, $0 > \lambda > \infty$.

Then for the semi-parametric MLE

$$\hat{\theta} = \arg\max_{\theta\in\Theta}\sum_{i=1}^{n}\ln\hat{f}(y_i|x_i, w_i;\theta) \tag{27}$$

we have

$$\hat{\theta}\xrightarrow{p}\theta_0 \tag{28}$$

Proof. See Appendix.

Assumption (A1) is sufficient but by no means necessary. It can be replaced by boundedness assumptions on the moment function and the Fréchet differential of the moment function (see the proof in the Appendix). However, we prefer to give sufficient conditions that can be verified more easily in most applications. In some cases, e.g. if y has unbounded support, the more complicated sufficient conditions must hold.

The next lemma shows that the two-step semi-parametric MLE has an asymptotically linear representation.

Lemma 2. If the assumptions of Theorem 2 hold and in addition

(A4) $\mathrm{E}(m(y, x, w, \theta_0, g_1, g_2)m(y, x, w, \theta_0, g_1, g_2)') > \infty$.

(A5) $T_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma}\right)$ with $\frac{1}{4q} > \gamma > \frac{1}{4(k_{x^*}+3)}$, and $S_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$ and $R_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$ with $\frac{1}{4q} > \gamma' > \frac{1}{4(k_\varepsilon+k_{x^*}+5)}$.

(A6) $g_1(\varepsilon)$ has $k_\varepsilon+1$ absolutely integrable derivatives and $g_2(x^*, w)$ has $k_{x^*}+1$ absolutely integrable derivatives with respect to x^*. The range-restricted distribution of x^* has support $\mathcal{X}^* = [L, U]$ where L can be $-\infty$ or U can be ∞ and the derivatives of the marginal density of x^* satisfy $g_2^{(k)}(L) = g_{2^{(k)}}(U) = 0$ for $k = 0, \ldots, k_{x^*}-1$. We assume that the partial derivatives of the joint density of x^*, w with respect to x^* satisfy[5] $g_2^{(k)}(L, w) = g_2^{(k)}(U, w) = 0$ for $k = 0, \ldots, k_{x^*}-1$ for all $w \in \mathcal{W}$. $f^*(y|x^*, w;\theta_0)$ and $\frac{\partial f^*(y|x^*, w;\theta_0)}{\partial\theta}$ have $\max\{k_\varepsilon+1, k_{x^*}+1\}$ absolutely integrable derivatives with respect to x^*.

Then

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}m(y_j, x_j, w_j, \theta_0, \hat{g}_1, \hat{g}_2)\xrightarrow{d}N(0, \Omega) \tag{29}$$

where

$$\Omega = \mathrm{E}[\psi(y, x, w)\psi(y, x, w)'] + \lambda\mathrm{E}[\varphi(x^*)\varphi(x^*)'] \tag{30}$$

[5] This is automatically satisfied if the order of range-restriction is 0, which is the leading case.

DOI: 10.1002/jae

$$\psi(y,x,w) = m(y,x,w,\theta_0,h_0) + \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d_2^*(r,s)\frac{\phi_{x^*}(r)}{\phi_x(r)}[e^{irx+isw} - \phi_{xw}(r,s)]\,\mathrm{d}s\,\mathrm{d}r$$
$$- \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d_2^*(r,s)\frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)\phi_x(r)}[e^{irx} - \phi_x(r)]\,\mathrm{d}s\,\mathrm{d}r$$
$$+ \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{d_1^*(t)}{\phi_{x^*}(t)}[e^{itx} - \phi_x(t)]\,\mathrm{d}t \tag{31}$$

$$\varphi(x^*) = -\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{d_1^*(t)}{\phi_{x^*}(t)}\frac{\phi_x(t)}{\phi_{x^*}(t)}[e^{itx^*} - \phi_{x^*}(t)]\,\mathrm{d}t$$
$$+ \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d_2^*(r,s)\frac{\phi_{xw}(r,s)}{\phi_x(r)}[e^{itx^*} - \phi_{x^*}(t)]\,\mathrm{d}s\,\mathrm{d}r \tag{32}$$

$$d_1^*(t) = E[\kappa_1^*(t,y,x,w)] \tag{33}$$

$$d_2^*(r,s) = E[\kappa_2^*(r,s,y,x,w)] \tag{34}$$

$$\kappa_1^*(t,y,x,w) = \int e^{-it(x-x^*)}\delta(y,x,w,x^*)g_2(x^*,w)\mathrm{d}x^* \tag{35}$$

$$\kappa_2^*(r,s,y,x,w) = \int e^{-irx^*-isw}\delta(y,x,w,x^*)g_1(x-x^*)\mathrm{d}x^* \tag{36}$$

$$\delta(y,x,w,x^*) = \frac{f^*(y|x^*,w)}{f(y,x,w)}\left(\frac{\frac{\partial}{\partial\theta}f^*(y|x^*,w)}{f^*(y|x^*,w)} - \frac{\frac{\partial}{\partial\theta}f(y,x,w)}{f(y,x,w)}\right) \tag{37}$$

Proof. See Appendix.

The influence function of the semi-parametric MLE is equal to $m(y,x,w,\theta_0,h_0) + \psi(y,x,w) + \lambda\varphi(x^*)$. The term $m(y,x,w,\theta_0,h_0) + \psi(y,x,w)$ is the influence function for the survey data and the term $\varphi(x^*)$ is that for the marginal sample. The next theorem is an easy implication of the lemma.

Theorem 3. If assumptions (A1)–(A5) are satisfied, then

$$\sqrt{n}(\hat{\theta} - \theta_0)\xrightarrow{d}N(0,V) \tag{38}$$

with $V = (M')^{-1}\Omega M^{-1}$ where

$$M = \mathrm{E}\left(\frac{\partial m(y,x,w,\theta_0,g_1,g_2)}{\partial\theta'}\right) \tag{39}$$

The matrix Ω is estimated by substituting estimates for unknown parameters and empirical for population characteristic functions. The matrix Ω has a closed-form representation. Although we do not need this expression to estimate Ω, we consider it to see how zeros in the cf of x and x^* may affect the asymptotic variance. To keep the discussion simple we consider the third term of $\psi(y,x,w)$, which has a variance equal to

$$\mathrm{E}\left[\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{d_1^*(t)}{\phi_{x^*}(t)}(e^{itx} - \phi_x(t))\mathrm{d}t\right)^2\right]$$

DOI: 10.1002/jae

$$
\begin{aligned}
&= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{d_1^*(t)d_1^*(s)}{\phi_{x^*}(t)\phi_{x^*}(s)}(\phi_x(t+s)-\phi_x(t)\phi_x(s))\ \mathrm{d}t\ \mathrm{d}s \\
&= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d_1^*(t)d_1^*(s)\phi_\varepsilon(t)\phi_\varepsilon(s)\left(\frac{\phi_x(t+s)}{\phi_x(t)\phi_x(s)}-1\right)\ \mathrm{d}t\ \mathrm{d}s \qquad (40)
\end{aligned}
$$

Now if for some finite t both ϕ_x and ϕ_{x^*} are 0, while ϕ_ε is bounded from 0, then the integral may diverge and in that case the asymptotic variance is infinite.

4. A MONTE CARLO SIMULATION

This section applies the method developed above to a probit model with a mismeasured explanatory variable. The conditional density function of the probit model is

$$
\begin{aligned}
f^*(y|x^*, w;\theta) &= P(y, x^*, w;\theta)^y(1-P(y, x^*, w;\theta))^{1-y} \qquad (41)\\
P(y, x^*, w;\theta) &= \Phi(\beta_0+\beta_1 x^*+\beta_2 w)
\end{aligned}
$$

where $\theta = (\beta_0, \beta_1, \beta_2)'$ and Φ is the standard normal cdf. The true value and the error both have a normal distribution truncated at plus and minus 4 standard deviations, which is practically the same as the original normal distribution in the small sample. Four estimators are considered: (i) the ML probit estimator that uses mismeasured covariate x in the primary sample as if it were accurate, i.e. it ignores the measurement error. The MLE is not consistent. The conditional density function in this case is written as $f^*(y|x, w;\theta)$; (ii) the infeasible ML probit estimator that uses the latent true x^* as covariate. This estimator is consistent and has the smallest asymptotic variance of all estimators that we consider. The conditional density function is $f^*(y|x^*, w;\theta)$; (iii) the mixture MLE that assumes that the density function of x^* given x, w is known and that uses this density to integrate out the latent x^*. This estimator is consistent, but it is less efficient than the MLE in (i); and (iv) the semi-parametric MLE developed above that uses both the primary sample y_i, x_i, w_i, $i = 1, 2, \ldots, n$ and the secondary sample x_j^*, $j = 1, 2, \ldots, n_1$.

For each estimator, we report root mean squared error (RMSE), the average bias of estimates, and the standard deviation of the estimates over the replications.

We consider three different values of the measurement error variance: large, moderate and small (relative to the variance of the latent true value). The results are summarized in Table I. In all cases the smoothing parameters S, T are chosen as suggested in Diggle and Hall (1993). The results are quite robust against changes in the smoothing parameters, and the same is true in our application in Section 5.

Table I shows that the MLE that ignores the measurement error is significantly biased as expected. The bias of the coefficient of the mismeasured independent variable is larger than the bias of the coefficient of the other covariate or the constant. Some of the consistent estimators have a small-sample bias that is significantly different from 0. In particular, the (small-sample) biases in the new semi-parametric MLE are similar to those of the other consistent estimators.

In all cases the MSE of the infeasible MLE is (much) smaller than that of the other consistent estimators. The loss of precision is associated with the fact that x^* is not observed, but that we must integrate with respect to its distribution given x, w. It does not seem to matter that in the semi-parametric MLE this density is estimated non-parametrically, because the MSE of the estimator with a known distribution of the latent true value given x, w is only marginally smaller than that of our proposed estimator.

We also tested whether the sampling distribution of the semi-parametric MLE is normal. Figure 3 shows the empirical distribution of 400 semi-parametric MLE estimates of β_1. It is close to a normal

DOI: 10.1002/jae

Table I. Simulation results, probit model: $n = 500$, $n_1 = 600$, number of repetitions 200

	β_1			β_2			β_0		
	Root MSE	Mean bias	SD	Root MSE	Mean bias	SD	Root MSE	Mean bias	SD
$\frac{\sigma_\varepsilon^2}{\sigma_{x^*}^2} = 1.96$[a]									
Ignoring meas. error	0.6909	−0.6871*	0.0730	0.1452	0.0679*	0.1283	0.0692	−0.0340*	0.0603
True x^*	0.1464	0.0221*	0.1447	0.1310	−0.0143	0.1302	0.0598	0.0056	0.0595
Known meas. error dist.	0.2862	0.0330	0.2843	0.1498	−0.0151	0.1491	0.0712	0.0077	0.0708
Marginal information	0.3288	−0.0923	0.3156	0.1886	−0.0197	0.1876	0.0815	0.0025	0.0815
$\frac{\sigma_\varepsilon^2}{\sigma_{x^*}^2} = 1$[b]									
Ignoring meas. error	0.5386	−0.5311*	0.0894	0.1546	0.0562*	0.1441	0.0698	−0.0177*	0.0675
True x^*	0.1407	0.0025	0.1407	0.1466	0.0007	0.1466	0.0705	0.0111*	0.0696
Known meas. error dist.	0.2218	0.0152	0.2213	0.1563	−0.0046	0.1563	0.0758	0.0135*	0.0746
Marginal information	0.2481	0.0082	0.2480	0.1701	−0.0158	0.1693	0.0873	0.0163*	0.0858
$\frac{\sigma_\varepsilon^2}{\sigma_{x^*}^2} = 0.36$[c]									
Ignoring meas. error	0.2938	−0.2723*	0.1103	0.1449	0.0174*	0.1439	0.0630	−0.0132*	0.0616
True x^*	0.1384	0.0123	0.1379	0.1477	−0.0130	0.1471	0.0642	0.0031	0.0641
Known meas. error dist.	0.1711	0.0336*	0.1678	0.1518	−0.0177	0.1507	0.0655	0.0042	0.0653
Marginal information	0.1764	−0.0325*	0.1733	0.1743	−0.0634	0.1624	0.0942	0.0206*	0.0919

[a] $\beta_1 = 1$, $\beta_2 = -1$, $\beta_0 = 0.5$; $x^* \sim N(0, 0.25)$, $w \sim N(0, 0.25)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$; smoothing parameters are $T_n = 0.7$ for density of ε and $S_n = R_n = 0.6$ for joint density of x^*, w.
[b] $\beta_1 = 1$, $\beta_2 = -1$, $\beta_0 = 0.5$; $x^* \sim N(0, 0.25)$, $w \sim N(0, 0.25)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$; smoothing parameters are $T_n = 0.6$ for density of ε and $S_n = R_n = 0.7$ for joint density of x^*, w.
[c] $\beta_1 = 1$, $\beta_2 = -1$, $\beta_0 = 0.5$; $x^* \sim N(0, 0.25)$, $w \sim N(0, 0.25)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$; smoothing parameters are $T_n = 0.75$ for density of ε and $S_n = R_n = 0.2$ for joint density of x^*, w.

density with the same mean and variance. The p-value of the normality test, the Shapiro–Wilk W test, is 0.21, and therefore one cannot reject the hypothesis that the distribution of $\hat{\beta}_1$ is normal.

