Estimation of Nonlinear Models with Measurement Error Using Marginal Information[1]

By Yingyao Hu[2] and Geert Ridder[3]

Abstract

We consider 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 and the other explanatory variables. The estimator is shown to be $\sqrt{n}$ consistent and normally distributed. The simulation results are in line with the asymptotic results. The semi-parametric MLE is applied to a duration model for AFDC welfare spells with misreported welfare benefits. The marginal distribution of welfare benefits is obtained from an administrative source.

JEL classification: C14, C41, I38.

Keywords: measurement error model, marginal information, deconvolution, Fourier transform, duration model, welfare spells.

[1]This paper is based on the first author's Ph.D. thesis. We thank Robert Moffitt, Cheng Hsiao, Susanne Schennach, Matt Shum, Quang Vuong, and seminar participants at Georgetown U, Johns Hopkins U, Ohio State U, Queen's U, UBC, UCLA, UCSD, U Texas, U Toronto, and U Western Ontario. All errors are our responsibility.

[2]Department of Economics, University of Texas at Austin, 1 University Station C3100, BRB 1.116 , Austin, TX78712, hu@eco.utexas.edu, http://www.eco.utexas.edu/~hu/

[3]Department of Economics, University of Southern California, Los Angeles, CA90089, ridder@usc.edu, http://www-rcf.usc.edu/~ridder/.

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 data in which one or more independent variables are measured with error. Measurement error is a pervasive problem in economic data (Bound, Brown, Duncan, and Mathiowetz, 2001)). The identification and estimation of models that are nonlinear in mismeasured variables is a notoriously difficult problem (see (Carroll, Ruppert, and Stefanski, 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 additional sample information, such as a validation sample or replicate measurements. 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. Validation studies show that assumptions that are routinely made in statistical measurement error models are often violated (see among others (Rodgers, Brown, and Duncan, 1993)).

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 (Hsiao, 1989, 1991, Wang, 1998, Hsiao and Wang, 2000). With this assumption the estimation problem is complicated, but fully parametric. In general, the distribution of the measurement errors is non-parametrically unidentified, so that this approach relies on identification by distributional assumptions. [4]

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. The IV method was developed for models that are linear in the mismeasured variables. In gen-

[4] The shape of the distribution of the measurement error plays an important role in measurement error models. In linear models the regression coefficients are not identified if the distribution of the measurement error is normal, but they are if that distribution is not normal. Lewbel (1997) discusses estimation in linear models using higher order moments, if the distribution of the measurement error is non-normal. His results only apply to linear models. Lewbel's method, and others cited later, point at a curious interaction between model and distributional assumptions.

eral, IV estimators are biased in nonlinear models. However, Amemiya and Fuller (1988) and Carroll and Stefanski (1990) obtain a consistent IV estimator in nonlinear models under the assumption that the measurement error vanishes if the sample size increases. Hausman, Ichimura, Newey, and Powell (1991) and Hausman, Newey, and Powell (1995) extend IV estimation to a polynomial regression model. Newey (2001) considers the nonlinear regression model, but he notes that there are no general results on the non-parametric identification of nonlinear models with mismeasured regressors by instrumental variables.

The third approach is to use additional sample information. The additional information can come in the form of replicate measurements or in the form of a validation sample. The sample contains replicate measurements if there are at least two mismeasured variables that correspond to the same latent true value. Li and Vuong (1998) show that if the measurement errors in the two measurements are stochastically independent (although zero correlation suffices), the distribution of the latent true value is non-parametrically identified. Schennach (2000) uses the same approach to obtain a general extremum estimator in models that are nonlinear in the mismeasured variables. Hausman, Newey and Powell (1995) discuss the use of replicate measurements in polynomial regression models. In practice replicate measurements with independent (or uncorrelated) measurement errors are rare.[5] A validation sample is a subsample of the original sample for which accurate measurements are available. Bound et al. (1989) discuss the use of validation data in linear models. Hsiao (1989) and Hausman, Ichimura, Newey, and Powell (1991) discuss the extension to nonlinear models. Pepe and Fleming (1991) and Carroll and Wand (1991) propose to estimate the joint density of the latent true value, the mismeasured value, and the other variables non-parametrically, and to use this estimated density to correct for the measurement error bias in nonlinear models (see also Lee and Sepanski, 1995). The approach taken in this paper is along these lines. Chen, Hong and Tamer (2003) note that with validation data the classical assumption that the measurement error is independent of the latent true value and of the other variables in the model can be relaxed.

[5] If the administrative data that we use below are measured with error, we can consider the sample and administrative reports as replicate measurements. In this case the errors are likely to be generated by different mechanisms, so that the independence assumption is more reasonable. The extension of our approach to this situation is left to future work.

A validation sample is the gold standard for estimation if the independent variables have measurement errors. In this paper we show that much 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. Actually, most validation samples are constructed by matching survey data to administrative data. Creating such matched samples is very costly, in particular in surveys with a national coverage. Moreover, it requires the cooperation of the owners of the administrative data who may be reluctant to give permission. Not all surveys collect unique identifiers, as the Social Security Number, that can be used to match the survey information to that in administrative records. Finally, the matching raises privacy issues that may be hard to resolve. 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.[6]

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, Brown, Duncan, and Rodgers, 1994), union coverage (Barron, Berger, and Black, 1997), and unemployment spells (Mathiowetz and Duncan, 1988). Tax returns have been used in studies of wage and income (Code, 1992), unemployment benefits (Dibbs, Hale, Loverock, and Michaud, 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, Rouse and Staiger, 1999). Card et al. (2001) examine Medicaid coverage using Medicaid data. Bound et al. (2001) give a survey of studies that use administrative data.

[6] In the 70's several attempts were made to combine survey and administrative data to create a matched sample using a method called statistical matching. One of the reasons for creating such matched samples was that supposedly inaccurate survey information, was combined with more accurate data from administrative sources.

A problem with administrative records is that they usually contain only a small number of variables. We show that under reasonable assumptions that is sufficient to correct for measurement error in parametric models.

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 data set 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 value from the model. The resulting mixture model (with estimated mixing distribution) can then be estimated by ML. The resulting semi-parametric MLE is $\sqrt{n}$ consistent. 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. Although in this paper we maintain the classical measurement error assumptions the same method can be used for the case that the measurement error is correlated with the true value and the other covariates. Validation studies have shown that is is often the case.

In this paper we only consider continuous mismeasured variables. The discrete case will be considered in a separate paper (see Ridder and Moffitt, 2003) for a discussion). In the continuous case the non-parametric estimator of the conditional density of the latent true value given the reported value and the other independent variables is obtained by two deconvolutions. The paper contributes to deconvolution theory in two respects. We show that two assumptions that are commonly made in the literature on nonparametric estimation by deconvolution, i.e. the assumption that the support of the random variables is bounded and the assumption that their characteristic functions are never 0, need not hold, and are indeed incompatible for symmetric distributions. It turns out that the assumption that the characteristic functions is never 0 is not necessary for deconvolution, and we develop the theory for the case that the set of (real) zeros of the characteristic function is a countable, non-dense set. The reason that there is a preference for distributions with a bounded support is that the derivation of the rate of convergence of the empirical characteristic function is rather simple in that case. As far as we know there did not exist a results for distributions with an unbounded support, and we derive such a rate. This corrects a result in Horowitz and Markatou (1996).

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 conclusions. The proofs are in the appendix.

2 Identification using marginal information

2.1 Linear regression with errors-in-variables

Consider the linear regression model

$$y = \beta_0 + \beta_1 x^* + \beta_2 w + u \tag{1}$$

with $\mathrm{E}(u|x^*, w) = 0$. We do not observed x^*, but x with

$$x = x^* + \varepsilon \tag{2}$$

The usual assumption is that $\varepsilon \perp x^*, w, u$. Hence the measurement error ε is independent of the latent true value, the other independent variables, and the random error of the linear regression. Measurement error that satisfies these assumptions is called classical measurement error.

If u is uncorrelated with the independent variables, the regression coefficients can be expressed as

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(x^*) & \mathrm{Cov}(x^*, w) \\ \mathrm{Cov}(x^*, w) & \mathrm{Var}(w) \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{Cov}(x^*, y) \\ \mathrm{Cov}(w, y) \end{pmatrix} \tag{3}$$

and

$$\beta_0 = \mathrm{E}(y) - \beta_1 \mathrm{E}(x^*) - \beta_2 \mathrm{E}(w) \tag{4}$$

If only (a random sample from the joint distribution of) y, x, w is observed, the regression coefficients can not be identified without further information. We have

$$\mathrm{Cov}(x, y) = \mathrm{Cov}(x^*, y) + \beta_0 \mathrm{E}(\varepsilon - \mathrm{E}(\varepsilon)) + \beta_1 \mathrm{Cov}(\varepsilon, x^*) + \beta_2 \mathrm{Cov}(\varepsilon, w) + \mathrm{Cov}(\varepsilon, u) \tag{5}$$

and under the classical measurement error assumptions the right-hand side is equal to the covariance of x^* and y[7]. If the measurement error is uncorrelated with w, then $\mathrm{Cov}(x, w) = \mathrm{Cov}(x^* w)$. Because the mean and variance of x^* cannot be identified from the distribution of x without further assumptions, the regression coefficients are not identified. For instance, if the expected value of the measurement error is 0, $\mathrm{E}(x) = \mathrm{E}(x^*)$. But even with this assumption, the classical measurement error assumptions are not sufficient to identify the variance of x^*, and hence the regression coefficients, although the classical errors-in-variables assumptions imply bounds on the regression coefficient (Gini (1921)). [8]

The regression parameters are identified if the marginal mean and variance of the latent true value x^* can be obtained from a secondary data set. This result extends to the polynomial regression model considered by Hausman, Ichimura, Newey, and Powell (1991). In that case higher order moments of x^* are needed (see Hu (2002)). It is natural to ask whether knowledge of the marginal distribution of the latent true variable is sufficient for

[7]Of course we only need uncorrelatedness for this result.

[8]If the measurement error is not normally distributed, the regression coefficients can be identified using higher order moments (Bekker, 1986; Lewbel, 1997).

the identification of a general nonlinear model with measurement error. The next section shows that this is indeed the case.

Before we discuss identification under the classical measurement error assumptions we show that marginal information is also useful in the case of non-classical measurement error. Consider the measurement error model

$$x = \gamma_1 x^* + \gamma_2 w + \varepsilon \tag{6}$$

with $\mathrm{E}(\varepsilon|x^*, w) = 0$. Then we have the following system of equations

$$\begin{aligned} \mathrm{Cov}(x, w) &= \gamma_1 \mathrm{Cov}(x^*, w) + \gamma_2 \mathrm{Var}(w) \\ \mathrm{Cov}(x, w^2) &= \gamma_1 \mathrm{Cov}(x^*, w^2) + \gamma_2 \mathrm{E}\left((w - \mathrm{E}(w))^3\right) \\ \mathrm{Cov}(x, y) &= \gamma_1 \mathrm{Cov}(x^*, y) + \gamma_2 \mathrm{Cov}(w, y) \end{aligned} \tag{7}$$

If we have marginal information on x^*, w we can solve this system for $\gamma_1, \gamma_2, \mathrm{Cov}(x^*, y)$ and this suffices to identify the regression coefficients.

2.2 Models nonlinear in mismeasured covariates

A parametric model for the relation between a dependent variable y, a latent true variable x^* and other independent variables 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

$$x = x^* + \varepsilon \tag{8}$$

with $\varepsilon \perp x^*, w, y$. 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 random error of the regression. If the nonlinear model is derived from a latent regression model, as in probit and tobit, the assumption implies that the random error of the latent regression and the measurement error are independent.

In this paper we consider that case that x^* (and hence x) is a continuous variable.[9] 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.

Efficient inference for the parameters θ is based on the likelihood function. The individual contribution to the likelihood function 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

$$f(y|x, w; \theta) = \int f^*(y|x^*, w; \theta) g(x^*|x, w) \mathrm{d}x^* \tag{9}$$

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 (9) in likelihood inference is that it requires knowledge of the density $g(x^*|x, w)$. This density can be expressed as

$$g(x^*|x, w) = \frac{g(x|x^*, w) g_2(x^*, w)}{g_3(x, w)} \tag{10}$$

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 that density are identified, and if so, whether the identification is by functional form. Mispecification 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

$$g(x|x^*, w) = g_1(x - x^*) \tag{11}$$

[9] If x^* is discrete the distribution of x^* given x, w is still identified, but the estimation procedure is different (and fully parametric).

with g_1 the density of ε. Let $\phi_x(t) = \mathrm{E}(\exp(itx))$ be the characteristic function of the random variable x. From (8) 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

$$\phi_\varepsilon(t) = \frac{\phi_x(t)}{\phi_{x^*}(t)} \tag{12}$$

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 $g_3(x, w)$ as

$$g_3(x, w) = \int g(x, x^*, w)\mathrm{d}x^* = \int g_1(x - x^*)g_2(x^*, w)\mathrm{d}x^* \tag{13}$$

If $\phi_{xw}(s, t) = \mathrm{E}(\exp(isx + itw))$ is the characteristic function of the joint distribution of x, w, then the integral equation (13) is equivalent to $\phi_{xw}(s, t) = \phi_\varepsilon(s)\phi_{x^*w}(s, t)$, so that

$$\phi_{x^*,w}(s, t) = \frac{\phi_{x,w}(s, t)}{\phi_\varepsilon(s)} = \frac{\phi_{x,w}(s, t)\phi_{x^*}(s)}{\phi_x(s)} \tag{14}$$

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 (14) contains only characteristic functions of distributions that can be observed in either sample.

The conditional density of y given x, w in (9) 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,[10] 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

[10] The measure is the product measure of the counting measure for the discrete variables in y, w and the Lebesgue measure for the continuous variables in y, x^*, w.

a unique and well-separated maximum in θ_0 (Van der Vaart (1998), Lemma 5.35).

Identification of θ in $f^*(y|x^*, w; \theta)$ implies identification of θ in $f(y|x, w; \theta)$. To see this assume that θ is observationally equivalent to θ_0. Then for all y, w, x

$$\begin{aligned} & f(y|x, w; \theta) - f(y|x, w; \theta_0) \\ = & \int (f^*(y|x^*, w; \theta) - f^*(y|x^*, w; \theta_0))g(x^*|x, w)\mathrm{d}x^* \equiv 0 \end{aligned} \tag{15}$$

After substitution of (10) and (11) and a change of variable in the integration, this is equivalent to

$$\int (f^*(y|x-\varepsilon, w; \theta) - f^*(y|x-\varepsilon, w; \theta_0))g_2(x-\varepsilon, w)g_1(\varepsilon)\mathrm{d}\varepsilon \equiv 0 \tag{16}$$

Without loss of generality we assume that the support x and x^* is $\Re$.[11] Now for fixed y, w, (16) is of the form $\mathrm{E}(h(x-\varepsilon)) \equiv 0$ for all $x \in \Re$, and this implies that $h \equiv 0$, so that for all y, x^*, w, $f^*(y|x^*, w; \theta) \equiv f^*(y|x^*, w; \theta_0)$ and this cannot hold if θ is identified in the original model. Because if θ and θ_0 are observationally equivalent in the original model, they are also observationally equivalent in the distribution of y given x, w, we have that θ is identified in the conditional density of y given x^*, w if and only if θ is identified in the conditional density of y given x, w.

The fact that the density of x^* given x, w is non-parametrically identified makes it possible to study e.g. 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 that 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.

[11] If x^* is bounded and ε is independent of x^*, then the support of x is larger than that of x^*. The argument remains valid.

3 Estimation of errors-in-variables models with marginal information

3.1 Non-parametric Fourier inversion estimators

The first step in the estimation is to obtain a non-parametric estimator of $g(x^*|x,w) = g_1(x-x^*)g_2(x^*,w)$. The density g_1 of the measurement error ε has characteristic function (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

$$g_1(x-x^*) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it(x-x^*)}\frac{\phi_x(t)}{\phi_{x^*}(t)}\mathrm{d}t \tag{17}$$

The joint characteristic function of x^*, w is $\phi_{x^*w}(s,t) = \frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s)}$. Again Fourier inversion gives the joint density of x^*, w as

$$g_2(x^*,w) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-isx^*-ivw}\frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s)}\mathrm{d}t\mathrm{d}s \tag{18}$$

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

$$\hat{\phi}_x(t) = \frac{1}{n}\sum_{i=1}^{n} e^{itx_i} \tag{19}$$

However, the estimators that we obtain if we substitute the ecf of x and x^* in (17) and the ecf of x, w, x^* and x in (18) are not well-defined. In particular, sampling variation makes that the integrals do not 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.[12] For these reasons we introduce

[12] Horowitz and Markatou (1994),Lemma 1, p. 164, invoke a result on the uniform rate of convergence of the ecf that is not correct (Fuerverger and Mureika (1977), p. 89). Lemma 1 below gives the correct rate.

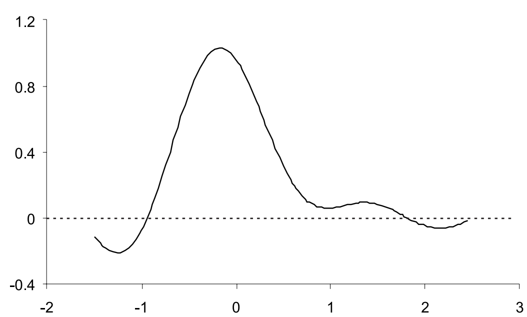
integration limits in the definition of the non-parametric density estimators[13]

$$\hat{g}_1(x - x^*) = \frac{1}{2\pi}\int_{-T_n}^{T_n} e^{-it(x-x^*)}\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t)}\mathrm{d}t \tag{20}$$

$$\hat{g}_2(x^*, w) = \frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw}\frac{\hat{\phi}_{xw}(s,t)\hat{\phi}_{x^*}(s)}{\hat{\phi}_x(s)}\mathrm{d}t\mathrm{d}s \tag{21}$$

S_n, T_n diverge at an appropriate rate to be defined below.[14] Although we integrate a complex-valued function the integrals are real. However, because we truncate the range of integration, the estimated densities need not be positive. Figure 1 illustrates this for our application.

Figure 1: Estimate of density of the measurement error with smoothing parameter $T = .7$



The non-parametric estimators in (20) and (21) cannot be used for all types of distributions. 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 $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\phi_{x^*w}(s,t)|\mathrm{d}t\mathrm{d}s < \infty$. A sufficient condition is that e.g. $\int_{-\infty}^{\infty}|g_1(\varepsilon)''|\mathrm{d}\varepsilon < \infty$ with g_1'' the second derivative of the pdf of ε , which a smoothness condition (and an analogous condition on the joint density of x, w).

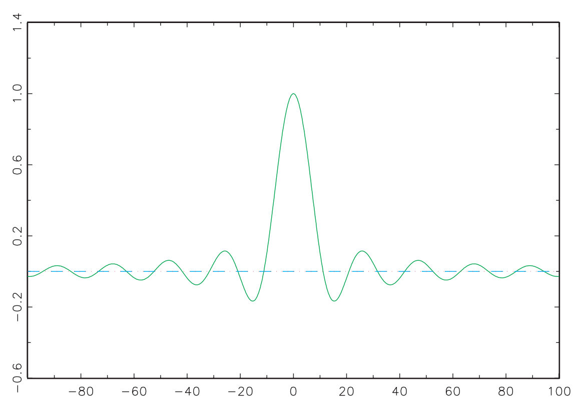
The second restriction is more important. Deconvolution is the division of a cf by another cf. Because division by 0 should be avoided, it is usually assumed that the cf in the denominator is nonzero for all $-\infty < t < \infty$. For instance the cf of the normal distribution with mean 0 (which is a real valued function) is greater than 0 for all t. This choice for the

[13] We could also multiply the integrand by a weight function that down weights the tails for finite n.

[14] The T_n in the integral that defines $\hat{g}_2$ need not be equal to the T_n that appears in the definition of $\hat{g}_1$. This economy in notation will not lead to confusion

cf in the denominator is the leading case in the signal processing literature where a signal is corrupted by mean 0 normal noise. In economic applications such an assumption is not reasonable, i.e. the distribution of x^* could well be nonnormal. In particular, this variable could be bounded. Lukacs (1970), Theorem 7.2.3, p. 202, shows that a distribution with bounded support has a cf that has (countably) infinitely many zeros, if we consider the cf as a function of a complex argument. If the distribution is symmetric (around some value, not necessarily 0) then the zeros will be on the real line. In Figure 2 we give the cf of a truncated (at -.5 and .5) Laplace distribution. The cf of the uniform distribution behaves in the same

Figure 2: Characteristic function of symmetrically truncated (at -3 and 3) Laplace distribution



way. Note that the zeros are 'isolated'. The zeros of asymmetric bounded distributions are usually not on the real line.[15] However, for the truncated Laplace distribution we found that the cf will be close to 0 if the truncation is not too asymmetric. For this reason we consider the case that the cf of the distribution in the denominator has countably many 'isolated' zeros.

Li and Vuong (1998) and Li (2002) assume that the cf is never 0 and that the distribution has bounded support, thereby excluding symmetric distributions with bounded support. They need this assumption to obtain a uniform almost sure bound on the ecf. As we shall see the assumption is also essential for their use of the Von Mises calculus to prove (uniform) consistency of their non-parametric density estimators.

[15] For the truncated Laplace distribution this can be easily shown if the lower truncation is 0. We could not find nor prove the result that asymmetric distributions with bounded support have no zeros on the real line.

Because $\phi_x(t) = \phi_{x^*}(t)\phi_\varepsilon(t)$ we have that $\phi_x(t) = 0$ if $\phi_{x^*}(t) = 0$. Hence in the ratio we divide 0 by 0 for countably many values of t. Without loss of generality we can define $\frac{0}{0} = 0$. The result will not affect the Fourier inversion formula, because it involves countably many values of the integrand and we can change $\phi_\varepsilon(t)$ for countable many t without changing the integral.

The division by 0 does affect the asymptotic analysis of the estimator. To keep things simple we consider the inversion estimator for the case that the distribution of x^* is known

$$\hat{g}_1(\varepsilon) = \frac{1}{2\pi}\int_{-T_n}^{T_n} \frac{\hat{\phi}_x(t)}{\phi_{x^*}(t)}\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T_n}^{T_n} \frac{e^{it(x-\varepsilon)}}{\phi_{x^*}(t)}\mathrm{d}t\mathrm{d}F_n(x) \tag{22}$$

with F_n the empirical cdf of x. The final expression involves a change in the order of integration. Hence we have expressed the estimator as a sample average. This is essentially an application of the Von Mises calculus (see e.g. Serfling (1980)), a technique that is employed by Li and Vuong (1998) and other authors. However

$$\int_{-T_n}^{T_n} \left|\frac{e^{it(x-\varepsilon)}}{\phi_{x^*}(t)}\right| \mathrm{d}t = \int_{-T_n}^{T_n} \left|\frac{1}{\phi_{x^*}(t)}\right| \mathrm{d}t \tag{23}$$

and the integral on the right-hand side diverges for a cf with zeros if T_n is large enough, e.g. if x^* has a symmetric distribution with bounded support. Hence the estimator is a weighted sample average with weights that have a diverging sum.