The computation of the Fourier inversion estimators in the simulation involves one-dimensional (distribution of ε) and two-dimensional (distribution of x^*, w) numerical integrals. In the simulations these are computed by Gauss–Laguerre quadrature. In the empirical application in Section 5 the second estimator involves a numerical integral of a dimension equal to the number of covariates in w plus 1. This numerical integral is computed by the Monte Carlo method (100 draws).

5. AN EMPIRICAL APPLICATION: THE DURATION OF WELFARE SPELLS

5.1. Background

The Aid to Families with Dependent Children (AFDC) program was created in 1935 to provide financial support to families with children who were deprived of the support of one biological parent by reason of death, disability, or absence from the home, and were under the care of the other parent or another relative. Only families with income and assets lower than a specified level

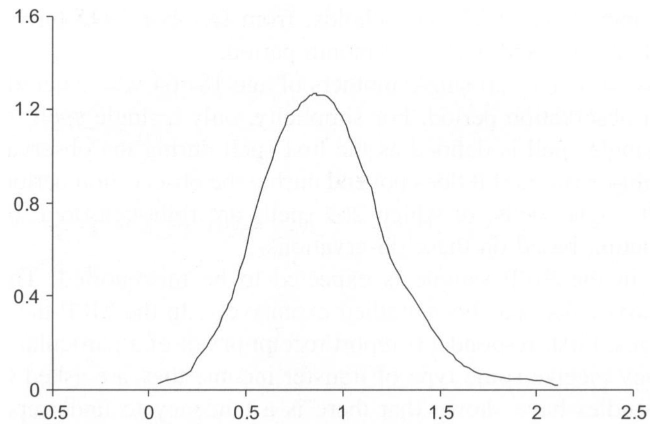

Figure 3. Sampling distribution of SPMLE of β_1 (200 repetitions)

are eligible. The majority of families of this type are single-mother families, consisting of a mother and her children. The AFDC benefit level is determined by maximum benefit level, the so-called guarantee, and deductions for earned income, child care, and work-related expenses. The maximum benefit level varies across the states, while the benefit reduction rate, sometimes called the tax rate, is set by the federal government. For example, the benefit reduction rate on earnings was reduced to 67% from 100% in 1967 and was raised back to 100% in 1981. AFDC was eliminated in 1996 and replaced by Temporary Assistance for Needy Families (TANF).

A review of the research on AFDC can be found in Moffitt (1992, 2002). In this application, we investigate to what extent the characteristics of the recipients, external economic factors, and the level of welfare benefits received influence the length of time spent on welfare. Most studies on welfare spells (Bane and Ellwood, 1994; Ellwood, 1986; O'Neill *et al.*, 1984; Blank, 1989; Fitzgerald, 1991) find that the level of benefits is negatively and significantly related to the probability of leaving welfare. Almost all studies use the AFDC guarantee rather than the reported benefit level as the independent variable. One reason for not using the reported benefit level is the fear of biases due to reporting error. The AFDC guarantee has less variation than the actual benefit level, as the AFDC guarantee is the same for all families with the same number of people who live in a particular state.

5.2. Data

The primary sample used here is extracted from the Survey of Income and Program Participation, a longitudinal survey that collects information on topics such as income, employment, health insurance coverage, and participation in government transfer programs. The SIPP population consists of persons resident in US households and persons living in group quarters. People selected for the SIPP sample are interviewed once every 4 months over the observation period. Sample members within each panel are randomly divided into four rotation groups of roughly equal size. Each month, the members of one rotation group are interviewed and information is collected about the previous 4 months, which are called reference months. Therefore, all rotation groups are interviewed every 4 months so that we have a panel with quarterly waves.

We use the 1992 and 1993 SIPP panels, each of which contains nine waves.[6] The SIPP 1992 panel follows 21,577 households from October 1991 through December 1994. The SIPP 1993

[6] The 1992 panel actually has 10 waves, but the 10th wave is only available in the longitudinal file. The original wave files are used here instead of the longitudinal file.

DOI: 10.1002/jae

panel contains information on 21,823 households, from October 1992 through December 1995. Each sample member is followed over a 36-month period.

We consider a flow sample of all single mothers of age 18–64 who entered the AFDC program during the 36-month observation period. For simplicity, only a single spell for each individual is considered here. A single spell is defined as the first spell during the observation period for each mother. A spell is right-censored if it does not end during the observation period. The SIPP duration sample contains 520 single spells, of which 269 spells are right-censored. Figure 4 presents the empirical hazard function based on these observations.

The benefit level in the SIPP sample is expected to be misreported. The reporting error in transfer income in survey data has been studied extensively. In the SIPP the reporting of transfer income is in two stages. First, respondents report receipt or not of a particular form of income, and if they report that they receive some type of transfer income they are asked the amount that they receive. Validation studies have shown that there is a tendency to underreport receipt, although for some types there is also evidence of overreporting receipt. The second source of measurement error is the response error in the amount of transfer income. Several studies find significant differences between survey reports and administrative records, but there are also studies that find little difference between reports and records. Most studies find that transfer income is underreported, and underreporting is particularly important for the AFDC program. A review of the research can be found in Bound *et al.* (2001).

The AFDC QC is a repeated cross-section that is conducted every month. Every month each state reports benefit amounts, last opening dates and other information from the case records of a randomly selected sample of the cases receiving cash payments in that state. Hence for the QC sample we know not only the true benefit level of a welfare recipient but also when the current welfare spell started. Therefore we can select from the QC sample all the women who enter the program in a particular month. The QC sample used here is restricted to the same population as the SIPP sample, which is all single mothers of age 18–64 who entered the program during the period from October 1991 to December 1995.

Because the welfare recipients can enter welfare in any month during the 51-month observation period, the distribution of the true benefits given the reported benefits and the other independent variables could be different for each of the 51 months. For instance, the composition of the families who go on welfare could have a seasonal or cyclical pattern. If this were the case we would have

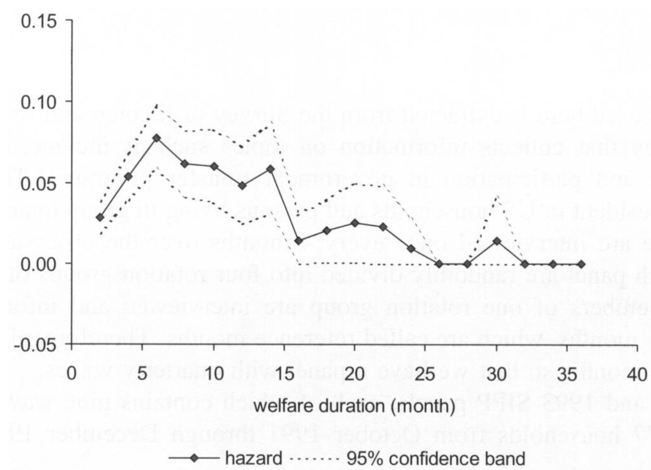

Figure 4. Empirical hazard rate of welfare durations in SIPP. This figure is available in color online at wileyonlinelibrary.com/journal/jae

DOI: 10.1002/jae

to estimate 51 distributions. Although this is feasible it is preferable to investigate first whether we can do with fewer. We test whether the distribution of the benefits is constant over the 51 months of entry or, if suspect cyclical shifts, the 4 years of the observation period. Table II reports the Kruskal–Wallis test for the null hypothesis of a constant distribution over the entry months (first row) and the entry years (second row). Table III reports the results of the Kolmogorov–Smirnov test of the hypothesis that the distribution of the welfare benefits in a particular month is the same as that in all other 50 months. The conclusion is that it is allowed to pool the 51 entry months and to estimate a single distribution of the true benefits given the reported benefits and the other independent variables.[7]

Since both the SIPP and AFDC QC samples come from the same population, we can compare the distributions of the nominal benefit levels in the two samples. Figure 5 shows the estimated density of log nominal benefit levels and Table IV reports summary statistics and the result of the Kolmogorov–Smirnov test of equality of the two distributions. A comparison of the

Table II. Stationarity of distribution of nominal benefits in QC sample: Kruskal–Wallis test, $n = 3318$

	Kruskal–Wallis statistic	Degrees of freedom	p-value
Nominal benefits between months	57.2	50	0.2254
Nominal benefits between years	6.1	4	0.1948

Table III. Stationarity of distribution nominal benefit levels in QC sample: Kolmogorov–Smirnov test distribution in indicated month vs. other months

Month	# obs.	K-S stat.	p-value	Month	# obs.	K-S stat.	p-value
1	82	0.077	0.725	27	80	0.078	0.727
2	82	0.062	0.923	28	48	0.094	0.793
3	75	0.105	0.391	29	67	0.120	0.301
4	64	0.082	0.798	30	67	0.112	0.383
5	67	0.106	0.455	31	63	0.096	0.623
6	63	0.089	0.711	32	54	0.137	0.273
7	58	0.127	0.319	33	62	0.091	0.694
8	55	0.172**	0.082	34	87	0.073	0.754
9	70	0.093	0.593	35	68	0.204*	0.008
10	68	0.071	0.889	36	66	0.119	0.317
11	68	0.120	0.293	37	68	0.136	0.168
12	67	0.076	0.840	38	81	0.090	0.551
13	69	0.142	0.132	39	62	0.146	0.151
14	59	0.102	0.589	40	45	0.117	0.573
15	61	0.123	0.329	41	72	0.057	0.975
16	62	0.110	0.449	42	50	0.141	0.279
17	57	0.103	0.594	43	61	0.137	0.208
18	47	0.106	0.677	44	55	0.166	0.101
19	59	0.074	0.905	45	68	0.113	0.364
20	52	0.105	0.623	46	57	0.110	0.507
21	43	0.109	0.694	47	63	0.088	0.724
22	69	0.125	0.242	48	83	0.117	0.221
23	70	0.041	1.000	49	80	0.140**	0.092
24	69	0.128	0.220	50	62	0.081	0.822
25	76	0.092	0.562	51	73	0.114	0.312
26	64	0.138	0.180				

Note: Significant at * 5% level; ** 10% level.

[7] In Table III we reject the null hypothesis once for the 51 tests. Although the test statistics are not independent, a rejection in a single case is to be expected.

DOI: 10.1002/jae

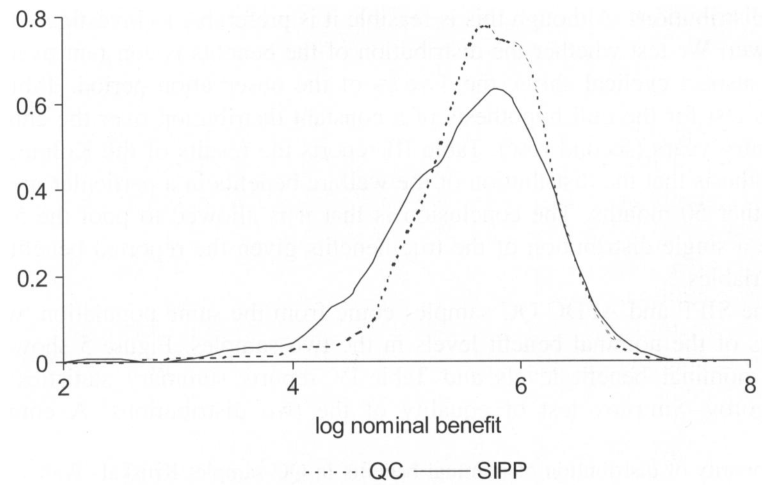

Figure 5. Density estimates of log benefits in SIPP and QC

Table IV. Comparison of the distribution of welfare benefits in SIPP and QC samples

	Real benefits		Nominal benefits	
	SIPP	QC	SIPP	QC
Mean	285.3	303.8	304.2	327.7
SD	169.6	156.9	180.9	169.4
Min.	9.3	9.6	10	10
Max.	959	1598	1025	1801
Skewness	1.08	1.27	1.07	1.33
Kurtosis	4.60	6.83	4.54	7.46
n	520	3318	520	3318
K-S statistic		0.123		0.128
p-value	0.0000		0.0000	

estimated densities and the sample means shows that benefits are indeed underreported. Indeed the Kolmogorov–Smirnov test confirms that the distribution in the SIPP sample is significantly different from the distribution in the AFDC QC. The variance of welfare benefits in the SIPP is larger than in the AFDC QC, which is a necessary condition for classical measurement error in the log benefits.

5.3. The Model and Estimation

We use a discrete duration model to analyze the grouped duration data, since the welfare duration is measured to the nearest month. As mentioned before, we consider a flow sample, and therefore we do not need to consider the sample selection problem that arises with stock sampling (Ridder, 1984). Let $[0, M]$ be the observation period, and let $t_{i0} \in [0, M]$ denote the month that individual i enters the welfare program, and $t_{i1} \in [0, M]$ the month that she leaves, if she leaves welfare during the observation period. If t_i^* is the length of the welfare spell in months, then the event t_{i0}, t_{i1} is equivalent to

$$t_{i1} - t_{i0} - 1 \leq t_i^* \leq t_{i1} - t_{i0} + 1 \tag{42}$$

DOI: 10.1002/jae

Also if the welfare spell is censored in month M, then

$$t_i^* \geq M - t_{i0} \tag{43}$$

Hence the censoring time is determined by the month of entry. We assume that this censoring time is independent of the welfare spell conditional on the (observed) covariates z_i and this is equivalent to the assumption that the month of entry is independent of the welfare spell conditional on these covariates.

The primary sample contains t_{i0}, t_{i1}, z_i, δ_i where δ_i is the censoring indicator. The latent t_i^* has a continuous conditional density that is assumed to be independent of the starting time, t_{i0}, conditional on the vector of observed covariates z_i. Let $\lambda(t, z, \theta)$ be a parametric hazard function and let $P_m(z_i, \theta)$ denote the probability that a welfare spell lasts at least m months, given that it has lasted $m - 1$ months. Then

$$P_m(z_i, \theta) = P(t_i^* \geq m | t_i^* \geq m - 1, z_i) = \exp\left(-\int_{m-1}^{m} \lambda(t, z_i, \theta)\, \mathrm{d}t\right) \tag{44}$$

If we allow for censored spells, the conditional density function for individual i with welfare spell t_i is

$$f^*(t_i, \delta_i, |z_i; \theta) = [1 - P_{t_i}(z_i, \theta)]^{\delta_i} \prod_{m=1}^{t_i - 1} P_m(z_i, \theta) \tag{45}$$

The hazard is specified as a proportional hazard model with a piece-wise constant baseline hazard:

$$\lambda(t, z_i, \theta) = \lambda_m \exp(z_i \beta), \quad m - 1 \leq t > m \tag{46}$$

This hazard specification implies that

$$P_m(z_i, \theta) = \exp[-\lambda_m \exp(z_i \beta)] \tag{47}$$

If the λ_m are unrestricted, then the covariates z_i cannot contain a constant term. For simplicity, define $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_M)'$. The unknown parameters then are $\theta = (\beta', \lambda')'$.

The covariates are $z_i = (x_i^*, w_i')'$, where the scalar x_i^* is the log real benefit level and the vector w_i contains the other covariates. The log real benefit level is defined as

$$x_i^* = \widetilde{x}_i^* - p, \tag{48}$$

where $\widetilde{x}_i^*$ is the log nominal benefit level and p is the log of the deflator.[8]

The measurement error ε_i is i.i.d. and the measurement error model is

$$\widetilde{x}_i = \widetilde{x}_i^* + \varepsilon_i, \quad \varepsilon_i \perp t_i, z_i, \delta_i \tag{49}$$

where $\widetilde{x}_i$ is the log reported nominal benefit level and ε_i is the individual reporting error. Note that error ε_i is not assumed to have a zero mean, and a non-zero mean can be interpreted as a systematic reporting error.

The variables involved in estimation are summarized in Table V. The MLE are reported in Table VI. We report the biased MLE that ignores the reporting error in the welfare benefits and the semi-parametric MLE that uses the marginal information in the AFDC QC. Note that the coefficient on the benefit level is larger for the semi-parametric MLE. This is in line with the bias

[8] We take the consumer price level as the deflator. We match the deflator to the month for which the welfare benefits are reported.

DOI: 10.1002/jae

Table V. Descriptive statistics, $n = 520$

	Mean	SD	Min.	Max.
Welfare spell (month)	9.07	8.25	1	35
Fraction censored	0.52	—	0	1
Age (years)	31.8	8.2	18	54
Disabled	0.84	—	0	1
Labor hours per week	13.3	17.6	0	70
Log real welfare benefits (week)	5.46	0.68	2.23	6.86
Log nominal welfare benefits (week)	5.52	0.68	2.30	6.93
Number of children under 18	1.92	1.02	1	7
Number of children under 5	0.60	0.76	0	4
Real non-benefits income ($1000/week)	0.234	0.402	0	0.360
State unemployment rate (%)	6.72	1.41	2.9	10.9
Education (years)	11.6	2.64	0	18

Table VI. Parameter estimates of duration model, $n = 520$, $n_1 = 3318$

Variable	MLE with marginal information		MLE ignoring measurement error	
	MLE	SE	MLE	SE
Log real benefits	−0.3368	0.1025	−0.2528	0.0877
Hours worked per week	0.2828	0.0955	0.2828	0.0938
Real non-benefits inc.	0.1891	0.1425	0.1842	0.1527
No. of children age >5	−0.1855	0.1095	−0.1809	0.1111
No. of children age >18	0.0724	0.0674	0.0712	0.0718
Years of education (/24)	−0.1803	0.9877	−0.3086	0.9663
Age (years/100)	−0.0692	0.0505	−0.0691	0.0481
State unempl. rate (%)	0.0112	0.0295	0.0082	0.0290
Disabled	−0.1093	0.1833	−0.1198	0.1867
Baseline hazard (weeks)				
1	0.0516	0.0097	0.0546	0.0105
2	0.0662	0.0120	0.0697	0.0127
3	0.0409	0.0097	0.0429	0.0104
4	0.1385	0.0203	0.1445	0.0211
5	0.0433	0.0121	0.0450	0.0128
6	0.0771	0.0169	0.0798	0.0177
7	0.0543	0.0151	0.0562	0.0156
8	0.0646	0.0180	0.0668	0.0186
9	0.0787	0.0211	0.0807	0.0217
10	0.0565	0.0189	0.0575	0.0195
11	0.0480	0.0184	0.0486	0.0186
12	0.0750	0.0250	0.0756	0.0252
13–14	0.0438	0.0146	0.0440	0.0144
15–16	0.0226	0.0113	0.0227	0.0114
17–18	0.0286	0.0143	0.0285	0.0143
19–20	0.0263	0.0152	0.0261	0.0150
21+	0.0116	0.0058	0.0114	0.0055

Note: The smoothing parameters are: distribution ε, $T_n = 0.7$, distribution of x^*, w, $S_n = 0.875$ and $R_n = 0.9$.

that we would expect in a linear model with a mismeasured covariate.[9] The other coefficients and the baseline hazard seem to be mostly unaffected by the reporting error. This may be due to the fact that the measurement error in this application is relatively small.

[9] There are no general results on the bias in nonlinear models and the bias could have been away from 0.

DOI: 10.1002/jae

6. CONCLUSION

This paper considers the problem of consistent estimation of nonlinear models with mismeasured explanatory variables, when marginal information on the true values of these variables is available. The marginal distribution of the true variables is used to identify the distribution of the measurement error, and the distribution of the true variables conditional on the mismeasured variables and the other explanatory variables. The estimator is shown to be $\sqrt{n}$ consistent and asymptotically normally distributed. The simulation results are in line with the asymptotic results. The semi-parametric MLE is applied to a duration model of AFDC welfare spells with misreported welfare benefits. The marginal distribution of welfare benefits is obtained from the AFDC Quality Control data. We find that the MLE that ignores the reporting error underestimates the effect of welfare benefits on probability of leaving welfare.

REFERENCES

Amemiya Y, Fuller WA. 1988. Estimation for the nonlinear functional relationship. *Annals of Statistics* **16**: 147–160.

Angrist J, Krueger A. 1999. Empirical strategies in labor economics. In *Handbook of Labor Economics*, Vol. 3A, Ashenfelter O, Card D (eds). North-Holland: Amsterdam; 1277–1366.

Augustin T. 2004. An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics* **31**: 43–50.

Bane MJ, Ellwood D. 1994. Understanding welfare dynamics. In *Welfare Realities: From Rhetoric to Reform*, Bane MJ, Ellwood D (eds). Harvard University Press: Cambridge, MA; 28–66.

Barron JM, Berger MC, Black DA. 1997. *On the Job Training*. WE Upjohn Institute for Employment Research: Kalamazoo, MI.

Bickel PJ, Ritov Y. 1987. Efficient estimation in the errors in variables model. *Annals of Statistics* **15**: 513–540.

Blank R. 1989. Analyzing the length of welfare spells. *Journal of Public Economics* **39**: 245–273.

Bound J, Brown C, Duncan GJ, Rodgers WL. 1989. *Measurement error in cross-sectional and longitudinal labor marker surveys: results from two validation studies*. NBER Working Paper 2884.

Bound J, Brown C, Duncan GJ, Rodgers WL. 1994. Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics* **12**: 345–368.

Bound J, Brown C, Mathiowetz N. 2001. Measurement error in survey data. In *Handbook of Econometrics*, Vol. 5, Heckman JJ, Leamer E (eds). Elsevier: Amsterdam; 3705–3843.

Card D, Hildreth A, Shore-Sheppard L. 2001. *The measurement of Medicaid coverage in the SIPP: evidence from California, 1990–1996*. NBER Working Paper 8514.

Carroll RJ, Stefanski LA. 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* **85**: 652–663.

Carroll RJ, Wand MP. 1991. Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society B* **53**: 573–585.

Carroll RJ, Ruppert D, Stefanski LA. 1995. *Measurement Error in Nonlinear Models*. Chapman & Hall: New York.

Chen X, Hong H, Tamer E. 2005. Measurement error models with auxiliary data. *Review of Economic Studies* **72**: 343–366.

Cheng CL, Van Ness JW. 1994. On estimating linear relationships when both variables are subject to errors. *Journal of Royal Statistical Society B* **56**: 167–183.

Code J. 1992. Using administrative record information to evaluate the quality of the income data collected in the survey of income and program participation. In *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, Ottawa; 295–306.

Cohen S, Carlson B. 1994. A comparison of household and medical provider reported expenditures in the 1987 NMES. *Journal of Official Statistics* **10**: 3–29.

Dibbs R, Hale A, Loverock R, Michaud S. 1995. *Some effects of computer assisted interviewing on the data quality of the survey of labour and income dynamics*. SLID research paper, series no. 95-07, Statistics Canada, Ottawa.