The solution that we propose for the division by 0 is simple. Instead of dividing by $\phi_{x^*}(t)$ we divide by $\phi_{x^*}(t,\eta_n)$ with

$$\phi_{x^*}(t,\eta_n) = \phi_{x^*}(t)I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right) + \frac{1}{2}\eta_n I\left(|\phi_{x^*}(t)| \le \frac{1}{2}\eta_n\right) \tag{24}$$

with η_n a sequence that converges to 0 at a rate to be specified below. Note that $|\phi_{x^*}(t,\eta_n)| = \frac{1}{2}\eta_n \neq 0$ if $|\phi_{x^*}(t)| \le \frac{1}{2}\eta_n$.
The function $\phi_{x^*}(t,\eta_n)$ is not continuous in t and hence is not a cf. We have

$$\sup_{-\infty<t<\infty} |\phi_{x^*}(t) - \phi_{x^*}(t,\eta_n)| \le \sup_{\{t||\phi_{x^*}(t)|\le\frac{1}{2}\eta_n\}} |\phi_{x^*}(t) - \phi_{x^*}(t,\eta_n)| \le \eta_n \tag{25}$$

and for all t

$$|\phi_{x^*}(t,\eta_n)| \geq \max\left\{|\phi_{x^*}(t)|\,,\frac{1}{2}\eta_n\right\} \tag{26}$$

Consider the estimator

$$\hat{g}_1(\varepsilon) = \mathrm{Re}\frac{1}{2\pi}\int_{-T_n}^{T_n}\frac{\hat{\phi}_x(t)}{\phi_{x^*}(t,\eta_n)}e^{-it\varepsilon}\mathrm{d}t \tag{27}$$

Note that we must take the real part of the function on the right-hand side, because the integrand is not necessarily real. Hence

$$\hat{g}_1(\varepsilon) - g_1(\varepsilon) = \mathrm{Re}\frac{1}{2\pi}\int_{-T_n}^{T_n}\left(\frac{\hat{\phi}_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right)e^{-it\varepsilon}\mathrm{d}t- \tag{28}$$

$$-\frac{1}{2\pi}\int_{|t|>T_n}\phi_\varepsilon(t)e^{-it\varepsilon}\mathrm{d}t$$

For the first term on the right-hand side

$$\left|\frac{1}{2\pi}\int_{-T_n}^{T_n}e^{-it\varepsilon}\left(\frac{\hat{\phi}_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right)\mathrm{d}t\right| \leq \frac{1}{2\pi}\int_{-T_n}^{T_n}\left|\frac{\hat{\phi}_x(t)-\phi_x(t)}{\eta_n}\right|\mathrm{d}t+ \tag{29}$$

$$+\frac{1}{2\pi}\int_{-T_n}^{T_n}\left|\frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right|\mathrm{d}t$$

First consider the second term on the right-hand side of (29) . For all t

$$\left|\frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right| \leq 2\left|\frac{\phi_x(t)}{\phi_{x^*}(t)}\right| \tag{30}$$

and

$$\int_{-\infty}^{\infty}\left|\frac{\phi_x(t)}{\phi_{x^*}(t)}\right|\mathrm{d}t = \int_{-\infty}^{\infty}|\phi_\varepsilon(t)|\,\mathrm{d}t < \infty \tag{31}$$

Hence by dominated convergence for all sequences T_n and $\eta_n = o(1)$, the second term on the right-hand side of (29) converges to 0. The rate of convergence of the first term is determined by the uniform rate of convergence of the empirical cf on intervals of diverging length.

The next lemma gives an almost sure rate of convergence that, as far as we know, is new. It corrects the result in Lemma 1 of Horowitz and Markatou (1996)

Lemma 1 *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*

$$\sup_{|t| \le T_n} \left| \hat{\phi}(t) - \phi(t) \right| = o(\alpha_n) \qquad \text{a.s.} \tag{32}$$

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

Proof See appendix.

This result is used to establish the rate of convergence of the nonparametric Fourier inversion estimators in the next two lemmas. The estimators are

$$\hat{g}_1(\varepsilon) = \mathrm{Re} \frac{1}{2\pi} \int_{-T_n}^{T_n} \frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t, \eta_n)} e^{-it\varepsilon} \mathrm{d}t \tag{33}$$

and

$$\hat{g}_2(x^*, w) = \mathrm{Re} \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^* - itw} \frac{\hat{\phi}_{xw}(s,t) \hat{\phi}_{x^*}(s)}{\hat{\phi}_x(s, \gamma_n)} \mathrm{d}t \mathrm{d}s \tag{34}$$

where the modified ecf is defined analogously to (24). We have

Lemma 2 *Let* ϕ_ε *be absolutely integrable and let* ϕ_{x^*} *be a cf with a countable number of 0's. Define for the sequence* T_n *that satisfies the restrictions of Lemma 1,* $\eta_n = |\phi_{x^*}(T_n)| \neq 0$, *and let* α_n *satisfy the restrictions of Lemma 1 and in addition* $\frac{\alpha_n}{\eta_n} = o(1)$. *Then a.s. for the estimator in (33)*

$$\sup_{(x,x^*) \in \mathcal{X} \times \mathcal{X}^*} |\hat{g}_1(x - x^*) - g_1(x - x^*)| = o\left(\frac{T_n \alpha_n}{\eta_n}\right) \tag{35}$$

with $\mathcal{X}, \mathcal{X}^*$ *the support of* x, x^*, *respectively. These supports may be bounded.*

and

Lemma 3 *Let* ϕ_{x^*w} *be absolutely integrable and let* ϕ_x *have a countable number of 0's. Define for the sequence* S_n *that satisfies the restrictions of Lemma 1,* $\gamma_n = |\phi_x(S_n)| \neq 0$, *and let* α_n *satisfy the restrictions of Lemma 1 and in addition* $\frac{\alpha_n}{\gamma_n} = o(1)$. *For some* $0 < \gamma < \frac{1}{2}$,

$T_n = o\left(\left(\frac{n}{\log n}\right)^{\gamma}\right)$. *Then a.s. for the estimator in (34)*

$$\sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}} |\hat{g}_2(x^*,w) - g_2(x^*,w)| = o\left(\frac{S_n T_n \alpha_n}{\gamma_n}\right) \tag{36}$$

The supports of x^*, x, w, denoted by $\mathcal{X}^*, \mathcal{X}, \mathcal{W}$ respectively, may be bounded.

Proof See appendix.

Comparison to the rate that can be obtained if the distribution of x^* is known reveals that the rate of convergence is not affected by the fact that distribution is estimated in Lemmas 2 and 3. This result is consistent with the result in Diggle and Hall (1993) who consider the Mean Integrated Squared Error of the Fourier inversion estimator.

3.2 The semi-parametric MLE

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

$$f(y|x,w;\theta_0) = \int_{\mathcal{X}^*} f^*(y|x^*,w;\theta_0)\frac{g_1(x-x^*)g_2(x^*,w)}{g(x,w)}\mathrm{d}x^* \tag{37}$$

in which $f^*(y|x^*,w;\theta)$ is the parametric model for the conditional distribution of y given w and the latent x^*. The scores of $f(y|x,w;\theta)$ and $f^*(y|x^*,w;\theta)$ are denoted by $s(y|x,w;\theta)$ and $s^*(y|x,w;\theta)$, respectively. The densities $f_x, f_{x^*}, f_{w|x}$ have support $\mathcal{X}, \mathcal{X}^*, \mathcal{W}$, respectively. These supports may be bounded. The unknown densities in the likelihood are either g_1, g_2 or $f_x, f_{x^*}, f_{w|x}$. We use h to denote either. The first choice is convenient in the consistency proof, while the second choice is appropriate in the computation of the asymptotic variance.

The semi-parametric MLE is defined as

$$\hat{\theta} = \arg\max_{\theta\in\Theta}\sum_{i=1}^{n} \ln \hat{f}(y_i|x_i,w_i;\theta) \tag{38}$$

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 semi-parametric MLE satisfies the moment condition

$$\sum_{i=1}^{n} m(y_i, x_i, w_i, \hat{\theta}, \hat{h}) = 0 \tag{39}$$

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

$$m(y, x, w, \theta, h) = \frac{\int_{\mathcal{X}^*} \frac{\partial f^*(y|x^*, w;\theta)}{\partial \theta} g_1(x - x^*) g_2(x^*, w) \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y|x^*, w;\theta) g_1(x - x^*) g_2(x^*, w) \mathrm{d}x^*} \tag{40}$$

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

Theorem 1 *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 $\theta \in \Theta$

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

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

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

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

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

(A2) The characteristic functions of ε and x^*, w are absolutely integrable.

(A3) For $0 < \gamma < \frac{1}{2}$, $T_n = o\left(\left(\frac{n}{\log n}\right)^{\gamma}\right)$, $S_n = o\left(\left(\frac{n}{\log n}\right)^{\gamma}\right)$, $\alpha_n = o(1)$, $\frac{\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma}}{\alpha_n} = O(1)$, $\theta_n = \inf_{|s| \leq T_n} |\phi_{x^*}(s)|$, $\gamma_n = \inf_{|s| \leq T_n} |\phi_x(s)|$, we have $\frac{T_n \alpha_n}{\theta_n} = O(1)$, $\frac{S_n T_n \alpha_n}{\gamma_n} = O(1)$.

then for the semi-parametric MLE

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

we have if $n, n_1 \to \infty$

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

A sufficient condition for assumption (A1) is that for some $0 < m_0, m_1 < \infty$ and all $(y, x^*, w) \in \mathcal{Y} \times \mathcal{X}^* \times \mathcal{W}$ and $\theta \in \Theta$

$$m_0 \leq f^*(y|x^*, w; \theta) \leq m_1 \tag{41}$$

For example, for a probit model these conditions are easily satisfied if the supports $\mathcal{X}^*$ and $\mathcal{W}$ are bounded.

Lemma 4 *If the assumptions of Theorem 1 hold and in addition*

(A4) $\lim_{n\to\infty} \frac{n}{n_1} = \lambda$, $0 < \lambda < \infty$, *and* $E(m(y, x, w, \theta_0, h_0)m(y, x, w, \theta_0, h_0)') < \infty$.

then (m_n is defined in the Appendix)

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_n(y_i, x_i, w_i, \theta_0, \hat{h}) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m(y_i, x_i, w_i, \theta_0, h_0) + \delta_x(x_i) + \delta_{xw}(x_i, w_i)) - \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} \delta_{x^*}(x_i^*) \right| = \tag{42}$$

$$= o_p(1)$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m(y_i, x_i, w_i, \theta_0, h_0) + \delta_x(x_i) + \delta_{xw}(x_i, w_i)) + \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n_1} \delta_{x^*}(x_i^*) \xrightarrow{d} N(0, \Omega) \tag{43}$$

$$\Omega = E\left[(m(y, x, w, \theta_0, h_0) + \delta_x(x) + \delta_{xw}(x, w)) (m(y, x, w, \theta_0, h_0) + \delta_x(x) + \delta_{xw}(x, w))'\right] +$$

$$+\lambda E[\delta_{x^*}(x^*)\delta_{x^*}(x^*)']$$

and

$$\delta_x(\tilde{x}) = -\int_{\mathcal{X}^*} \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} f^*(y|x^*, w) s_{0n}(y|x, w) f_0(x, w).$$

$$.(K_{1xn}(\tilde{x}, x - x^*) g_{2n}(x^*, w) + K_{2xn}(\tilde{x}, x^*, w) g_{1n}(x - x^*))\, dw\, dx\, dy\, dx^*$$

$$K_{1xn}(\tilde{x}, x - x^*) = \frac{1}{2\pi} \int_{-T_n}^{T_n} \frac{e^{-it(x-x^*)+it\tilde{x}}}{\phi_{x^*}(t, \eta_n)} dt$$

$$K_{2xn}(\tilde{x}, x^*, w) = -\frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^* - itw + is\tilde{x}} I\left(|\phi_x(s)| > \frac{1}{2}\gamma_n\right) \frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s,\gamma_n)^2} ds\, dt$$

and

$$\delta_{x^*}(\tilde{x}^*) = -\int_{\mathcal{X}^*} \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} f^*(y|x^*, w) s_{0n}(y|x, w) f_0(x, w).$$

$$.(K_{1x^*n}(\tilde{x}^*, x - x^*) g_{2n}(x^*, w) + K_{2x^*n}(\tilde{x}^*, x^*, w) g_{1n}(x - x^*))\, dw\, dx\, dy\, dx^*$$

$$K_{1x^*n}(\tilde{x}, x - x^*) = -\frac{1}{2\pi} \int_{-T_n}^{T_n} I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right) e^{-it(x-x^*) + it\tilde{x}} \frac{\phi_x(t)}{\phi_{x^*}(t, \eta_n)^2} dt$$

$$K_{2x^*n}(\tilde{x}, x^*, w) = \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^* - itw + is\tilde{x}} \frac{\phi_{xw}(s,t)}{\phi_x(t,\gamma_n)} ds\, dt$$

and

$$\delta_{xw}(\tilde{x}, \tilde{w}) = -\int_{\mathcal{X}^*} \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} f^*(y|x^*, w) s_{0n}(y|x, w) f_0(x, w).$$

$$.K_{2xwn}(\tilde{x}, \tilde{w}, x^*, w) g_{1n}(x - x^*))\, dw\, dx\, dy\, dx^*$$

$$K_{1xwn}(\tilde{x}, \tilde{w}, x - x^*) \equiv 0$$

$$K_{2xwn}(\tilde{x}, \tilde{w}, x^*, w) = \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^* - itw + is\tilde{x} + it\tilde{w}} \frac{\phi_{x^*}(s)}{\phi_x(s,\gamma_n)} ds\, dt$$

and

$$g_{1n}(x - x^*) = \frac{1}{2\pi} \int_{-T_n}^{T_n} e^{-it(x-x^*)} \frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)} dt$$

$$g_{2n}(x^*, w) = \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^* - itw} \frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s,\gamma_n)} ds\, dt$$

Theorem 2 *If assumptions (A1)-(A4) are satisfied, then if* $n, n_1 \to \infty$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V) \tag{44}$$

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

$$M = E\left(\frac{\partial m(y, z, h_0)}{\partial \theta'}\right)$$

Proof See appendix.