Diggle PJ, Hall P. 1993. A Fourier approach to nonparametric deconvolution of a density estimate. *Journal of Royal Statistical Society B* **55**: 523–531.

DOI: 10.1002/jae

Ellwood D. 1986. Targeting *'Would be' long-term recipients of AFDC*. Mathematica Policy Research, Princeton, NJ.

Fan J. 1991. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics* **19**: 1257–1272.

Fitzgerald J. 1991. Welfare durations and the marriage marker: evidence from the Survey of Income and Program Participation. *Journal of Human Resources* **26**: 545–561.

Grondin C, Michaud S. 1994. Data quality of income data using computer-assisted interview: the experience of the Canadian survey of labour and income dynamics. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA; 830–835.

Hausman J, Ichimura H, Newey W, Powell J. 1991. Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics* **50**: 273–295.

Hausman J, Newey W, Powell J. 1995. Nonlinear errors in variables: estimation of some Engel curves. *Journal of Econometrics* **65**: 205–233.

Horowitz J, Markatou M. 1996. Semiparametric estimation of regression models for panel data. *Review of Economic Studies* **63**: 145–168.

Hsiao C. 1989. Consistent estimation for some nonlinear errors-in-variables models. *Journal of Econometrics* **41**: 159–185.

Hsiao C. 1991. Identification and estimation of dichotomous latent variables models using panel data. *Review of Economic Studies* **58**: 717–731.

Hsiao C, Wang QK. 2000. Estimation of structural nonlinear errors-in-variables models by simulated least-squares method. *International Economic Review* **41**: 523–542.

Hu Y, Ridder G. 2010. On deconvolution as a first stage nonparametric estimator. *Econometric Reviews* **29**: 1–32.

Hu Y, Schennach S. 2008. Instrumental variable treatment of nonclassical measurement error models. *Econometrica* **76**: 195–216.

Johnson A, Sanchez ME. 1993. Household and medical provider reports on medical conditions: national medical expenditure survey 1987. *Journal of Economic and Social Measurement* **19**: 199–223.

Kane TJ, Rouse CE, Staiger D. 1999. *Estimating the returns to schooling when schooling is misreported*. NBER Working Paper 7235.

Kong FH, Gu M. 1999. Consistent estimation in Cox proportional hazards model with covariate measurement errors. *Statistica Sinica* **9**: 953–969.

Lee LF, Sepanski JH. 1995. Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association* **90**: 130–140.

Li T, Vuong Q. 1998. Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis* **65**: 139–165.

Luenberger DG. 1969. *Optimization by Vector Space Methods*. Wiley: New York.

Mathiowetz N, Duncan G. 1988. Out of work, out of mind: response errors in retrospective reports of unemployment. *Journal of Business and Economic Statistics* **6**: 221–229.

Moffitt R. 1992. Incentive effects of the U.S. welfare system: a review. *Journal of Economic Literature* **30**: 1–61.

Moffitt R. 2002. *The Temporary Assistance for Needy Families Program*. NBER working paper 8749.

Murphy SA, Van Der Vaart AW. 1996. Likelihood inference in the errors-in-variables model. *Journal of Multivariate Analysis* **59**: 81–108.

Newey W. 1994. The asymptotic variance of semiparametric estimators. *Econometrica* **62**: 1349–1382.

Newey W. 2001. Flexible simulated moment estimation of nonlinear errors-in-variables models. *Review of Economics and Statistics* **83**: 616–627.

O'Neill J, Wolf D, Bassi L, Hannan M. 1984. *An analysis of time on welfare*. Urban Institute: Washington, DC.

Pepe MS, Fleming TR. 1991. A general nonparametric method for dealing with errors in missing or surrogate covariate data. *Journal of the American Statistical Association* **86**: 108–113.

Pollard D. 1984. *Convergence of Stochastic Processes*. Springer: New York.

Ridder G. 1984. The distribution of single-spell duration data. In *Studies in Labor Market Analysis*, Neumann GR, Westergard-Nielsen N (eds). Springer: Berlin; 45–74.

Ridder G, Moffitt R. 2007. The econometrics of data combination. In *Handbook of Econometrics*, Vol. 6, Part 2, Heckman J, Leamer E (eds). North-Holland: Amsterdam; 5469–5547.

Schennach S. 2004a. Estimation of nonlinear models with measurement error. *Econometrica* **72**: 33–75.

Schennach S. 2004b. Nonparametric estimation in the presence of measurement error. *Econometric Theory* **20**: 1046–1093.

DOI: 10.1002/jae

Schennach S. 2007. Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* **75**: 201–239.
Taupin ML. 2001. Semi-parametric estimation in the nonlinear structural errors-in-variables model. *Annals of Statistics* **29**: 66–93.
Van Der Vaart AW. 1998. *Asymptotic Statistics*. Cambridge University Press: Cambridge, UK.
Wang L. 1998. Estimation of censored linear errors-in-variables models. *Journal of Econometrics* **84**: 383–400.

APPENDIX

Proof of Theorem 1

Assume that θ is observationally equivalent to θ_0. Then for all y, w, x

$$\begin{aligned} & f(y|x, w; \theta) - f(y|x, w; \theta_0) \\ &= \int_{\mathcal{X}^*} (f^*(y|x^*, w; \theta) - f^*(y|x^*, w; \theta_0)) g(x^*|x, w)\, \mathrm{d}x^* \equiv 0 \end{aligned} \tag{50}$$

After substitution of (3) and (4) and a change of variable in the integration, this is equivalent to

$$\int_{\mathcal{E}} (f^*(y|x - \varepsilon, w; \theta) - f^*(y|x - \varepsilon, w; \theta_0)) g_2(x - \varepsilon, w) g_1(\varepsilon)\, \mathrm{d}\varepsilon \equiv 0 \tag{51}$$

By the convolution theorem this implies that

$$h^*(t, y, w, \theta)\phi_\varepsilon(t) = 0 \tag{52}$$

for all t, y, w, θ, with

$$h^*(t, y, w, \theta) = \int_{\mathcal{X}^*} e^{itx^*} h(y, w, x^*, \theta)\, \mathrm{d}x^* \tag{53}$$

and

$$h(y, w, x^*, \theta) = (f^*(y|x^*, w; \theta) - f^*(y|x^*, w; \theta_0)) g_2(x^*, w) \tag{54}$$

so that

$$h^*(t, y, w, \theta) \equiv 0 \tag{55}$$

except possibly for a countable number of values of t. Because $h^*(t, y, w, \theta)$ is absolute integrable with respect to t under the assumptions, we have by the Fourier inversion theorem that for all y, w, x^*, θ

$$h(y, w, x^*, \theta) = (f^*(y|x^*, w; \theta) - f^*(y|x^*, w; \theta_0)) g_2(x^*, w) = 0 \tag{56}$$

Hence on the support of x^*, w we have

$$f^*(y|x^*, w; \theta) = f^*(y|x^*, w; \theta_0) \tag{57}$$

so that $\theta = \theta_0$. □

The next lemma gives an almost sure rate of convergence for the empirical characteristic function without any restriction on the support of the distribution, which is related to Lemma 6 in Schennach (2004a). It can be compared to the result in Lemma 1 of Horowitz and Markatou (1996).

Lemma 3.

(i) Let $\hat{\phi}(t) = \int_{-\infty}^{\infty} e^{itx}\, \mathrm{d}F_n(x)$ be the empirical characteristic function of a random sample from a distribution with cdf F and with $\mathrm{E}(|x|) > \infty$. For $0 > \gamma > \frac{1}{2}$, let $T_n = o\left(\left(\frac{n}{\log n}\right)^{\gamma}\right)$. Then

$$\sup_{|t| \le T_n} |\hat{\phi}(t) - \phi(t)| = o(\alpha_n) \qquad \text{a.s.} \tag{58}$$

DOI: 10.1002/jae

with $\alpha_n = o(1)$ and $\dfrac{\left(\dfrac{\log n}{n}\right)^{\frac{1}{2}-\gamma}}{\alpha_n} = O(1)$, i.e. the rate of convergence is at most $\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma}$.

(ii) Let $\hat{\phi}(s, t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{isx+ity}\,\mathrm{d}F_n(x, y)$ be the empirical characteristic function of a random sample from a bivariate distribution with cdf F and with $\mathrm{E}(|x| + |y|) > \infty$. For $0 > \gamma' > \frac{1}{2}$, let[10] $R_n = o\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$ and $S_n = o\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$. Then

$$\sup_{|r|\le R_n|,s|\le S_n} \left|\hat{\phi}(r, s) - \phi(r, s)\right| = o(\alpha_n) \qquad \text{a.s.} \tag{59}$$

with $\alpha_n = o(1)$ and $\dfrac{\left(\dfrac{\log n}{n}\right)^{\frac{1}{2}-\gamma'}}{\alpha_n} = O(1)$, i.e. the rate is the same as in the one-dimensional case.

The lemma ensures that the Fourier inversion estimators $\hat{g}_1$ and $\hat{g}_2$ are well defined if n is sufficiently large, because the denominators of the integrands are bounded from 0 except possibly on a set that has probability 0.

Proof of Lemma 3

For part (i) consider the parametric class of functions $\mathcal{G}_n = \{e^{itx}||t| \le T_n\}$. The first step is to find the L_1 covering number of $\mathcal{G}_n$. Because $e^{itx} = \cos(tx) + i\sin(tx)$, we need covers of $\mathcal{G}_{1n} = \{\cos(tx)||t| \le T_n\}$ and $\{\mathcal{F}_{2n} = \sin(tx)||t| \le T_n\}$. Because $|\cos(t_2x) - \cos(t_1x)| \le |x||t_2 - t_1|$ and $\mathrm{E}(|x|) > \infty$, an $\frac{\varepsilon}{2}\mathrm{E}(|x|)$ cover (with respect to the L_1 norm) of $\mathcal{G}_{1n}$ is obtained from an $\frac{\varepsilon}{2}$ cover of $\{t||t| \le T_n\}$ by choosing t_k, $k = 1, \ldots, K$ arbitrarily from the distinct covering sets, where K is the smallest integer larger than $\frac{2T_n}{\varepsilon}$. Because $|\sin(t_2x) - \sin(t_1x)| \le |x||t_2 - t_1|$, the functions $\sin(t_kx)$, $k = 1, \ldots, K$ are an $\frac{\varepsilon}{2}\mathrm{E}(|x|)$ cover of $\mathcal{F}_{2n}$. Hence $\cos(t_kx) + i\sin(t_kx)$, $k = 1, \ldots, K$ is an $\varepsilon\mathrm{E}(|x|)$ cover of $\mathcal{G}_n$, and we conclude that

$$\mathcal{N}_1(\varepsilon, P, \mathcal{G}_n) \le A\frac{T_n}{\varepsilon} \tag{60}$$

with P an arbitrary probability measure such that $\mathrm{E}(|x|) > \infty$ and $A > 0$, a constant that does not depend on n. The next step is to apply the argument that leads to Theorem 2.37 in Pollard (1984). The theorem cannot be used directly, because the condition $\mathcal{N}_1(\varepsilon, P, \mathcal{G}_n) \le A\varepsilon^{-W}$ is not met. In Pollard's proof we set $\delta_n = 1$ for all n, and $\varepsilon_n = \varepsilon\alpha_n$. Equations (30) and (31) in Pollard (1984, p. 31) are valid for $\mathcal{N}_1(\varepsilon, P, \mathcal{G}_n)$ defined above. Hence we have as in Pollard's proof using his (31)

$$\Pr(\sup_{|t|\le T_n} |\hat{\phi}(t) - \phi(t)| > 2\varepsilon_n) \le 2A\left(\frac{\varepsilon_n}{T_n}\right)^{-1} \exp\left(-\frac{1}{128}n\varepsilon_n^2\right) + \Pr(\sup_{|t|\le T_n} \hat{\phi}(2t) > 64) \tag{61}$$

The second term on the right-hand side is obviously 0. The first term on the right-hand side is bounded by

$$2A\varepsilon^{-1}\exp\left(\log\left(\frac{T_n}{\alpha_n}\right) - \frac{1}{128}n\varepsilon^2\alpha_n^2\right) \tag{62}$$

[10] We could allow for different growth in S_n and T_n, but nothing is gained by this.

The restrictions on α_n and T_n imply that $\frac{T_n}{\alpha_n} = o\left(\sqrt{\frac{n}{\log n}}\right)$, and hence $\log\left(\frac{T_n}{\alpha_n}\right) - \frac{1}{2}\log n \to -\infty$. The same restrictions imply that $\frac{n\alpha_n^2}{\log n} \to \infty$. The result now follows from the Borel–Cantelli lemma.

For part (ii) we note that the $\frac{\varepsilon}{2}$ covers of $|s| \le S_n$ and $|r| \le R_n$ generate $\frac{\varepsilon}{2}\mathrm{E}(|x| + |y|)$ covers of $\cos(sx + ty)$, and $\sin(sx + ty)$ and an $\varepsilon\mathrm{E}(|x| + |y|)$ cover of $e^{isx+ity}$. Hence (60) becomes

$$\mathcal{N}_1(\varepsilon, P, \mathcal{G}_n) \le A\frac{R_n S_n}{\varepsilon^2} \tag{63}$$

Hence in (61) we must replace $\frac{\varepsilon_n}{T_n}$ by $\frac{\varepsilon_n}{R_n}\frac{\varepsilon_n}{S_n}$ and in the next equation $\log\left(\frac{T_n}{\alpha_n}\right)$ by $\log\left(\frac{R_n}{\alpha_n}\right) + \log\left(\frac{S_n}{\alpha_n}\right)$. □

Lemma 3 suggests that we can choose $T_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma}\right)$, $R_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$, $S_n = O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$, and $\alpha_n = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma'-\eta}\right)$ for any arbitrarily small $\eta > 0$.

Proof of Lemma 1

(i) Define $\varepsilon = x - x^*$. Then

$$\sup_{\varepsilon\in\mathcal{E}} |\hat{g}_1(\varepsilon) - g_1(\varepsilon)| \le \sup_{\varepsilon\in\mathcal{E}} \left| \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it\varepsilon}\left(\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right) K_n^*(t)\,\mathrm{d}t\right|$$
$$+ \sup_{\varepsilon\in\mathcal{E}} \left|\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it\varepsilon}\phi_\varepsilon(t)[1 - K_n^*(t)]\,\mathrm{d}t\right| \tag{64}$$

We give bounds on the terms that are uniform over $\varepsilon \in \mathcal{E}$. Using the identity

$$\frac{\widehat{a}}{\widehat{b}} - \frac{a}{b} = \frac{1}{\widehat{b}}(\widehat{a} - a) - \frac{a}{\widehat{b}b}(\widehat{b} - b) \tag{65}$$

we bound the first term on the right-hand side, the variance term, by ($K_n^*(t) = 0$ for $|t| > T_n$):

$$\frac{1}{2\pi}\int_{-T_n}^{T_n} \left|\frac{1}{\frac{\hat{\phi}_{x^*}(t)}{\phi_{x^*}(t)}}\right| \left|\frac{\hat{\phi}_x(t) - \phi_x(t)}{\phi_{x^*}(t)}\right| |K_n^*(t)|\mathrm{d}t + \frac{1}{2\pi}\int_{-T_n}^{T_n}\left|\frac{\phi_x(t)}{\phi_{x^*}(t)}\right| \left|\frac{1}{\frac{\hat{\phi}_{x^*}(t)}{\phi_{x^*}(t)}}\right|$$
$$\left|\frac{\hat{\phi}_{x^*}(t) - \phi_{x^*}(t)}{\phi_{x^*}(t)}\right| |K_n^*(t)|\mathrm{d}t \tag{66}$$

Because $|\phi_{x^*}(t)| > 0$ and $\phi_{x^*}(t)$ is absolute integrable so that $\lim_{|t|\to\infty} |\phi_{x^*}(t)| = 0$, we have that $\inf_{|t|\le T_n} |\phi_{x^*}(t)| = |\phi_{x^*}(T_n)|$ if n is sufficiently large. Note also that because $\int_{-\infty}^{\infty} |K(z)|^2\,\mathrm{d}z > \infty$

$$\int_{-T_n}^{T_n} |K_n^*(t)|\mathrm{d}t = \int_{-T_n}^{T_n} \left|K^*\left(\frac{t}{T_n}\right)\right| \mathrm{d}t = T_n \int_{-1}^{1} |K^*(s)|\mathrm{d}s \leq T_n \int_{-1}^{1} \int_{-\infty}^{\infty} |K(z)|\mathrm{d}z \ \mathrm{d}s \leq CT_n \tag{67}$$

Using this and Lemma 3 we find that (66) is a.s. bounded by (the first term dominates the second because ϕ_ε is absolutely integrable and $|K_n^*(t)| \leq 1$)

$$O\left(\frac{T_n \alpha_n}{|\phi_{x^*}(T_n)|\left(1 - o\left(\frac{\alpha_n}{|\phi_{x^*}(T_n)|}\right)\right)}\right) = O\left(T_n^{k_{x^*}+2}\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma-\eta}\right) \tag{68}$$

where $T_n = O\left(\left(\frac{n}{\log n}\right)^\gamma\right)$ and the distribution of x^* is range-restricted of order k_{x^*}. Consider the second term in (64), i.e. the bias term. Because $K^*\left(\frac{t}{T_n}\right) = \int_{-\infty}^{\infty} e^{-itz} K(T_n z) \ \mathrm{d}z$ we have by the convolution theorem

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it\varepsilon}\phi_\varepsilon(t)[1 - K_n^*(t)] \ \mathrm{d}t = g_1(\varepsilon) - \int_{-\infty}^{\infty} g_1(\varepsilon - z)K(T_n z) \ \mathrm{d}z = g_1(\varepsilon) - \int_{-\infty}^{\infty} g_1\left(\varepsilon - \frac{z}{T_n}\right) K(z) \ \mathrm{d}z \tag{69}$$

Expanding $g_1\left(\varepsilon - \frac{z}{T_n}\right)$ in a qth order Taylor series we have, because K is a qth order kernel and the qth derivative of g_1 is bounded

$$\left|\frac{1}{2\pi}\int e^{-it\varepsilon}\phi_\varepsilon(t)[1 - K_n^*(t)] \ \mathrm{d}t\right| \leq CT_n^{-q}\int_{-\infty}^{\infty} |z|^q K(z) \ \mathrm{d}z \tag{70}$$

Therefore the bias term is $O(T_n^{-q})$. Hence we have the combined bound

$$\sup_{(x,x^*)\in\mathcal{X}\times\mathcal{X}^*} |\hat{g}_1(x - x^*) - g_1(x - x^*)| = O\left(T_n^{k_{x^*}+2}\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma-\eta}\right) + O(T_n^{-q}) \tag{71}$$

(ii) We have

$$\begin{aligned}
&\sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}} |\hat{g}_2(x^*, w) - g_2(x^*, w)| \\
&\leq \sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}} \left|\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-irx^*-isw}\left(\frac{\hat{\phi}_{xw}(r,s)\hat{\phi}_{x^*}(r)}{\hat{\phi}_x(r)} - \frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)}\right)K_n^*(r,s) \ \mathrm{d}s \ \mathrm{d}r\right| \\
&+ \sup_{(x^*,w)\in\S^*\times\mathcal{W}} \left|\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-irx^*-isw}\frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)}[1 - K_n^*(r,s)]\mathrm{d}s \ \mathrm{d}r\right|
\end{aligned} \tag{72}$$

Using the identity

$$\frac{\widehat{a}\widehat{c}}{\widehat{b}} - \frac{ac}{b} = \frac{\widehat{c}}{\widehat{b}}(\widehat{a} - a) + \frac{a}{\widehat{b}}(\widehat{c} - c) - \frac{ac}{\widehat{b}b}(\hat{b} - b) \tag{73}$$

the first term, i.e. the variance term, is bounded by

$$
\begin{aligned}
&\frac{1}{(2\pi)^2}\int_{-R_n}^{R_n}\int_{-S_n}^{S_n}\left|\frac{\hat{\phi}_{x^*}(r)}{\hat{\phi}_x(r)}\right||\hat{\phi}_{xw}(r,s)-\phi_{xw}(r,s)||K_n^*(r,s)|\mathrm{d}s\ \mathrm{d}r\\
&+\frac{1}{(2\pi)^2}\int_{-R_n}^{R_n}\int_{-S_n}^{S_n}\left|\frac{\phi_{xw}(r,s)}{\hat{\phi}_x(r)}\right||\hat{\phi}_{x^*}(r)-\phi_{x^*}(r)||K_n^*(r,s)|\mathrm{d}s\ \mathrm{d}r\\
&+\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\hat{\phi}_x(r)\phi_x(r)}\right||\hat{\phi}_x(r)-\phi_x(r)||K_n^*(r,s)|\mathrm{d}s\ \mathrm{d}r\\
&\le\frac{1}{(2\pi)^2}\int_{-R_n}^{R_n}\int_{-S_n}^{S_n}|\hat{\phi}_{x^*}(r)|\frac{1}{\left||\phi_\varepsilon(r)|-\dfrac{|\hat{\phi}_x(r)-\phi_x(r)|}{|\phi_{x^*}(r)|}\right|}\frac{|\hat{\phi}_{xw}(r,s)-\phi_{xw}(r,s)|}{|\phi_{x^*}(r)|}|K_n^*(r,s)|\mathrm{d}s\ \mathrm{d}r\\
&+\frac{1}{(2\pi)^2}\int_{-R_n}^{R_n}\int_{-S_n}^{S_n}|\phi_{xw}(r,s)|\frac{1}{\left||\phi_\varepsilon(r)|-\dfrac{|\hat{\phi}_x(r)-\phi_x(r)|}{|\phi_{x^*}(r)|}\right|}\frac{|\hat{\phi}_{x^*}(r)-\phi_{x^*}(r)|}{|\phi_{x^*}(r)|}|K_n^*(r,s)|\mathrm{d}s\ \mathrm{d}r\\
&+\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\phi_{x^*w}(r,s)|\frac{1}{\left||\phi_\varepsilon(r)|-\dfrac{|\hat{\phi}_x(r)-\phi_x(r)|}{|\phi_{x^*}(r)|}\right|}\frac{|\hat{\phi}_x(r)-\phi_x(r)|}{|\phi_{x^*}(r)|}|K_n^*(r,s)|\mathrm{d}s\ \mathrm{d}r \quad (74)
\end{aligned}
$$

Note that by a similar argument to that in (67)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|K_n^*(r,s)|\mathrm{d}r\ \mathrm{d}s\le CR_nS_n \quad (75)$$

Using the same method of proof as in part (i), the bound is (note that the final two terms are dominated by the first)

$$O\left(\frac{\alpha_nR_nS_n}{\phi_\varepsilon(R_n)\phi_{x^*}(R_n)}\right)$$

where $S_n=O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$, $R_n=O\left(\left(\frac{n}{\log n}\right)^{\gamma'}\right)$, and $\alpha_n=O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma'-\eta}\right)$. For the bias term we have by the convolution theorem

$$
\begin{aligned}
&\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-irx^*-isw}\phi_{x^*w}(r,s)[1-K_n^*(r,s)]\mathrm{d}s\ \mathrm{d}r\\
&=g_2(x^*,w)-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g_2\left(x^*-\frac{r}{R_n},w-\frac{s}{S_n}\right)K(r,s)\mathrm{d}r\ \mathrm{d}s \quad (76)
\end{aligned}
$$

Because $K(r,s)$ is a qth order kernel and all the qth order derivatives of $g_2(x^*,w)$ are bounded, we have by a qth order Taylor series expansion of g_2

$$\left|\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-irx^*-isw}\phi_{x^*w}(r,s)[1-K_n^*(r,s)]\ \mathrm{d}s\ \mathrm{d}r\right|\le CR_n^{-q_1}S_n^{-q_2} \quad (77)$$

with $q_1 + q_2 = q$. Combining the bounds on the variance and bias terms we find, if ε is range restricted of order k_ε and x^* is range-restricted of order k_{x^*},

$$\sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}} |\hat{g}_2(x^*, w) - g_2(x^*, w)| = O\left(R_n^{k_{x^*}+k_\varepsilon+3} S_n \left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma'-\eta}\right) + O(R_n^{-q_1} S_n^{-q_2}) \quad (78)$$

with $q_1 + q_2 = q$.[11] □

Proof of Theorem 2

First we linearize the moment function. Let h_0 be the joint density of x^*, x, w, i.e. $h_0(x^*, x, w) = g_1(x - x^*)g_2(x^*, w)$. We have $\hat{h}(x^*, x, w) = \hat{g}_1(x - x^*)\hat{g}_2(x^*, w)$. Both the population densities g_1, g_2 and their estimators are obtained by Fourier inversion. Because the corresponding characteristic functions are assumed to be absolutely integrable, g_1, g_2 are bounded on their support. Their estimators are bounded for finite n. Hence without loss of generality we can restrict g_1, g_2 and hence h to the set of densities that are bounded on their support.