We have left the variance in a form that can be easily estimated. Some simplifications occur is we let $n, n_1 \to \infty$, but the resulting expressions are not so easily estimated.

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

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

where $\theta = (\beta_0, \beta_1, \beta_2)'$ and Φ is the standard normal cdf. 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 (ii),[16] (iv) the semi-parametric MLE developed above that uses both the primary sample $\{y_i, x_i, w_i\}$ for $i = 1, 2, ..., n$ and the secondary sample $\{x_j\}$ for $j = 1, 2, ..., n_1$.

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 1.

[16] We could also have considered the estimator in which the density of x^* given x, w is specified up to a vector of parameters that are estimated together with the regression parameters in the probit. This estimator will perform worse than the one we consider.

Table 1: Simulation results Probit model: $n = 500$, $n_1 = 600$, number of repetitions 200.

	β_1			β_2			β_0		
$\frac{\sigma^2_\varepsilon}{\sigma^2_{x^*}} = 1.96^a$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
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
$\frac{\sigma^2_\varepsilon}{\sigma^2_{x^*}} = 1^b$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
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
$\frac{\sigma^2_\varepsilon}{\sigma^2_{x^*}} = .36^c$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
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] $\beta_1 = 1$,$\beta_2 = -1$,$\beta_0 = .5$; $x^* \sim N(0, .25)$, $w \sim N(0, .25)$, $\varepsilon \sim N(0, \sigma^2_\varepsilon)$; the smoothing parameters are $T = .7$ for the density of ε and $S = T = .6$ for the joint density of x^*, w.
[b] $\beta_1 = 1$,$\beta_2 = -1$,$\beta_0 = .5$; $x^* \sim N(0, .25)$, $w \sim N(0, .25)$, $\varepsilon \sim N(0, \sigma^2_\varepsilon)$; the smoothing parameters are $T = .6$ for the density of ε and $S = T = .7$ for the joint density of x^*, w.
[c] $\beta_1 = 1$,$\beta_2 = -1$,$\beta_0 = .5$; $x^* \sim N(0, .25)$, $w \sim N(0, .25)$, $\varepsilon \sim N(0, \sigma^2_\varepsilon)$; the smoothing parameters are $T = .75$ for the density of ε and $S = T = .2$ for the joint density of x^*, w.

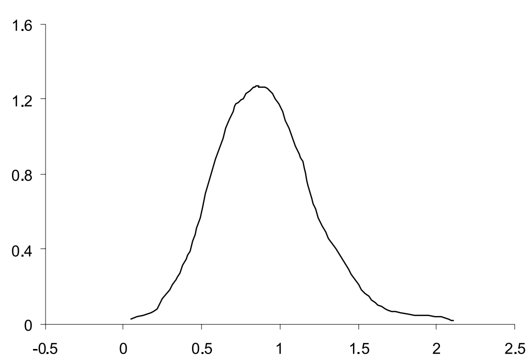
In all cases the smoothing parameters S, T are chosen as suggested in Diggle and Hall (1993). The results are quite robust against changes in the smoothing parameters, and the same is true in our application in section 5.

Table 1 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. As the measurement error variance decreases the MSE of the semi-parametric MLE becomes close to that of the infeasible efficient estimator, so that there is no downside to its use.

We also present the empirical distribution of the semi-parametric MLE. Figure 3 shows the empirical distribution of 200 semi-parametric MLE estimates of β_1. It is close to a normal

Figure 3: Estimate of density of sampling distribution of SPMLE $\hat{\beta}_1$, 200 repetitions



density with the same mean and variance.

The computation of the Fourier inversion estimators in the simulation involve one dimensional (distribution of ε) and two dimensional (distribution of x^*, w) numerical integrals. In the simulations these are computed by Gauss-Laguerre quadrature.[17] 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 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 percent from 100 percent in 1967 and was raised back to 100 percent 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

[17] All computations were performed in Gauss.

related to the probability of leaving welfare. Almost all studies use the AFDC guarantee rather than the reported benefit level of 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 U.S. households and persons living in group quarters. People selected for the SIPP sample are interviewed once every four 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 four months, which are called reference months. Therefore, all rotation groups are interviewed every four months so that we have a panel with quarterly waves.

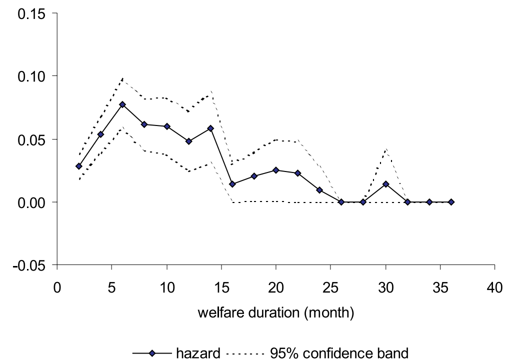
We use the 1992 and 1993 SIPP panels, each of which contains 9 waves.[18] The SIPP 1992 panel follows 21,577 households from October 1991 through December 1994. The SIPP 1993 panel contains information on 21,823 households, from October 1992 through December 1995. Each sample member is followed over a 36-month period.

We consider a flow sample of all single mothers with age 18 to 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

[18] 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.

Figure 4: Empirical hazard rate of welfare durations in SIPP



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 with age 18 to 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 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 5 years of the observation period. Table 2 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 3 reports the results of the Kolmogorov-Smirnov test of the hypothesis that the

Table 2: 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

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.[19]

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 4 reports summary statistics and the result of the Kolmogorov-Smirnov test of equality of the two distributions. A comparison of the 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

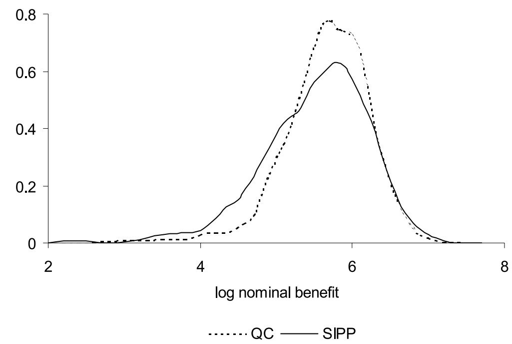
[19] In table 3 we reject the null hypothesis once for the 51 tests. Although the test statistics are not independent, a rejection in a single case is to be expected.

Table 3: Stationarity of distribution nominal benefit levels in QC sample: Kolmogorov-Smirnov test distribution in indicated month vs. the 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				

$*$ significant at 5% level
$**$ significant at 10% level

Figure 5: Density estimates log benefits in SIPP and QC



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

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

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

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

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.

Table 4: 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
Std. Dev.	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
Kolmogorov-Smirnov statistic	.123		.128	
p-value	.0000		.0000	

The primary sample sample contains $t_{i0}, t_{i1}, z_i, \delta_i$ 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

$$P_m(z_i, \theta) = P(t_i^* \geq m | t_i^* \geq m-1, z_i) = \exp\left(-\int_{m-1}^{m} \lambda(t, z_i, \theta) dt\right), \tag{45}$$

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

$$f^*(t_i, \delta_i, |z_i; \theta) = [1 - P_{t_i}(z_i, \theta)]^{\delta_i} \prod_{m=1}^{t_i - 1} P_m(z_i, \theta). \tag{46}$$

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

$$\lambda(t, z_i, \theta) = \lambda_m \exp(z_i \beta), \quad m-1 \leq t < m.$$

This hazard specification implies that

$$P_m(z_i, \theta) = \exp[-\lambda_m \exp(z_i \beta)],$$

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

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

where $\widetilde{x}_i^*$ is the log nominal benefit level and p is the log of the deflator[20].

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

$$\widetilde{x}_i = \widetilde{x}_i^* + \varepsilon_i, \quad \varepsilon_i \perp t_i, z_i, \delta_i, \tag{47}$$

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 5. The MLE are reported in table 6. 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 in line with the bias that we would expect in a linear model with a mismeasured covariate.[21] The other coefficients and the baseline hazard seems 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.

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

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

Table 5: Descriptive statistics, $n = 520$.

	Mean	Std. Dev.	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 (month)	5.46	0.68	2.23	6.86
Log nominal welfare benefits (month)	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/month)	0.234	0.402	0	0.360
State unemployment rate (perc.)	6.72	1.41	2.9	10.9
Education (years)	11.6	2.64	0	18

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.

Table 6: Parameter estimates of duration model, $n = 520, n_1 = 3318$.

	MLE with marginal information		MLE ignoring measurement error	
Variable	MLE	Stand. Error	MLE	Stand. Error
Log real benefits	-0.3368	0.1025	-0.2528	0.0877
Hours worked per week(/24)	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
Age (years/100)	-0.1803	0.9877	-0.3086	0.9663
State unempl. rate (perc.)	-0.0692	0.0505	-0.0691	0.0481
Years of education	0.0112	0.0295	0.0082	0.0290
Disabled	-0.1093	0.1833	-0.1198	0.1867
Baseline hazard (months)				
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

The smoothing parameters are: distribution ε, $T = .7$, distribution of x^*, w, $S = .875$ and $T = .9$.

APPENDIX

1 Notation

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

$$f(y|x,w;\theta_0) = \int_{\mathcal{X}^*} f^*(y|x^*,w;\theta_0)\frac{g_1(x-x^*)g_2(x^*,w)}{g(x,w)}\mathrm{d}x^* \tag{48}$$

in which $f^*(y|x^*,w;\theta)$ is the parametric model for the conditional distribution of y given w and the latent x^*. The scores of $f(y|x,w;\theta)$ and $f^*(y|x^*,w;\theta)$ are denoted by $s(y|x,w;\theta)$ and $s^*(y|x,w;\theta)$, respectively.

The population densities $f_x, f_{x^*}, f_{w|x}$ have support $\mathcal{X}, \mathcal{X}^*, \mathcal{W}$, respectively. The densities are assumed to be bounded on their support. The supports can be bounded or unbounded. Often the assumption of bounded supports is made to obtain simple a.s. rates of convergence (see below).

The moment function $m(y,x,w,\theta,f_x,f_{x^*},f_{w|x})$ is the score of the integrated likelihood

$$m(y,x,w,\theta,f_x,f_{x^*},f_{w|x}) = \frac{\int_{\mathcal{X}^*}\frac{\partial f^*(y|x^*,w;\theta)}{\partial\theta}g_1(x-x^*)g_2(x^*,w)\mathrm{d}x^*}{\int_{\mathcal{X}^*}f^*(y|x^*,w;\theta)g_1(x-x^*)g_2(x^*,w)\mathrm{d}x^*} \tag{49}$$

with

$$g_1(x-x^*) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-it(x-x^*)}\frac{\phi_x(t)}{\phi_{x^*}(t)}\mathrm{d}t \tag{50}$$

$$g_2(x^*,w) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-iux^*-ivw}\frac{\phi_{xw}(u,v)\phi_{x^*}(u)}{\phi_x(u)}\mathrm{d}u\mathrm{d}v \tag{51}$$

and

$$\phi_x(t) = \int_{\mathcal{X}} e^{itx} f_x(x)\mathrm{d}x \tag{52}$$

$$\phi_{x^*}(t) = \int_{\mathcal{X}^*} e^{itx^*} f_{x^*}(x^*)\mathrm{d}x^* \tag{53}$$

$$\phi_{xw}(t) = \int_{\mathcal{X}} \int_{\mathcal{W}} e^{itx} f_x(x)\mathrm{d}x \tag{54}$$

2 Organization of the proof

The first step is to give conditions under which the non-parametric estimators of $g_1(x-x^*)$ and $g_2(x^*,w)$ are uniformly (in x, x^* and x^*, w, respectively) consistent. In Lemma 1 we give a new a.s. bound on the empirical characteristic function that does not require that the support is bounded. We also establish consistency if the support of the random variables is bounded.

The second step of the proof is to establish Fréchet differentiability of the moment function (or functional) with respect to $f_x, f_{x^*}, f_{w|x}$. The Fréchet differential linearizes the moment function(al) in $f_x, f_{x^*}, f_{w|x}$ and this is needed to prove asymptotic normality of the semi-parametric MLE. The expected value of the Fréchet derivative is the term that is added to the moment function evaluated in the population densities to obtain the influence function of the estimator.

3 Rate of convergence of the empirical characteristic function

We first prove a general result on the a.s. rate of convergence of the empirical characteristic function.