The moment function is

$$m(y, x, w, \theta, h) = \frac{\int_{\mathcal{X}^*} \frac{\partial}{\partial\theta} f^*(y|x^*, w; \theta) h(x^*, x, w)\, \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y|x^*, w; \theta) h(x^*, x, w)\, \mathrm{d}x^*} \quad (79)$$

The joint density of y, x, w is denoted by $f(y, x, w; \theta)$. The population density of x^*, x, w is denoted by $h_0(x^*, x, w)$, $f_0(y, x, w, \theta) = \int_{\mathcal{X}^*} f^*(y|x^*, w; \theta) h_0(x^*, x, w)\, \mathrm{d}x^*$, and $f(y, x, w, \theta) = \int_{\mathcal{X}^*} f^*(y|x^*, w; \theta) h(x^*, x, w)\, \mathrm{d}x^*$.

Both the numerator and denominator in (79) are linear in h. Hence m is Fréchet differentiable in h and

$$\begin{aligned}
&\sup_{(y,x,w)\in\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} |m(y, x, w, \theta, h) - m(y, x, w, \theta, h_0) \\
&- \int_{\mathcal{X}^*} \left[\frac{f^*(y|x^*, w; \theta)}{f_0(y, x, w; \theta)} (s^*(y|x^*, w; \theta) - s_0(y|x, w; \theta))\right] (h(x^*, x, w) - h_0(x^*, x, w))\, \mathrm{d}x^*| \\
&= o(||h - h_0||)
\end{aligned} \quad (80)$$

with s^* and s_0 the scores of $f^*(y|x^*, w; \theta)$ and $f_0(y|x, w; \theta)$ respectively.

To prove consistency we need that for all $\theta \in \Theta$

$$|m(y, x, w, \theta, h_0)| \le b_1(y, x, w) \quad (81)$$

[11] Although the 'range-restricted' assumption rules out distributions that are 'supersmooth' like the normal distribution, the desirable nonparametric rate of convergence is still feasible in the case where both the error distribution and that of the latent true values are supersmooth. In that case, the bias terms converge to zero exponentially with respect to the smoothing parameters T_n, R_n, and S_n. This fact allows those smoothing parameters to diverge very slowly with respect to the sample size n, for example, $O(\log n)$, to still achieve the desirable convergence rate of both the variance terms and the bias terms. The rest of the proofs could be modified similarly to allow this case (see Schennach, 2004b, for further discussion).

DOI: 10.1002/jae

with $\mathrm{E}(b_1(y, w, x)) > \infty$, and that for all h in a (small) neighborhood of h_0 and all $\theta \in \Theta$, the Fréchet differential in h satisfies

$$\left| \int_{\mathcal{X}^*} \frac{f^*(y|x^*, w; \theta)}{f(y, x, w; \theta)} (s^*(y|x^*, w; \theta) - s(y|x, w; \theta))\, \mathrm{d}x^* \right| \leq b_2(y, w, x) \tag{82}$$

with $\mathrm{E}(b_2(y, w, x)) > \infty$.

The following weak restrictions on the parametric model are sufficient. There are constants $0 > m_0 > m_1 > \infty$ such that for all $(y, x^*, w) \in \mathcal{Y} \times \mathcal{X}^* \times \mathcal{W}$ and $\theta \in \Theta$

$$m_0 \leq f^*(y|x^*, w; \theta) \leq m_1 \tag{83}$$

$$\left| \frac{\partial f^*(y|x^*, w; \theta)}{\partial \theta} \right| \leq m_1 \tag{84}$$

This is sufficient for (81). For (82) we need in addition that for all $(y, w) \in \mathcal{Y} \times \mathcal{W}$ and $\theta \in \Theta$

$$\int_{\mathcal{X}^*} f^*(y|x^*, w; \theta)\, \mathrm{d}x^* > \infty \tag{85}$$

$$\left| \int_{\mathcal{X}^*} \frac{\partial f^*(y|x^*, w; \theta)}{\partial \theta}\, \mathrm{d}x^* \right| > \infty \tag{86}$$

If (82) holds then by Proposition 2 in Luenberger (1969, p. 176)

$$|m(y, x, w, \theta, h) - m(y, x, w, \theta, h_0)| \leq b_2(y, x, w) \sup_{(y, x^*, w) \in \mathcal{Y} \times \mathcal{X}^* \times \mathcal{W}} |h(x^*, x, w) - h_0(x^*, x, w)| \tag{87}$$

Hence Assumptions 5.4 and 5.5. in Newey (1994) are satisfied and we conclude that the semiparametric MLE is consistent if we use a (uniformly in x^*, x, w) consistent estimator for h.

Proof of Lemma 2

The derivation consists of a number of steps. We first linearize the score with respect to θ. Next we express the score at the population parameter as the sum of the population score and a correction term that accounts for the nonparametric estimates of the density functions. The correction term is further linearized in three steps. In the first step, the estimated score is linearized w.r.t. its numerator and denominator. In the second step, the leading terms in the first step are linearized w.r.t. the estimated densities $\hat{g}_1$ and $\hat{g}_2$. In the third step, the leading terms left in the previous step are linearized w.r.t. the empirical characteristic functions $\hat{\phi}_{x^*}(r)$, $\hat{\phi}_x(r)$, and $\hat{\phi}_{xw}(r, s)$. In each step, we show that the remainder terms are asymptotically negligible. The resulting expression is rewritten as the sum of five U-statistics. The asymptotic variances of these U-statistics are shown to be finite.

In the sequel, the moment functions are evaluated at $\theta = \theta_0$, and the dependence on θ_0 is suppressed in the notation, e.g. $f^*(y|x^*, w) = f^*(y|x^*, w; \theta_0)$ etc.

The semi-parametric MLE satisfies

$$\sum_{j=1}^{n} m(y_j, x_j, w_j, \hat{\theta}, \hat{g}_1, \hat{g}_2) = \sum_{j=1}^{n} \frac{\int_{\mathcal{X}^*} \frac{\partial}{\partial \theta} f^*(y_j|x^*, w_j, \hat{\theta}) \hat{g}_1(x_j - x^*) \hat{g}_2(x^*, w_j)\, \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y_j|x^*, w_j, \hat{\theta}) \hat{g}_1(x_j - x^*) \hat{g}_2(x^*, w_j)\, \mathrm{d}x^*} = 0 \tag{88}$$

DOI: 10.1002/jae

Linearization with respect to θ gives (the integration region is $\mathcal{X}^*$)

$$0 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\int \frac{\partial}{\partial \theta} f^*(y_j|x^*, w_j, \theta_0)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*}{\int f^*(y_j|x^*, w_j, \theta_0)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*}$$

$$+ \left(\frac{1}{n} \sum_{j=1}^{n} \frac{\int f^*(y_j|x^*, w_j, \overline{\theta})\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^* \int \frac{\partial^2}{\partial\theta\partial\theta'} f^*(y_j|x^*, w_j, \overline{\theta})\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^* - \int \frac{\partial}{\partial \theta} f^*(y_j|x^*, w_j, \overline{\theta})\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j \mathrm{d}x^* \int \frac{\partial}{\partial \theta'} f^*(y_j|x^*, w_j, \overline{\theta}))\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*}{\left(\int f^*(y_j|x^*, w_j, \overline{\theta})\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^* \right)^2} \right)$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \tag{89}$$

By Lemma 1 $\hat{g}_1$ and $\hat{g}_2$ converge uniformly, and this ensures that the matrix in the second term on the right-hand side converges to a matrix that is nonsingular, because the model is identified.

Hence we concentrate on the first term on the right-hand side:

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\int_{\mathcal{X}^*} \frac{\partial}{\partial \theta} f^*(y_j|x^*, w_j, \theta_0)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y_j|x^*, w_j, \theta_0)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*} \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} m(y_j, x_j, , w_j, \theta_0, \hat{g}_1, \hat{g}_2) \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} m(y_j, x_j, , w_j, \theta_0, g_1, g_2) + B
\end{aligned} \tag{90}$$

with

$$B = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\frac{\int_{\mathcal{X}^*} \frac{\partial}{\partial \theta} f^*(y_j|x^*, w_j)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y_j|x^*, w_j)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*} - \frac{\int_{\mathcal{X}^*} \frac{\partial}{\partial \theta} f^*(y_j|x^*, w_j)g_1(x_j - x^*)g_2(x^*, w_j)\ \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y_j|x^*, w_j)g_1(x_j - x^*)g_2(x^*, w_j)\ \mathrm{d}x^*} \right) \tag{91}$$

where B is the correction term that accounts for the estimated g_1 and g_2. This term is analyzed first. We use the following identities repeatedly:

$$\widehat{ab} = ab + b(\widehat{a} - a) + a(\widehat{b} - b) + (\widehat{a} - a)(\widehat{b} - b) \tag{92}$$

$$\frac{\widehat{a}}{\widehat{b}} = \frac{a}{b} + \frac{1}{b}(\widehat{a} - a) - \frac{a}{b^2}(\widehat{b} - b) + \frac{a}{\widehat{b}b^2}(\widehat{b} - b)^2 - \frac{1}{\widehat{b}b}(\widehat{a} - a)(\widehat{b} - b) \tag{93}$$

To simplify the notation we define

$$\frac{\frac{\partial}{\partial\theta}\widehat{f}(y_j, x_j, w_j)}{\widehat{f}(y_j, x_j, w_j)} \equiv \frac{\int_{\mathcal{X}^*} \frac{\partial}{\partial\theta} f^*(y_j|x^*, w_j)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y_j|x^*, w_j)\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j)\ \mathrm{d}x^*} \tag{94}$$

$$\frac{\frac{\partial}{\partial\theta} f(y_j, x_j, w_j)}{f(y_j, x_j, w_j)} \equiv \frac{\int_{\mathcal{X}^*} \frac{\partial}{\partial\theta} f^*(y_j|x^*, w_j)g_1(x_j - x^*)g_2(x^*, w_j)\ \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y_j|x^*, w_j)g_1(x_j - x^*)g_2(x^*, w_j)\ \mathrm{d}x^*} \tag{95}$$

First, using identity (93), we linearize the estimated score w.r.t the numerator $\frac{\partial}{\partial\theta}\widehat{f}(y_j, x_j, w_j)$ and the denominator $\widehat{f}(y_j, x_j, w_j)$. B then becomes (the integration region is $\mathcal{X}^*$)

$$\begin{aligned}
B = {} & \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{f(y_j, x_j, w_j)} \int \frac{\partial}{\partial\theta} f^*(y_j|x^*, w_j)[\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j) - g_1(x_j - x^*)g_2(x^*, w_j)]\mathrm{d}x^* \\
& - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\frac{\partial}{\partial\theta} f(y_j, x_j, w_j)}{f^2(y_j, x_j, w_j)} \int f^*(y_j|x^*, w_j)[\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j) - g_1(x_j - x^*)g_2(x^*, w_j)]\mathrm{d}x^* \\
& + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\frac{\partial}{\partial\theta} f(y_j, x_j, w_j)}{\widehat{f}(y_j, x_j, w_j) f^2(y_j, x_j, w_j)} \\
& \times \left(\int f^*(y_j|x^*, w_j)[\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j) - g_1(x_j - x^*)g_2(x^*, w_j)]\mathrm{d}x^* \right)^2 \\
& - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{\widehat{f}(y_j, x_j, w_j) f(y_j, x_j, w_j)} \\
& \times \left(\int \frac{\partial}{\partial\theta} f^*(y_j|x^*, w_j)[\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j) - g_1(x_j - x^*)g_2(x^*, w_j)]\mathrm{d}x^* \right) \\
& \times \left(\int_{\mathcal{X}^*} f^*(y_j|x^*, w_j)[\hat{g}_1(x_j - x^*)\hat{g}_2(x^*, w_j) - g_1(x_j - x^*)g_2(x^*, w_j)]\mathrm{d}x^* \right) \\
& \equiv D_1 - D_2 + D_3 - D_4
\end{aligned} \tag{96}$$

D_3 is bounded by

$$|D_3| \leq \sup_{(y,x,w)\in\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} \left| \frac{\frac{\partial}{\partial\theta} f(y,x,w)}{\hat{f}(y,x,w) f(y,x,w)^2} \right| \sup_{(y,w)\in\mathcal{Y}\times\mathcal{W}} \left| \int_{\mathcal{X}^*} f^*(y|x^*,w)\mathrm{d}x^* \right|^2$$

$$\times \sqrt{n} \sup_{(x,x^*,w)\in\mathcal{X}\times\mathcal{X}^*\times\mathcal{W}} |\hat{g}_1(x-x^*)\hat{g}_2(x^*,w) - g_1(x-x^*)g_2(x^*,w)|^2 \tag{97}$$

By assumption (A1) $f(y,x,w)$ is bounded from 0 on its support so that its uniform consistent estimator $\hat{f}(y,x,w)$ is also bounded from 0 on that support for sufficiently large n. By (A1) $|\frac{\partial}{\partial\theta} f(y,x,w)|$ is bounded on the support, and so is $|\int_{\mathcal{X}^*} f^*(y|x^*,w)\mathrm{d}x^*|$. By (92)

$$\begin{aligned}
&n^{\frac{1}{4}} \sup_{(x,x^*,w)\in\mathcal{X}\times\mathcal{X}^*\times\mathcal{W}} |\hat{g}_1(x-x^*)\hat{g}_2(x^*,w) - g_1(x-x^*)g_2(x^*,w)| \\
&= n^{\frac{1}{4}} \sup_{(x,x^*,w)\in\mathcal{X}\times\mathcal{X}^*\times\mathcal{W}} |g_2(x^*,w)(\hat{g}_1(x-x^*) - g_1(x-x^*))| \\
&+ n^{\frac{1}{4}} \sup_{(x,x^*,w)\in\mathcal{X}\times\mathcal{X}^*\times\mathcal{W}} |g_1(x-x^*)(\hat{g}_2(x^*,w) - g_2(x^*,w))| \\
&+ n^{\frac{1}{4}} \sup_{(x,x^*,w)\in\mathcal{X}\times\mathcal{X}^*\times\mathcal{W}} |(\hat{g}_1(x-x^*) - g_1(x-x^*))(\hat{g}_2(x^*,w) - g_2(x^*,w))|
\end{aligned} \tag{98}$$

If assumption (A5) holds, these expressions are $o_p(1)$. In the same way we show that $D_4 = o_p(1)$.

Next we consider

$$D_1 - D_2 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j,x_j,w_j,x^*)[\hat{g}_1(x_j-x^*)\hat{g}_2(x^*,w_j) - g_1(x_j-x^*)g_2(x^*,w_j)]\mathrm{d}x^* \tag{99}$$

where

$$\delta(y_j,x_j,w_j,x^*) = \frac{f^*(y_j|x^*,w_j)}{f(y_j,x_j,w_j)} \left(\frac{\frac{\partial}{\partial\theta} f^*(y_j|x^*,w_j)}{f^*(y_j|x^*,w_j)} - \frac{\frac{\partial}{\partial\theta} f(y_j,x_j,w_j)}{f(y_j,x_j,w_j)} \right) \tag{100}$$

Using identity (92) we obtain

$$\begin{aligned}
D_1 - D_2 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j,x_j,w_j,x^*) g_2(x^*,w_j)[\hat{g}_1(x_j-x^*) - g_1(x_j-x^*)]\mathrm{d}x^* \\
&+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j,x_j,w_j,x^*) g_1(x_j-x^*)[\hat{g}_2(x^*,w_j) - g_2(x^*,w_j)]\mathrm{d}x^* \\
&+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j,x_j,w_j,x^*)[\hat{g}_1(x_j-x^*) - g_1(x_j-x^*)][\hat{g}_2(x^*,w_j) - g_2(x^*,w_j)]\mathrm{d}x^* \\
&\equiv E_1 + E_2 + E_3
\end{aligned} \tag{101}$$

DOI: 10.1002/jae

Again, because $|\delta(y, x, w, x^*)|$ is bounded by (A1), we have using the same argument as above that $E_3 = o_p(1)$ by (A5). Next $E_1 + E_2$ is decomposed into the variance part and the bias part as follows:

$$\begin{aligned}
E_1 + E_2 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j, x_j, w_j, x^*) g_2(x^*, w_j)[\hat{g}_1(x_j - x^*) - \tilde{g}_1(x_j - x^*)]\mathrm{d}x^* \\
&+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j, x_j, w_j, x^*) g_1(x_j - x^*)[\hat{g}_2(x^*, w_j) - \tilde{g}_2(x^*, w_j)]\mathrm{d}x^* \\
&+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j, x_j, w_j, x^*) g_2(x^*, w_j)[\tilde{g}_1(x_j - x^*) - g_1(x_j - x^*)]\mathrm{d}x^* \\
&+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j, x_j, w_j, x^*) g_1(x_j - x^*)[\tilde{g}_2(x^*, w_j) - g_2(x^*, w_j)]\mathrm{d}x^* \\
&\equiv F_1 + F_2 + F_3 + F_4 \qquad (102)
\end{aligned}$$

where

$$\tilde{g}_1(x - x^*) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it(x-x^*)} \phi_\varepsilon(t) K_n^*(t)\mathrm{d}t \qquad (103)$$

$$\tilde{g}_2(x^*, w) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-irx^* - isw} \phi_{x^*w}(r, s) K_n^*(r, s)\mathrm{d}s\mathrm{d}r \qquad (104)$$

As shown in Lemma 1 we have

$$\sup_{\varepsilon} |\tilde{g}_1(\varepsilon) - g_1(\varepsilon)| = O_p(T_n^{-q}) \qquad (105)$$

$$\sup_{x^*, w} |\tilde{g}_2(x^*, w) - g_2(x^*, w)| = O_p(R_n^{-q_1} S_n^{-q_2}) \qquad (106)$$

with $q_1 + q_2 = q$. Therefore if (A5) holds then $|F_3| = o(1)$ and $|F_4| = o(1)$ because by (A1) δ is bounded. Hence we only need to consider F_1 and F_2 that we linearize w.r.t the ecf's $\hat{\phi}_{x^*}(r)$, $\hat{\phi}_x(r)$, and $\hat{\phi}_{xw}(r, s)$.

We have for F_1

$$\begin{aligned}
F_1 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j, x_j, w_j, x^*) g_2(x^*, w_j)[\hat{g}_1(x_j - x^*) - \tilde{g}_1(x_j - x^*)]\mathrm{d}x^* \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j, x_j, w_j, x^*) g_2(x^*, w_j) \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it(x_j - x^*)} \left(\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t)} - \frac{\phi_x(t)}{\phi_{x^*}(t)} \right) K_n^*(t)\mathrm{d}t \right] \mathrm{d}x^* \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa_1^*(t, y_j, x_j, w_j) \left(\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t)} - \frac{\phi_x(t)}{\phi_{x^*}(t)} \right) K_n^*(t)\ \mathrm{d}t \qquad (107)
\end{aligned}$$

where

$$\kappa_1^*(t, y_j, x_j, w_j) = \int_{\mathcal{X}^*} e^{-it(x_j - x^*)} \delta(y_j, x_j, w_j, x^*) g_2(x^*, w_j)\ \mathrm{d}x^* \qquad (108)$$

F_1 can be linearized further using identity (93):

$$F_1 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa_1^*(t, y_j, x_j, w_j) \frac{\hat{\phi}_x(t) - \phi_x(t)}{\phi_{x^*}(t)} K_n^*(t)\, \mathrm{d}t$$
$$- \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa_1^*(t, y_j, x_j, w_j) \frac{\phi_x(t)}{\phi_{x^*}(t)} \frac{\hat{\phi}_{x^*}(t) - \phi_{x^*}(t)}{\phi_{x^*}(t)} K_n^*(t)\, \mathrm{d}t$$
$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa_1^*(t, y_j, x_j, w_j) \frac{\phi_x(t)}{\hat{\phi}_{x^*}(t)} \left(\frac{\hat{\phi}_{x^*}(t) - \phi_{x^*}(t)}{\phi_{x^*}(t)} \right)^2 K_n^*(t)\, \mathrm{d}t$$
$$- \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa_1^*(t, y_j, x_j, w_j) \frac{1}{\hat{\phi}_{x^*}(t)\phi_{x^*}(t)} [\hat{\phi}_x(t) - \phi_x(t)][\hat{\phi}_{x^*}(t) - \phi_{x^*}(t)] K_n^*(t)\, \mathrm{d}t$$
$$= F_{11} - F_{12} + F_{13} - F_{14} \tag{109}$$

Consider $\kappa_1^*(t, y_j, x_j, w_j)$, which we denote by $\kappa_j^*(t)$. We also use the notation $\delta_j(x^*) = \delta(y_j, x_j, w_j, x^*)$ and $\kappa_j(x^*) = \delta_j(x^*) g_2(x^*, w_j)$. A superscript (k) indicates that we consider the kth derivative of a function. If we integrate by parts $k_{x^*} + 1$ times we obtain, if $\mathcal{X}^* = [L, U]$,

$$\kappa_j^*(t) = \sum_{l=1}^{k_{x^*}+1} \frac{(-1)^{l-1}}{(it)^l} e^{-it(x_j - x^*)} \kappa_j^{(l-1)}(x^*)|_L^U + \frac{(-1)^{k_{x^*}+1}}{(it)^{k_{x^*}+1}} \int_L^U e^{-it(x_j - x^*)} \kappa_j^{(k_{x^*}+1)}(x^*)\, \mathrm{d}x^* \tag{110}$$

with

$$\kappa_j^{(k)}(x^*) = \sum_{l=0}^{l} \binom{k}{l} \delta_j^{(l)}(x^*) g_2^{(k-l)}(x^*, w_j)\, \mathrm{d}x^* \tag{111}$$

If x^* is range-restricted of order k_{x^*}, then $g_2^{(k)}(L) = g_2^{(k)}(U) = 0$ for $k = 0, \ldots, k_{x^*} - 1$, so that if, as assumed, this implies that $g_2^{(k)}(L, w) = g_2^{(k)}(U, w) = 0$ for $k = 0, \ldots, k_{x^*} - 1$ and all $w \in \mathcal{W}$, we have that

$$|\kappa_j^*(t)| \le C|t|^{k_{x^*}+1} \tag{112}$$

We conclude that $\left| \frac{\kappa_1^*(t, y_j, x_j, w_j)}{\phi_{x^*}(t)} \right|$ is bounded.