Lemma 1 *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*

$$\sup_{|t|\leq T_n}\left|\hat{\phi}(t)-\phi(t)\right| = o(\alpha_n) \qquad \text{a.s.} \tag{55}$$

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

Proof. Consider the parametric class of functions $\mathcal{G}_n = \{e^{itx}||t| \leq T_n\}$. 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 $\mathcal{G}_{1n} = \{\cos(tx)||t| \leq T_n\}$ and $\{\mathcal{F}_{2n} = \sin(tx)||t| \leq T_n\}$. Because $|\cos(t_2x) - \cos(t_1x)| \leq |x||t_2 - t_1|$ 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 $\{t||t| \leq T_n\}$ 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 $|\sin(t_2x)-\sin(t_1x)| \leq |x||t_2-t_1|$, 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

$$\mathcal{N}_1(\varepsilon, P, \mathcal{G}_n) \leq A\frac{T_n}{\varepsilon} \tag{56}$$

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 $\mathcal{N}_1(\varepsilon, P, \mathcal{G}_n) \leq A\varepsilon^{-W}$ 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)

$$\Pr\left(\sup_{|t|\leq T_n}|\hat{\phi}(t)-\phi(t)| > 2\varepsilon_n\right) \leq 2A\left(\frac{\varepsilon_n}{T_n}\right)^{-1}\exp\left(-\frac{1}{128}n\varepsilon_n^2\right) + \tag{57}$$

$$+\Pr\left(\sup_{|t|\leq T_n}\hat{\phi}(2t) > 64\right)$$

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

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

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. □

Remark 1 Horowitz and Markatou (1996), Lemma 1, p. 164, claim that

$$\sup_{|t|<\infty}\left|\hat{\phi}(t) - \phi(t)\right| = o\left(\sqrt{\frac{\log n}{n}}\right) \qquad \text{a.s.} \tag{58}$$

This cannot be correct, because it would imply uniform convergence of the empirical characteristic function without bounds on t, a result that does not hold (see e.g. Feuerverger and Mureika (1977), p. 89). The problem with their proof is that they assume that the functions e^{itx} have a finite covering number if there is no restriction on t, a statement that is obviously not true. The rate result above does not require any assumption on the tail of F (except existence of the mean). Such assumptions seem necessary, if one uses the usual proof for uniform convergence to obtain a bound on the rate of convergence.

Remark 2 If the support of x is bounded we can obtain a slightly faster rate of convergence. The proof of Theorem 1 in Csörgö (1980) shows that with bounded support the a.s. bound is $T_n(\frac{\log\log n}{n})^{\frac{1}{2}}$. Hence if for $0 < \gamma < \frac{1}{2}$, $T_n = o\left((\frac{n}{\log\log n})^{\gamma}\right)$, then the rate of convergence is at most $\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}-\gamma}$. Because the assumptions that ensure convergence with unbounded support are stronger than those for the case of bounded support we only consider the former case.

Using the same method of proof we obtain the rate of uniform convergence for a bivariate empirical characteristic function.

Lemma 2 *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[22] $S_n = o\left(\left(\frac{n}{\log n}\right)^\gamma\right)$ *and* $T_n = o\left(\left(\frac{n}{\log n}\right)^\gamma\right)$. *Then*

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

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

Proof. The $\frac{\varepsilon}{2}$ covers of $|s| \le S_n$ and $|t| \le T_n$ 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 (56) becomes

$$\mathcal{N}_1(\varepsilon, P, \mathcal{G}_n) \le A\frac{S_n T_n}{\varepsilon^2} \tag{60}$$

Hence in (57) we must replace $\frac{\varepsilon_n}{T_n}$ by $\frac{\varepsilon_n}{S_n}\frac{\varepsilon_n}{T_n}$ and in the next equation $\log\left(\frac{T_n}{\alpha_n}\right)$ by $\log\left(\frac{S_n}{\alpha_n}\right) + \log\left(\frac{T_n}{\alpha_n}\right)$ $\square$.

In the sequel we also need the a.s. rate of convergence of $\frac{\hat{\phi}(t)}{\phi(t)}$. This rate depends on a lower bound on $\phi(t)$ for t large. Define $K_1(t) = \inf_{|s|\le t}|\phi(s)|$. If $\phi(t) \ne 0$ for all t, then continuity of ϕ implies that $K_1(t) > 0$ for all t. Hence we have the following obvious result

Lemma 3 *Under the conditions of Lemma 1 we have for* $0 < \gamma < \frac{1}{2}$ *and* $T_n = o\left(\left(\frac{n}{\log n}\right)^\gamma\right)$

$$\sup_{|t|\le T_n} \left|\frac{\hat{\phi}(t) - \phi(t)}{\phi(t)}\right| = o\left(\frac{\alpha_n}{\theta_n}\right) \qquad \text{a.s.} \tag{61}$$

with $\alpha_n = o(1)$ *and* $\frac{\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma}}{\alpha_n} = O(1)$, *and* $\theta_n = K_1(T_n)$.

For convergence θ_n must go to 0 at a rate that is the same as that of α_n or slower, i.e. the rate is at most $\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma}$. This implies a restriction on the rate of T_n that depends on the tail behavior of ϕ. For instance, if $\phi(t) \ge C_1 t^{-\theta}$, then $\gamma \le \frac{1}{2(\theta+1)}$. If ϕ is absolutely integrable, then $\theta > 1$ and this implies that $\gamma < \frac{1}{4}$.

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

4 Nonparametric estimators of $g_1(x-x^*)$ and $g_2(x^*,w)$

The nonparametric estimator of the density $g_1(x-x^*)$ is

$$\hat{g}_1(x-x^*) = \frac{1}{2\pi}\int_{-T_n}^{T_n} e^{-it(x-x^*)}\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t)}\mathrm{d}t \tag{62}$$

Lemma 4 *Let ϕ_ε be absolutely integrable and let $\phi_{x^*}(t) \neq 0$ for all t. Define $K_{1x^*}(t) = \inf_{|s|\le t}|\phi_{x^*}(s)|$ and $\theta_n = K_{1x^*}(T_n)$, and let T_n, α_n satisfy the restrictions of Lemma 1. Then a.s.*

$$\sup_{-\infty<x,x^*<\infty}|\hat{g}_1(x-x^*) - g_1(x-x^*)| = o\left(\frac{T_n\alpha_n}{\theta_n^2}\right)$$

Proof. Define $z = x - x^*$. Then

$$\sup_{-\infty<z<\infty}|\hat{g}_1(z)-g_1(z)| \le \sup_{-\infty<z<\infty}\left|\frac{1}{2\pi}\int_{-T_n}^{T_n} e^{-itz}\left(\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right)\mathrm{d}t\right| \tag{63}$$

$$+\sup_{-\infty<z<\infty}\left|\frac{1}{2\pi}\int_{-\infty}^{-T_n} e^{-itz}\phi_\varepsilon(t)\mathrm{d}t\right| + \sup_{-\infty<z<\infty}\left|\frac{1}{2\pi}\int_{T_n}^{\infty} e^{-itz}\phi_\varepsilon(t)\mathrm{d}t\right|$$

We give bounds on the terms that are uniform over $-\infty < z < \infty$. First, we consider the first term on the right-hand side that is bounded by

$$\frac{1}{2\pi}\int_{-T_n}^{T_n}\left|\frac{\hat{\phi}_x(t)-\phi_x(t)}{\phi_{x^*}(t)}\right|\left|\frac{1}{\frac{\hat{\phi}_{x^*}(t)}{\phi_{x^*}(t)}}\right|\mathrm{d}t+ \tag{64}$$

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

By Lemma 2 we have that a.s. with $K_{1x^*}(t) = \inf_{|s|\le t}|\phi_{x^*}(s)|$ and $\theta_n = K_{1x^*}(T_n)$ Hence (64) is a.s. bounded by (α_n satisfies the restrictions of Lemma 1)

$$\frac{T_n o(\alpha_n)}{\theta_n\left(1-o\left(\frac{\alpha_n}{\theta_n}\right)\right)} + \frac{T_n o\left(\frac{\alpha_n}{\theta_n}\right)}{\theta_n\left(1-o\left(\frac{\alpha_n}{\theta_n}\right)\right)}$$

The other (non-stochastic) terms in (63) are bounded by (note that $|\phi_\varepsilon(t)|$ is symmetric around 0)

$$O\left(\int_{T_n}^{\infty} |\phi_\varepsilon(t)|\mathrm{d}t\right) \tag{65}$$

which is $o(1)$ if ϕ_ε is absolutely integrable $\Box$.

Remark The nonparametric estimator converges a.s. uniformly for all x, x^* if $\frac{\alpha_n}{\theta_n^2} = O(1)$. Also note that the result does not require an assumption on the support of x^*.

Next we consider the case that the the cf of x^* has a countable number of 'isolated' zeros. For all t

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

Hence, if the number of 0's of $\phi_{x^*}(t)$ is countable, we have for all t, if we define $\frac{0}{0} = 0$,

$$\phi_\varepsilon(t) = \frac{\phi_x(t)}{\phi_{x^*}(t)} \tag{66}$$

Hence, if we define $\varepsilon = x - x^*$, the Fourier inverse

$$g_1(\varepsilon) = \frac{1}{2\pi}\int_{\infty}^{\infty} e^{-it\varepsilon}\frac{\phi_x(t)}{\phi_{x^*}(t)}\mathrm{d}t \tag{67}$$

is well-defined. An estimator is obtained if the cf of x is replaced by the empirical cf $\hat{\phi}_x(t) = \int_{-\infty}^{\infty} e^{itx}\mathrm{d}F_n(x)$ and we integrate over $[-T_n, T_n]$. A change in the order of integration gives

$$\hat{g}_1(\varepsilon) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T_n}^{T_n} \frac{e^{it(x-\varepsilon)}}{\phi_{x^*}(t)}\mathrm{d}t\mathrm{d}F_n(x) \tag{68}$$

Hence, we can express the estimator as a sample average. However,

$$\int_{-T_n}^{T_n}\left|\frac{e^{it(x-\varepsilon)}}{\phi_{x^*}(t)}\right|\mathrm{d}t = \int_{-T_n}^{T_n}\left|\frac{1}{\phi_{x^*}(t)}\right|\mathrm{d}t$$

and the latter integral diverges if the cf can be 0. For instance, the cf of the uniform distribution on $[-a, a]$ is $\phi_{x^*}(t) = \frac{\sin ta}{ta}$ and $\int_{-T_n}^{T_n} \frac{at}{\sin at}\mathrm{d}t$ diverges if $T_n > \frac{\pi}{a}$. In general, symmetric bounded distributions have cf's with infinitely, but countably many 0's. Hence the integral of the inverse cf diverges if T_n is large enough. This precludes the use of e.g. the Von Mises calculus (see e.g. Serfling (1980)), because the corresponding derivatives are infinite.

The next step is to propose a solution to this problem. First, we assume that the distribution is known, a common assumption in the deconvolution literature. Let $\eta > 0$ and define

$$\phi_{x^*}(t,\eta) = \phi_{x^*}(t) I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta\right) + \frac{1}{2}\eta I\left(|\phi_{x^*}(t)| \leq \frac{1}{2}\eta\right) \tag{69}$$

The function $\phi_{x^*}(t,\eta)$ is not continuous in t and hence is not a cf. We have

$$\sup_{-\infty<t<\infty} |\phi_{x^*}(t) - \phi_{x^*}(t,\eta)| \leq \sup_{\{t||\phi_{x^*}(t)|\frac{1}{2}\eta\}} |\phi_{x^*}(t) - \phi_{x^*}(t,\eta)| \leq \eta \tag{70}$$

and for all t

$$|\phi_{x^*}(t,\eta)| \geq \max\left\{|\phi_{x^*}(t)|, \frac{1}{2}\eta\right\} \tag{71}$$

Consider the estimator

$$\hat{g}_1(\varepsilon) = \mathrm{Re}\frac{1}{2\pi}\int_{-T_n}^{T_n} \frac{\hat{\phi}_x(t)}{\phi_{x^*}(t,\eta_n)} e^{-it\varepsilon}\mathrm{d}t \tag{72}$$

Note that we must take the real part of the function on the right-hand side, because the integrand is not necessarily real. Hence

$$\hat{g}_1(\varepsilon) - g_1(\varepsilon) = \mathrm{Re}\frac{1}{2\pi}\int_{-T_n}^{T_n} \left(\frac{\hat{\phi}_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right) e^{-it\varepsilon}\mathrm{d}t - \tag{73}$$

$$-\frac{1}{2\pi}\int_{|t|>T_n} \phi_\varepsilon(t) e^{-it\varepsilon}\mathrm{d}t$$

Hence for the first term on the right-hand side

$$\left|\frac{1}{2\pi}\int_{-T_n}^{T_n} e^{-it\varepsilon}\left(\frac{\hat{\phi}_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right)\mathrm{d}t\right| \leq \frac{1}{2\pi}\int_{-T_n}^{T_n} \left|\frac{\hat{\phi}_x(t) - \phi_x(t)}{\phi_{x^*}(t,\eta_n)}\right| \mathrm{d}t + \tag{74}$$

$$+\frac{1}{2\pi}\int_{-T_n}^{T_n} \left|\frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right| \mathrm{d}t$$

First consider the second term on the right-hand side. For all t

$$\left|\frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)}\right| \leq 2\left|\frac{\phi_x(t)}{\phi_{x^*}(t)}\right|$$

and

$$\int_{-\infty}^{\infty}\left|\frac{\phi_x(t)}{\phi_{x^*}(t)}\right| \mathrm{d}t = \int_{-\infty}^{\infty}|\phi_\varepsilon(t)|\,\mathrm{d}t < \infty$$

Hence by dominated convergence for all sequences T_n and $\eta_n = o(1)$, the second term on the right-hand side of (74) converges to 0.