Hence by Lemma 3 and (67)

$$|F_{13}| = O_p\left(\sqrt{n} \frac{T_n}{|\phi_{x^*}(T_n)|} \alpha_n^2 \right) = O_p\left(\sqrt{n} \left(\frac{\log n}{n} \right)^{1-(k_{x^*}+4)\gamma-\eta} \right) \tag{113}$$

and

$$|F_{14}| = O_p\left(\sqrt{n} \frac{T_n}{|\phi_{x^*}(T_n)|} \alpha_n^2 \right) = O_p\left(\sqrt{n} \left(\frac{\log n}{n} \right)^{1-(k_{x^*}+4)\gamma-\eta} \right) \tag{114}$$

so that these terms are $o_p(1)$ if (A5) holds.

F_{11} can be written as

$$F_{11} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa_1^*(t, y_j, x_j, w_j) \frac{\hat{\phi}_x(t) - \phi_x(t)}{\phi_{x^*}(t)} K_n^*(t)\, \mathrm{d}t$$

$$= \frac{1}{n\sqrt{n}} \sum_{k=1}^{n} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\kappa_1^*(t, y_k, x_k, w_k)}{\phi_{x^*}(t)} [e^{itx_j} - \phi_x(t)] K_n^*(t) \, \mathrm{d}t \tag{115}$$

If we omit the terms $k = j$ in the summations (these terms are $o_p(1)$) the resulting expression is a one-sample U-statistic. Using the same line of proof as in Hu and Ridder (2010), we can show that this term is $O_p(1)$. Hence by projection

$$F_{11} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{d_1^*(t)}{\phi_{x^*}(t)} [e^{itx_j} - \phi_x(t)] K_n^*(t) \mathrm{d}t + o_p(1) \tag{116}$$

with

$$d_1^*(t) = E[\kappa_1^*(t, y, x, w))] \tag{117}$$

Similarly, we have

$$F_{12} = \sqrt{\frac{n}{n_1}} \frac{1}{\sqrt{n_1}} \sum_{j=1}^{n_1} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{d_1^*(t)}{\phi_{x^*}(t)} \frac{\phi_x(t)}{\phi_{x^*}(t)} [e^{itx_j^*} - \phi_{x^*}(t)] K_n^*(t) \, \mathrm{d}t + o_p(1) \tag{118}$$

The next step is to analyze F_2:

$$\begin{aligned} F_2 &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_{\mathcal{X}^*} \delta(y_j, x_j, w_j, x^*) g_1(x_j - x^*) [\hat{g}_2(x^*, w_j) - \widetilde{g}_2(x^*, w_j)] \mathrm{d}x^* \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \kappa_2^*(r, s, y_j, x_j, w_j) \left(\frac{\hat{\phi}_{xw}(r, s) \hat{\phi}_{x^*}(r)}{\hat{\phi}_x(r)} - \frac{\phi_{xw}(r, s) \phi_{x^*}(r)}{\phi_x(r)} \right) \\ &\quad \cdot K_n^*(r, s) \, \mathrm{d}s \, \mathrm{d}r \end{aligned} \tag{119}$$

where

$$\kappa_2^*(r, s, y_j, x_j, w_j) = \int_{\mathcal{X}^*} e^{-irx^* - isw_j} \delta(y_j, x_j, w_j, x^*) g_1(x_j - x^*) \, \mathrm{d}x^* \tag{120}$$

To linearize $\dfrac{\hat{\phi}_{x^*}(r) \hat{\phi}_{xw}(r, s)}{\hat{\phi}_x(r)}$, we use the identity

$$\begin{aligned} \frac{\widehat{a}\widehat{c}}{\widehat{b}} &= \frac{ac}{b} + \frac{1}{b}(\widehat{a}\widehat{c} - ac) - \frac{ac}{b^2}(\widehat{b} - b) + \frac{ac}{\widehat{b}b^2}(\widehat{b} - b)^2 - \frac{1}{\widehat{b}b}(\widehat{a}\widehat{c} - ac)(\hat{b} - b) \\ &= \frac{ac}{b} + \frac{c}{b}(\widehat{a} - a) + \frac{a}{b}(\widehat{c} - c) - \frac{ac}{b^2}(\widehat{b} - b) + \frac{ac}{\widehat{b}b^2}(\widehat{b} - b)^2 + \frac{1}{b}(\widehat{a} - a)(\widehat{c} - c) \\ &\quad - \frac{c}{\widehat{b}b}(\widehat{a} - a)(\widehat{b} - b) - \frac{a}{\widehat{b}b}(\widehat{c} - c)(\widehat{b} - b) - \frac{1}{\widehat{b}b}(\widehat{a} - a)(\widehat{c} - c)(\widehat{b} - b) \end{aligned} \tag{121}$$

Therefore, we have

$$F_2 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \kappa_2^*(r, s, y_j, x_j, w_j) \frac{\phi_{xw}(r, s)}{\phi_x(r)} [\hat{\phi}_{x^*}(r) - \phi_{x^*}(r)] K_n^*(r, s) \, \mathrm{d}s \, \mathrm{d}r$$

$$
\begin{aligned}
&+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\kappa_2^*(r,s,y_j,x_j,w_j)\frac{\phi_{x^*}(r)}{\phi_x(r)}[\hat{\phi}_{xw}(r,s)-\phi_{xw}(r,s)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r\\
&-\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\kappa_2^*(r,s,y_j,x_j,w_j)\frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)\phi_x(r)}[\hat{\phi}_x(r)-\phi_x(r)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r\\
&+F_{24}\\
&\equiv F_{21}+F_{22}-F_{23}+F_{24}
\end{aligned}
\tag{122}
$$

where $F_{24}=G_1+G_2+G_3+G_4+G_5$ contains all the other (quadratic) terms in the linearization.

Integration by parts as for $\kappa_1^*(t,y_j,x_j,w_j)$ shows that

$$
|\kappa_2^*(r,s,y_j,x_j,w_j)|\le C|r|^{k_\varepsilon+1}
\tag{123}
$$

if ε is range-restricted of order k_ε and $\delta(y,x,w,x^*)$, $g_1(\varepsilon)$ have absolutely integrable derivatives of order $k_\varepsilon+1$ with respect to x^* and ε, respectively. This implies that $\left|\frac{\kappa_2^*(r,s,y,x,w)}{\phi_\varepsilon(r)}\right|$ is bounded in r, s.

Hence

$$
\begin{aligned}
G_1=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}&\frac{\kappa_2^*(r,s,y_j,x_j,w_j)}{\phi_\varepsilon(r)}\frac{\phi_{x^*w}(r,s)}{\left(\dfrac{\hat{\phi}_x(r)-\phi_x(r)}{\phi_{x^*}(r)}+\phi_\varepsilon(r)\right)\phi_{x^*}(r)^2}\\
&\cdot(\hat{\phi}_x(r)-\phi_x(r))^2K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r=O_p\left(\sqrt{n}\frac{R_nS_n}{\phi_{x^*}(R_n)^2|\phi_\varepsilon(R_n)|}\alpha_n^2\right)
\end{aligned}
\tag{124}
$$

Next

$$
\begin{aligned}
G_2&=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\kappa_2^*(r,s,y_j,x_j,w_j)}{\phi_\varepsilon(r)}\frac{\hat{\phi}_{x^*}(r)-\phi_{x^*}(r)}{\phi_{x^*}(r)}\\
&\quad\cdot(\hat{\phi}_{xw}(r,s)-\phi_{xw}(r,s))K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r\\
&=O_p\left(\sqrt{n}\frac{R_nS_n}{|\phi_{x^*}(R_n)|}\alpha_n^2\right)
\end{aligned}
\tag{125}
$$

Further

$$
\begin{aligned}
G_3&=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\kappa_2^*(r,s,y_j,x_j,w_j)}{\phi_\varepsilon(r)}\frac{\phi_{x^*w}(r,s)}{\left(\dfrac{\hat{\phi}_x(r)}{\phi_{x^*}(r)}-\phi_\varepsilon(r)\right)+\phi_\varepsilon(r)}\\
&\quad\cdot\frac{(\hat{\phi}_x(r)-\phi_x(r))(\hat{\phi}_{x^*}(r)-\phi_{x^*}(r))}{\phi_{x^*}(r)^2}K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r\\
&=O_p\left(\sqrt{n}\frac{R_nS_n}{\phi_{x^*}(R_n)^2}\alpha_n^2\right)
\end{aligned}
\tag{126}
$$

DOI: 10.1002/jae

Also

$$G_4 = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\kappa_2^*(r,s,y_j,x_j,w_j)}{\phi_\varepsilon(r)}\frac{\phi_{x^*}(r)}{\left(\dfrac{\hat\phi_x(r)}{\phi_\varepsilon(r)}-\phi_{x^*}(r)\right)+\phi_{x^*}(r)}$$

$$\cdot\frac{(\hat\phi_x(r)-\phi_x(r))(\hat\phi_{xw}(r,s)-\phi_{xw}(r,s))}{\phi_{x^*}(r)\phi_\varepsilon(r)}K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r$$

$$= O_p\left(\sqrt{n}\frac{R_nS_n}{|\phi_{x^*}(R_n)\phi_\varepsilon(R_n)|}\alpha_n^2\right) \tag{127}$$

The final term is dominated by the others, so that the final bound is

$$O_p\left(\sqrt{n}\frac{R_nS_n}{|\phi_\varepsilon(R_n)|\phi_{x^*}(R_n)^2}\alpha_n^2\right)$$

which is $o_p(1)$ if (A5) holds.

Using the same line of proof as in Hu and Ridder (2010), we can show that F_{21}, F_{22}, and F_{23} are $O_p(1)$. Moreover, they can be written as U-statistics, so that

$$F_{21} = \sqrt{\frac{n}{n_1}}\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d_2^*(r,s)\frac{\phi_{xw}(r,s)}{\phi_x(r)}$$

$$[e^{itx_j^*}-\phi_{x^*}(t)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r+o_p(1) \tag{128}$$

$$F_{22} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d_2^*(r,s)\frac{\phi_{x^*}(r)}{\phi_x(r)}$$

$$[e^{irx_j+isw_j}-\phi_{xw}(r,s)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r+o_p(1) \tag{129}$$

$$F_{23} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d_2^*(r,s)\frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)\phi_x(r)}$$

$$[e^{irx_j}-\phi_x(r)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r+o_p(1) \tag{130}$$

where

$$d_2^*(r,s) = E[\kappa_2^*(r,s,y_j,x_j,w_j)] \tag{131}$$

Substitution in (90) gives the asymptotically linear expression for the score

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}m(y_j,x_j,w_j,\theta_0,\hat g_1,\hat g_2)$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}m(y_j,x_j,w_j,\theta_0,g_1,g_2)+F_{11}-F_{12}+F_{21}+F_{22}-F_{23}+o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}m(y_j,x_j,w_j,\theta_0,g_1,g_2)+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{d_1^*(t)}{\phi_{x^*}(t)}[e^{itx_j}-\phi_x(t)]K_n^*(t)\,\mathrm{d}t$$

DOI: 10.1002/jae

$$
\begin{aligned}
&-\frac{\sqrt{n}}{\sqrt{n_1}}\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{d_1^*(t)}{\phi_{x^*}(t)}\frac{\phi_x(t)}{\phi_{x^*}(t)}[e^{itx_j^*}-\phi_{x^*}(t)]K_n^*(t)\,\mathrm{d}t\\
&+\frac{\sqrt{n}}{\sqrt{n_1}}\frac{1}{\sqrt{n_1}}\sum_{j=1}^{n_1}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d_2^*(r,s)\frac{\phi_{xw}(r,s)}{\phi_x(r)}[e^{itx_j^*}-\phi_{x^*}(t)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r\\
&+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d_2^*(r,s)\frac{\phi_{x^*}(r)}{\phi_x(r)}[e^{irx_j+isw_j}-\phi_{xw}(r,s)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r\\
&-\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}d_2^*(r,s)\frac{\phi_{xw}(r,s)\phi_{x^*}(r)}{\phi_x(r)\phi_x(r)}[e^{irx_j}-\phi_x(r)]K_n^*(r,s)\,\mathrm{d}s\,\mathrm{d}r+o_p(1)
\end{aligned}
$$

By the triangular array central limit theorem this expression converges in distribution to a normal random variable with mean 0 and variance matrix Ω given in Lemma 2.