For first term on the right-hand side of (74) we have

$$\frac{1}{2\pi}\int_{-T_n}^{T_n}\left|\frac{\hat{\phi}_x(t)-\phi_x(t)}{\phi_{x^*}(t,\eta_n)}\right| \mathrm{d}t \leq CT_n\frac{\sup_{|t|\leq T_n}\left|\hat{\phi}_x(t)-\phi_x(t)\right|}{\inf_{|t|\leq T_n}|\phi_{x^*}(t,\eta_n)|} = o\left(\frac{\alpha_n T_n}{\eta_n}\right) \tag{75}$$

because for T_n sufficiently large the interval $[-T_n, T_n]$ always contains 0's of $\phi_{x^*}(t)$. Without loss of generality we can choose $\eta_n = |\phi_{x^*}(T_n)| \neq 0$. The rate of convergence is essentially the same as in the case that $\phi_{x^*}(t)$ is nowhere 0.

Hence we have proved

Lemma 5 *Let ϕ_ε be absolutely integrable and let $\phi_{x^*}(t)$ be a known cf with a countable number of zeros. Define for the sequence T_n that satisfies the restrictions of Lemma 1, $\eta_n = |\phi_{x^*}(T_n)| \neq 0$, and let α_n satisfy the restrictions of Lemma 1 and in addition $\frac{\alpha_n}{\eta_n} = o(1)$. Then a.s. for the estimator in (72)*

$$\sup_{(x,x^*)\in\mathcal{X}\times\mathcal{X}^*}|\hat{g}_1(x-x^*) - g_1(x-x^*)| = o\left(\frac{T_n\alpha_n}{\eta_n}\right)$$

with $\mathcal{X}, \mathcal{X}^*$ the support of x, x^*, respectively. These supports may be bounded.

Next we consider the case that the cf of x^* is estimated. We redefine $\phi_{x^*}(t,\eta_n)$ as

$$\phi_{x^*}(t,\eta_n) = \phi_{x^*}(t)I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right) + \frac{1}{2}\eta_n\mathrm{sign}(\phi_{x^*}(t))I\left(|\phi_{x^*}(t)| \leq \frac{1}{2}\eta_n\right) \tag{76}$$

with $\operatorname{sign}(\phi(t))$ short-hand for $\operatorname{sign}(\operatorname{Re}\phi(t))$.[23] Note that this change of definition leaves $|\phi_{x^*}(t,\eta_n)|$ unchanged, and all statements that depend on $\phi_{x^*}(t,\eta_n)$ only through $|\phi_{x^*}(t,\eta_n)|$ still apply. We also define

$$\hat{\phi}_{x^*}(t,\eta_n) = \hat{\phi}_{x^*}(t) I\left(\left|\hat{\phi}_{x^*}(t)\right| > \frac{1}{2}\eta_n\right) + \frac{1}{2}\eta_n \operatorname{sign}(\hat{\phi}_{x^*}(t)) I\left(\left|\hat{\phi}_{x^*}(t)\right| \leq \frac{1}{2}\eta_n\right) \tag{77}$$

$$\tilde{\phi}_{x^*}(t,\eta_n) = \hat{\phi}_{x^*}(t) I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right) + \frac{1}{2}\eta_n \operatorname{sign}(\phi_{x^*}(t)) I\left(|\phi_{x^*}(t)| \leq \frac{1}{2}\eta_n\right) \tag{78}$$

From Lemma 1 we have for all $|t| \leq T_n$

$$|\tilde{\phi}_{x^*}(t)| \geq |\phi_{x^*}(t)| - o(\alpha_n) \qquad \text{a.s.} \tag{79}$$

By (78) we have that $|\tilde{\phi}_{x^*}(t,\eta_n)| \geq \frac{1}{2}\eta_n$ unless both $|\phi_{x^*}(t)| > \frac{1}{2}\eta_n$ and $|\tilde{\phi}_{x^*}(t)| < \frac{1}{2}\eta_n$. In that case by (79)

$$|\tilde{\phi}_{x^*}(t,\eta_n)| = |\tilde{\phi}_{x^*}(t)| \geq \frac{1}{2}\eta_n - o(\alpha_n) \qquad \text{a.s.} \tag{80}$$

Hence

$$\inf_{|t|\leq T_n} |\tilde{\phi}_{x^*}(t,\eta_n)| \geq \frac{1}{2}\eta_n - o(\alpha_n) \qquad \text{a.s.} \tag{81}$$

Obviously

$$\sup_{|t|\leq T_n} |\tilde{\phi}_{x^*}(t,\eta_n) - \phi_{x^*}(t,\eta_n)| = o(\alpha_n) \qquad \text{a.s.} \tag{82}$$

Also

$$\begin{aligned} \hat{\phi}_{x^*}(t,\eta_n) - \tilde{\phi}_{x^*}(t,\eta_n) &= \hat{\phi}_{x^*}(t) - \operatorname{sign}(\phi_{x^*}(t))\frac{1}{2}\eta_n \ \text{ if } \ |\hat{\phi}_{x^*}(t)| \geq \frac{1}{2}\eta_n,\ |\phi_{x^*}(t)| < \frac{1}{2}\eta_n \\ &= \operatorname{sign}(\hat{\phi}_{x^*}(t))\frac{1}{2}\eta_n - \hat{\phi}_{x^*}(t) \ \text{ if } \ |\hat{\phi}_{x^*}(t)| < \frac{1}{2}\eta_n,\ |\phi_{x^*}(t)| \geq \frac{1}{2}\eta_n \end{aligned} \tag{83}$$

and 0 otherwise. By Lemma 1 for $|t| \leq T_n$, $|\hat{\phi}_{x^*}(t) - \phi_{x^*}(t)| = o(\alpha_n)$ with probability 1. Hence, if $\frac{\alpha_n}{\eta_n} = o(1)$, then for all t such that $|t| \leq T_n, |\hat{\phi}_{x^*}(t)| \geq \frac{1}{2}\eta_n,\ |\phi_{x^*}(t)| < \frac{1}{2}\eta_n$ the sup of $|\hat{\phi}_{x^*}(t,\eta_n) - \tilde{\phi}_{x^*}(t,\eta_n)|$ is $o(\alpha_n)$ almost surely. For the second line in (83) we prove in the

[23] Without loss of generality, we set $sign(0) = 1$.

same way that sup of the absolute deviation is $o(\alpha_n)$ almost surely, so that we have

$$\sup_{|t|\leq T_n} |\hat{\phi}_{x^*}(t,\eta_n) - \tilde{\phi}_{x^*}(t,\eta_n)| = o(\alpha_n) \qquad \text{a.s.} \tag{84}$$

The estimator is

$$\hat{g}_1(\varepsilon) = \mathrm{Re}\frac{1}{2\pi}\int_{-T_n}^{T_n} \frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t,\eta_n)} e^{-it\varepsilon} \mathrm{d}t \tag{85}$$

First we consider the infeasible estimator

$$\hat{g}_1(\varepsilon) = \mathrm{Re}\frac{1}{2\pi}\int_{-T_n}^{T_n} \frac{\hat{\phi}_x(t)}{\tilde{\phi}_{x^*}(t,\eta_n)} e^{-it\varepsilon} \mathrm{d}t \tag{86}$$

and we determine its rate of convergence. We need a bound on

$$\left| \frac{1}{2\pi}\int_{-T_n}^{T_n} e^{-it\varepsilon} \left(\frac{\hat{\phi}_x(t)}{\hat{\phi}_{x^*}(t,\eta_n)} - \frac{\phi_x(t)}{\phi_{x^*}(t)} \right) \mathrm{d}t \right| \leq \tag{87}$$

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

First, consider the second term on the right-hand side. The integrand in that term is bounded by

$$\left| \frac{\phi_x(t)}{\tilde{\phi}_{x^*}(t,\eta_n)} \right| + \left| \frac{\phi_x(t)}{\phi_{x^*}(t)} \right| \tag{88}$$

For the first term

$$\left| \frac{\phi_x(t)}{\tilde{\phi}_{x^*}(t,\eta_n)} \right| = \left| \frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)} \right| \frac{1}{\left| \frac{\tilde{\phi}_{x^*}(t,\eta_n) - \phi_{x^*}(t,\eta_n)}{\phi_{x^*}(t,\eta_n)} + 1 \right|} \leq$$

$$\leq |\phi_\varepsilon(t)| \frac{1}{1 - \left| \frac{\tilde{\phi}_{x^*}(t,\eta_n) - \phi_{x^*}(t,\eta_n)}{\phi_{x^*}(t,\eta_n)} \right|} \leq |\phi_\varepsilon(t)| \frac{1}{1 - o\left(\frac{\alpha_n}{\eta_n}\right)}$$

for $|t| \leq T_n$ and almost surely. Hence (88) is bounded by

$$|\phi_\varepsilon(t)| \frac{2 - o\left(\frac{\alpha_n}{\eta_n}\right)}{1 - o\left(\frac{\alpha_n}{\eta_n}\right)} \qquad \text{a.s.} \tag{89}$$

By dominated convergence we have that if $\frac{\alpha_n}{\eta_n} = o(1)$, then the second term on the right-hand side of (87) is $o(1)$ almost surely.

By Lemma 1 and (81) the first term on the right-hand side of (87) is bounded by

$$C \sup_{|t| \leq T_n} \left| \hat{\phi}_x(t) - \phi_x(t) \right| \frac{T_n}{\eta_n} = o\left(\frac{\alpha_n T_n}{\eta_n} \right) \qquad \text{a.s.} \tag{90}$$

Finally

$$|\hat{g}_1(\varepsilon) - \tilde{g}_1(\varepsilon)| \leq \frac{1}{2\pi} \int_{|t| \leq T_n} \left| \frac{\hat{\phi}_x(t)}{\tilde{\phi}_{x^*}(t, \eta_n)} \right| \left| \frac{\hat{\phi}_{x^*}(t, \eta_n) - \tilde{\phi}_{x^*}(t, \eta_n)}{\hat{\phi}_{x^*}(t, \eta_n)} \right| \mathrm{d}t \leq \tag{91}$$

$$\leq o\left(\frac{\alpha_n}{\eta_n} \right) \int_{|t| \leq T_n} \left| \frac{\hat{\phi}_x(t)}{\tilde{\phi}_{x^*}(t, \eta_n)} \right| \mathrm{d}t \qquad \text{a.s.}$$

with

$$\int_{|t| \leq T_n} \left| \frac{\hat{\phi}_x(t)}{\tilde{\phi}_{x^*}(t, \eta_n)} \right| \mathrm{d}t \leq \int_{|t| \leq T_n} \left| \frac{\hat{\phi}_x(t) - \phi_x(t)}{\tilde{\phi}_{x^*}(t, \eta_n)} \right| \mathrm{d}t + \int_{|t| \leq T_n} \left| \frac{\phi_x(t)}{\tilde{\phi}_{x^*}(t, \eta_n)} \right| \mathrm{d}t \leq \tag{92}$$

$$o\left(\frac{\alpha_n T_n}{\eta_n} \right) + \frac{\int_{|t| \leq T_n} |\phi_\varepsilon(t)| \mathrm{d}t}{1 - o\left(\frac{\alpha_n}{\eta_n} \right)} < \infty \qquad \text{a.s}$$

We have proved

Lemma 6 *Let ϕ_ε be absolutely integrable and let $\phi_{x^*}(t)$ be a cf with a countable number of 0's. Define for the sequence T_n that satisfies the restrictions of Lemma 1, $\eta_n = |\phi_{x^*}(T_n)| \neq 0$, and let α_n satisfy the restrictions of Lemma 1 and in addition $\frac{\alpha_n}{\eta_n} = o(1)$. Then a.s. for the estimator in (85)*

$$\sup_{(x, x^*) \in \mathcal{X} \times \mathcal{X}^*} |\hat{g}_1(x - x^*) - g_1(x - x^*)| = o\left(\frac{T_n \alpha_n}{\eta_n} \right)$$

with $\mathcal{X}, \mathcal{X}^*$ the support of x, x^*, respectively. These supports may be bounded.

Remark This proof applies to the case considered in Lemma 4 as well. Hence, a more careful analysis reveals that we can use the absolute integrability of the cf of ε to deal with the second term in the bound which is $o(1)$ a.s. under the conditions that ensure a.s. convergence to 0 of the first term. This result is consistent with the result in Diggle and Hall (1993) who indicate that the convergence speed is not affected by the estimation of the denominator. In the sequel we will use the rate established in this lemma.

For the nonparametric estimator of $g_2(x^*, w)$ we use the same method of proof as in Lemma 5. This means that there are no restrictions on the support of the random variables. The estimator is

$$\hat{g}_2(x^*, w) = \mathrm{Re}\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw}\frac{\hat{\phi}_{xw}(s,t)\hat{\phi}_{x^*}(s)}{\hat{\phi}_x(s,\gamma_n)}\mathrm{d}t\mathrm{d}s \tag{93}$$

with an obvious definition of $\hat{\phi}_x(s,\gamma_n)$.

Lemma 7 *Let $\phi_{x^*w}(t)$ be absolutely integrable and let $\phi_x(t)$ have a countable number of 0's. Define for the sequence S_n that satisfies the restrictions of Lemma 1, $\gamma_n = |\phi_x(S_n)| \neq 0$, and let α_n satisfy the restrictions of Lemma 1 and in addition $\frac{\alpha_n}{\gamma_n} = o(1)$. Then a.s. for the estimator in (93)*

$$\sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}} |\hat{g}_2(x^*,w) - g_2(x^*,w)| = o\left(\frac{S_nT_n\alpha_n}{\gamma_n}\right)$$

The supports of x^*, x, w, denoted by $\mathcal{X}^*, \mathcal{X}, \mathcal{W}$ respectively, may be bounded.

Proof. We have

$$\sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}} |\hat{g}_2(x^*,w) - g_2(x^*,w)| \leq \tag{94}$$

$$\leq \sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}}\left|\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw}\left(\frac{\hat{\phi}_{xw}(s,t)\hat{\phi}_{x^*}(s)}{\hat{\phi}_x(s)} - \frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s)}\right)\mathrm{d}s\mathrm{d}t\right| +$$

$$+ \sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}}\left|\int_{|s|>S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw}\phi_{x^*w}(s,t)\mathrm{d}s\mathrm{d}t\right| +$$

$$+ \sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}}\left|\int_{-S_n}^{S_n}\int_{|t|>T_n} e^{-isx^*-itw}\phi_{x^*w}(s,t)\mathrm{d}s\mathrm{d}t\right| +$$

$$+ \sup_{(x^*,w)\in\mathcal{X}^*\times\mathcal{W}}\left|\int_{|s|>S_n}\int_{|t|>T_n} e^{-isx^*-it}\phi_{x^*w}(s,t)\mathrm{d}s\mathrm{d}t\right|$$

If $\phi_{x^*w}(t)$ is absolutely integrable, then the final three terms are $o(1)$. The first term is bounded by

$$\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n}\left|\frac{\hat{\phi}_{x^*}(s)}{\hat{\phi}_x(s,\gamma_n)}\right||\hat{\phi}_{xw}(s,t)-\phi_{xw}(s,t)|\mathrm{d}s\mathrm{d}t+ \tag{95}$$

$$+\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n}\left|\frac{\phi_{xw}(s,t)}{\hat{\phi}_x(s,\gamma_n)}\right||\hat{\phi}_{x^*}(s)-\phi_{x^*}(s)|\mathrm{d}s\mathrm{d}t+$$

$$+\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n}\left|\frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\hat{\phi}_x(s,\gamma_n)}-\frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s,\gamma_n)}\right|\mathrm{d}s\mathrm{d}t$$

Using the same line of proof as in Lemma 5, the final term is $o(1)$ almost surely, if $\frac{\alpha_n}{\gamma_n}=o(1)$. Using Lemma 2 the first two terms are almost surely of order $\frac{o(\alpha_n)T_nS_n}{\gamma_n-o(\alpha_n)}=o\left(\frac{\alpha_nT_nS_n}{\gamma_n}\right)$. $\square$

5 Consistency

First we linearize of the moment function. Let h be the joint density of x^*,x,w. Under the assumptions made $h(x^*,x,w)=g_1(x-x^*)g_2(x^*,w)$. Both the population densities g_{10},g_{20} and their estimators are obtained by Fourier inversion. Because the corresponding characteristic functions are assumed to be absolutely integrable, g_{10},g_{20} 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

$$m(y,x,w,\theta,h)=\frac{\int_{\mathcal{X}^*}\frac{\partial f^*}{\partial\theta}(y|x^*,w;\theta)h(x^*,x,w)\mathrm{d}x^*}{\int_{\mathcal{X}^*}f^*(y|x^*,w;\theta)h(x^*,x,w)\mathrm{d}x^*} \tag{96}$$

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 and $f_0(y,x,w,\theta)=\int_{\mathcal{X}^*}f^*(y|x^*,w;\theta)h_0(x^*,x,w)\mathrm{d}x^*$.

Both the numerator and denominator in (96) are linear in h. Hence m is Fréchet differentiable in h and

$$\sup_{y,x,w}|m(y,x,w,\theta,h)-m(y,x,w,\theta,h_0)- \tag{97}$$

$$-\int_{\mathcal{X}^*}\left[\frac{f^*(y|x^*,w;\theta)}{f_0(y,x,w;\theta)}(s^*(y|x^*,w;\theta)-s_0(y|x,w;\theta))\right].$$

$$.(h(x^*,x,w)-h_0(x^*,x,w))\mathrm{d}x^*| = o(||h-h_0||)$$

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

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

$$|m(y,x,w,\theta,h_0)| \leq b_1(y,x,w) \tag{98}$$

with $\mathrm{E}(b_1(y,w,x)) < \infty$. For all h in a (small) neighborhood of h_0 and all $\theta \in \mathcal{B}$

$$\left| \int_{\mathcal{X}^*} \frac{f^*(y|x^*,w;\theta)}{f(y,x,w;\theta)} (s^*(y|x^*,w;\theta) - s(y|x,w;\theta))\mathrm{d}x^* \right| \leq b_2(y,w,x) \tag{99}$$

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 $\theta \in \Theta$

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

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

This is sufficient for (98). For (99) we need in addition that for all $(y,w) \in \mathcal{Y} \times \mathcal{W}$ and $\theta \in \Theta$

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

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

It may be possible to relax this assumption using the type of expansion considered in the asymptotic normality proof.

If assumption (99) holds then by Proposition 2, p. 176 in Luenberger (1969)

$$|m(z,\theta,h) - m(z,\theta,h_0)| \leq b_2(y,x,w) \sup_{x^*,x,w} |h(x^*,x,w) - h_0(x^*,x,w)| \tag{100}$$

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) estimator for h.

Theorem 1 *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* $\theta \in \Theta$

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

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

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

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

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

(A2) The characteristic functions of ε *and* x^*, w *are absolutely integrable.*

(A3) For $0 < \gamma < \frac{1}{2}$, $T_n = o\left(\left(\frac{n}{\log n}\right)^\gamma\right)$, $S_n = o\left(\left(\frac{n}{\log n}\right)^\gamma\right)$, $\alpha_n = o(1)$, $\frac{\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\gamma}}{\alpha_n} = O(1)$, $\theta_n = \inf_{|s| \leq T_n} |\phi_{x^*}(s)|$, $\gamma_n = \inf_{|s| \leq T_n} |\phi_x(s)|$, *we have* $\frac{T_n \alpha_n}{\theta_n} = O(1)$, $\frac{S_n T_n \alpha_n}{\gamma_n} = O(1)$.

then for the semi-parametric MLE

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

we have

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

6 Asymptotic distribution

The first step is to derive the correction term that accounts for the fact that the density $h(x^*, x, w)$ is estimated. This requires to linearize the moment function with respect to the estimated densities. 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.

In the consistency proof we linearized with respect to the densities g_1 and g_2 that appear in the moment function. This linear approximation does not give the asymptotic linear representation of the moment function, because it involves estimated densities. For that reason we linearize here with respect to the densities f_x, f_{x^*}, f_{xw}. The linearization should avoid the infinite derivatives associated with distributions of bounded support. This leads us to linearize $m_n(y,x,w,h)$ where $h = (f_x \; f_{x^*} \; f_{xw})'$, and

$$m_n(y,x,w,h) = \frac{\int_{\mathcal{X}^*} \frac{\partial f^*(y|x^*,w)}{\partial \theta} g_{1n}(x-x^*) g_{2n}(x^*,w) \mathrm{d}x^*}{\int_{\mathcal{X}^*} f^*(y|x^*,w) g_{1n}(x-x^*) g_{2n}(x^*,w) \mathrm{d}x^*} \tag{101}$$

and

$$g_{1n}(x-x^*) = \frac{1}{2\pi} \int_{-T_n}^{T_n} e^{-it(x-x^*)} \frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)} \mathrm{d}t \tag{102}$$

$$g_{2n}(x^*,w) = \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^*-itw} \frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s,\gamma_n)} \mathrm{d}s\mathrm{d}t \tag{103}$$

Because the moment function is itself a function of (modified) characteristic functions, its Fréchet derivatives can be computed using the chain rule (Luenberger (1969), Proposition 1, p. 176). The characteristic functions are linear functionals. The Fréchet differential of the modified characteristic function defined in (69) is at f_{x^*0}

$$\mathrm{D}\phi_{x^*0}(t,\eta_n)(f_{x^*} - f_{x^*0}) = I\left(|\phi_{x^*0}(t)| > \frac{1}{2}\eta_n\right) \int_{\mathcal{X}^*} e^{itx}(f_{x^*}(\tilde{x}) - f_{x^*0}(\tilde{x}))\mathrm{d}\tilde{x} \tag{104}$$

with $\phi_{x^*0}(t)$ the characteristic function of f_{x^*0}. The Fréchet derivative does not exist if $\phi_{x^*0}(t) = \frac{1}{2}\eta_n$, but we can set that derivative to any value without changing the following results. An analogous result holds for $\phi_x(t,\gamma_n)$.

Although we formally take the derivative with respect to f_x, f_{x^*}, f_{xw}, we can also take the derivative withe respect to the corresponding characteristic functions $\phi_x, \phi_{x^*}, \phi_{xw}$. This gives the same result. In the sequel we express the differential as a linear functional of f_x, f_{x^*}, f_{xw}, but if convenient as a linear functional of $\phi_x, \phi_{x^*}, \phi_{xw}$.

As an intermediate step we list the Fréchet derivatives of the densities g_{1n}, g_{2n} with respect to f_x, f_{x^*}, f_{xw}.

$$\mathrm{D}g_{1n}(x-x^*)(h_x) = \int_{\mathcal{X}} K_{1xn}(\tilde{x}, x-x^*)h_x(\tilde{x})\mathrm{d}\tilde{x}$$

$$\mathrm{D}g_{2n}(x^*, w)(h_x) = \int_{\mathcal{X}} K_{2xn}(\tilde{x}, x^*, w)h_x(\tilde{x})\mathrm{d}\tilde{x}$$

$$\mathrm{D}g_{1n}(x-x^*)(h_{x^*}) = \int_{\mathcal{X}^*} K_{1x^*n}(\tilde{x}, x-x^*)h_{x^*}(\tilde{x})\mathrm{d}\tilde{x}$$

$$\mathrm{D}g_{2n}(x^*, w)(h_{x^*}) = \int_{\mathcal{X}^*} K_{2x^*n}(\tilde{x}, x^*, w)h_{x^*}(\tilde{x})\mathrm{d}\tilde{x}$$

$$\mathrm{D}g_{1n}(x^*, w)(h_{xw}) = \int_{\mathcal{W}}\int_{\mathcal{X}} K_{1xwn}(\tilde{x}, \tilde{w}, x-x^*)h_{xw}(\tilde{x}, \tilde{w})\mathrm{d}\tilde{x}\mathrm{d}\tilde{w}$$

$$\mathrm{D}g_{2n}(x^*, w)(h_{xw}) = \int_{\mathcal{W}}\int_{\mathcal{X}} K_{2xwn}(\tilde{x}, \tilde{w}, x^*, w)h_{xw}(\tilde{x}, \tilde{w})\mathrm{d}\tilde{x}\mathrm{d}\tilde{w}$$

with

$$K_{1xn}(\tilde{x}, x-x^*) = \frac{1}{2\pi}\int_{-T_n}^{T_n} \frac{e^{-it(x-x^*)+it\tilde{x}}}{\phi_{x^*}(t, \eta_n)}\mathrm{d}t$$

$$K_{2xn}(\tilde{x}, x^*, w) = -\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}} I\left(|\phi_x(s)| > \frac{1}{2}\gamma_n\right)\frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s,\gamma_n)^2}\mathrm{d}s\mathrm{d}t$$

$$K_{1x^*n}(\tilde{x}, x-x^*) = -\frac{1}{2\pi}\int_{-T_n}^{T_n} I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right)e^{-it(x-x^*)+it\tilde{x}}\frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)^2}\mathrm{d}t$$

$$K_{2x^*n}(\tilde{x}, x^*, w) = \frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}}\frac{\phi_{xw}(s,t)}{\phi_x(t,\gamma_n)}\mathrm{d}s\mathrm{d}t$$

$$K_{1xwn}(\tilde{x}, \tilde{w}, x-x^*) \equiv 0$$

$$K_{2xwn}(\tilde{x}, \tilde{w}, x^*, w) = \frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}+it\tilde{w}}\frac{\phi_{x^*}(s)}{\phi_x(s,\gamma_n)}\mathrm{d}s\mathrm{d}t$$

Note that e.g. K_{1xn} is the Fréchet derivative of g_1 with respect to f_x. If $T_n \to \infty$ and $\eta_n \to 0$, then possibly $|K_{1nx}(\tilde{x}, x-x^*)| \to \infty$ if x^* has bounded support. However, if f_x is the population density of x, then

$$\mathrm{D}g_{1n}(x-x^*)(f_x) = \frac{1}{2\pi}\int_{-T_n}^{T_n} e^{-it(x-x^*)}\frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)}\mathrm{d}t$$

which is finite if $n \to \infty$, because of the absolute integrability of the characteristic function of ε. The same remark applies to $K_{2xn}(\tilde{x}, x^*, w), K_{1x^*n}(\tilde{x}, x-x^*), K_{2x^*n}(\tilde{x}, x^*, w), K_{2xwn}(\tilde{x}, \tilde{w}, x^*, w)$.

A second application of the chain rule gives the Fréchet derivatives of the moment function with respect to h.

$$\mathrm{D}m(y,x,w,h)(h_x) = \int_{\mathcal{X}} L_{xn}(\tilde{x},y,x,w)h_x(\tilde{x})\mathrm{d}\tilde{x}$$

$$\mathrm{D}m(y,x,w,h)(h_{x^*}) = \int_{\mathcal{X}^*} L_{x^*n}(\tilde{x},y,x,w)h_{x^*}(\tilde{x})\mathrm{d}\tilde{x}$$

$$\mathrm{D}m(y,x,w,h)(h_{xw}) = \int_{\mathcal{W}}\int_{\mathcal{X}} L_{xwn}(\tilde{x},\tilde{w},y,x,w)h_{xw}(\tilde{w},\tilde{x})\mathrm{d}\tilde{x}\mathrm{d}\tilde{w}$$

with

$$L_{xn}(\tilde{x},y,x,w) = \int_{\mathcal{X}^*} \frac{f^*(y|x^*,w)}{f_{0n}(y|x,w)}(s^*(y|x^*,w) - s_{0n}(y|x,w)).$$

$$.(K_{1xn}(\tilde{x},x-x^*)g_{2n}(x^*,w) + K_{2xn}(\tilde{x},x^*,w)g_{1n}(x-x^*))\mathrm{d}x^*$$

$$L_{x^*n}(\tilde{x},y,x,w) = \int_{\mathcal{X}^*} \frac{f^*(y|x^*,w)}{f_{0n}(y|x,w)}(s^*(y|x^*,w) - s_{0n}(y|x,w)).$$

$$.(K_{1x^*n}(\tilde{x},x-x^*)g_{2n}(x^*,w) + K_{2x^*n}(\tilde{x},x^*,w)g_{1n}(x-x^*))\mathrm{d}x^*$$

$$L_{xwn}(\tilde{x},\tilde{w},y,x,w) = \int_{\mathcal{X}^*} \frac{f^*(y|x^*,w)}{f_{0n}(y|x,w)}(s^*(y|x^*,w) - s_{0n}(y|x,w)).$$

$$.K_{2xwn}(\tilde{x},\tilde{w},x^*,w)g_{1n}(x-x^*)\mathrm{d}x^*$$

We also need the second-order Fréchet derivatives. The first step is to obtain the derivatives of $f_0(y|x,w)$ with respect to f_x,f_{x^*} and $f_{w|x}$

$$M_{kn}(\tilde{x},y,x,w) = \int_{\mathcal{X}^*} f^*(y|x^*,w).$$

$$.\left(g_{2n}(x^*,x)K_{1kn}(\tilde{x},x-x^*) + g_{1n}(x-x^*)K_{2kn}(\tilde{x},x^*,w)\right)\mathrm{d}x^*$$

for $k=x,x^*$, and

$$M_{xwn}(\tilde{x},\tilde{w},y,x,w) = \int_{\mathcal{X}^*} f^*(y|x^*,w).$$

$$.g_{1n}(x-x^*)K_{2xwn}(\tilde{x},\tilde{w},x^*,w)\mathrm{d}x^*$$

We also need the derivatives of $K_{1xn}, K_{2xn}, K_{1x^*n}, K_{2x^*n}, K_{2xwn}$ with respect to f_x,f_{x^*} and

$f_{w|x}$

$$K_{1x,xn}(\tilde{x}_1, \tilde{x}_2, x - x^*) \equiv 0$$

$$K_{1x,x^*n}(\tilde{x}_1, \tilde{x}_2, x - x^*) = -\frac{1}{2\pi} \int_{-T_n}^{T_n} I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right) \frac{e^{-it(x-x^*)+it\tilde{x}_1+it\tilde{x}_2}}{\phi_{x^*}(t,\eta_n)^2} \mathrm{d}t$$

$$K_{1x,xwn}(\tilde{x}_1, \tilde{x}_2, \tilde{w}_2, x - x^*) \equiv 0$$

$$K_{2x,xn}(\tilde{x}_1, \tilde{x}_2, x^*, w) = -\frac{2}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}_1+is\tilde{x}_2} I\left(|\phi_x(s)| > \frac{1}{2}\gamma_n\right) \frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s,\gamma_n)^3} \mathrm{d}s\mathrm{d}t$$

$$K_{2x,x^*n}(\tilde{x}_1, \tilde{x}_2, x^*, w) = -\frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}_1+is\tilde{x}_2} I\left(|\phi_x(s)| > \frac{1}{2}\gamma_n\right) \frac{\phi_{xw}(s,t)}{\phi_x(s,\gamma_n)^2} \mathrm{d}s\mathrm{d}t$$

$$K_{2x,xwn}(\tilde{x}_1, \tilde{x}_2, \tilde{w}_2, x^*, w) = -\frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}_1+its\tilde{x}_2+it\tilde{w}_2}.$$

$$.I\left(|\phi_x(s)| > \frac{1}{2}\gamma_n\right) \frac{\phi_{x^*}(s)}{\phi_x(s,\gamma_n)^2} \mathrm{d}s\mathrm{d}t$$

$$K_{1x^*,x^*n}(\tilde{x}, x - x^*) = \frac{1}{\pi} \int_{-T_n}^{T_n} I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right) e^{-it(x-x^*)+it\tilde{x}_1+it\tilde{x}_2} \frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)^3} \mathrm{d}t$$

$$K_{1x^*,xwn}(\tilde{x}_1, \tilde{x}_2, \tilde{w}_2, x - x^*) \equiv 0$$

$$K_{2x^*,x^*n}(\tilde{x}_1, \tilde{x}_2, x^*, w) \equiv 0$$

$$K_{2x^*,xwn}(\tilde{x}_1, \tilde{x}_2, x^*, w) = \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}_1+is\tilde{x}_2+it\tilde{w}_2} \frac{1}{\phi_x(t,\gamma_n)} \mathrm{d}s\mathrm{d}t$$

$$K_{1xw,xwn}(\tilde{x}_1, \tilde{w}_1, \tilde{x}_2, \tilde{w}_2, x - x^*) \equiv 0$$

$$K_{2xw,xwn}(\tilde{x}_1, \tilde{w}_1, \tilde{x}_2, \tilde{w}_2, x^*, w) \equiv 0$$

Finally define

$$N_{kn}(\tilde{x}, x^*, y, x, w) = -\frac{f^*(y|x^*, w)}{f_{0n}(y|x, w)^2}(s^*(y|x^*, w) - s_{0n}(y|x, w))M_{kn}(\tilde{x}, y, x, w) -$$

$$-\frac{f^*(y|x^*, w)}{f_{0n}(y|x, w)} L_{kn}(\tilde{x}, y, x, w)$$

for $k = x, x^*$, and

$$N_{xwn}(\tilde{x}, \tilde{w}, x^*, y, x, w) = -\frac{f^*(y|x^*, w)}{f_{0n}(y|x, w)^2}(s^*(y|x^*, w) - s_{0n}(y|x, w))M_{xwn}(\tilde{x}, \tilde{w}, y, x, w) -$$

$$-\frac{f^*(y|x^*,w)}{f_{0n}(y|x,w)}L_{xwn}(\tilde{x},\tilde{w},y,x,w)$$

The second-order Fréchet derivatives of the moment function are

$$L_{kln}(\tilde{x}_1,\tilde{x}_2,y,w,x)=\int_{\mathcal{X}^*}N_{ln}(\tilde{x}_2,x^*,y,x,w)\left((K_{1kn}(\tilde{x}_1,x-x^*)g_{2n}(x^*,w)+\right.$$

$$+K_{2kn}(\tilde{x}_1,x^*,w)g_{1n}(x-x^*))\,\mathrm{d}x^*+\int_{\mathcal{X}^*}\frac{f^*(y|x^*,w)}{f_{0n}(y|x,w)}(s^*(y|x^*,w)-s_{0n}(y|x,w)).$$

$$.(K_{1kn}(\tilde{x}_1,x-x^*)K_{2ln}(\tilde{x}_2,x^*,w)+K_{1kln}(\tilde{x}_1,\tilde{x}_2,x-x^*)g_{2n}(x^*,w)+$$

$$+K_{2kn}(\tilde{x}_1,x^*,w)K_{1ln}(\tilde{x}_2,x-x^*)+K_{2kln}(\tilde{x}_1,\tilde{x}_2,x^*,w)g_{1n}(x-x^*))\mathrm{d}x^*$$

for $k,l=x,x^*$, and

$$L_{xw,ln}(\tilde{x}_1,\tilde{w}_1,\tilde{x}_2,y,w,x)=\int_{\mathcal{X}^*}N_{ln}(\tilde{x}_2,x^*,y,x,w)\left((K_{2xwn}(\tilde{x}_1,x^*,w)g_{1n}(x-x^*)\right)\mathrm{d}x^*+$$

$$+\int_{\mathcal{X}^*}\frac{f^*(y|x^*,w)}{f_{0n}(y|x,w)}(s^*(y|x^*,w)-s_{0n}(y|x,w)).$$

$$.(K_{2xwn}(\tilde{x}_1,\tilde{w}_1,x^*,w)K_{1ln}(\tilde{x}_2,x-x^*)+K_{2xwln}(\tilde{x}_1,\tilde{w}_1,\tilde{x}_2,x^*,w)g_{1n}(x-x^*))\mathrm{d}x^*$$

for $l=x,x^*$.

The first-order Fréchet derivative is used to derive the correction term that accounts for the fact that the densities f_x,f_{x^*} and f_{xw} are estimated. This correction term is equal to the sum of the conditional expectations of the Fréchet derivatives of the moment function with respect to these densities evaluated in θ_0 and in the population f_x, f_{x^*}, f_{xw} given x,w, where the expectation is taken with respect to the population distribution of y,x,w with density $f(y,x,w;\theta_0)=f_0(y,x,w)$. We find

$$\delta_x(\tilde{x})=E(L_{xn}(\tilde{x},y,x,w))=-\int_{\mathcal{X}^*}\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}}f^*(y|x^*,w)s_{0n}(y|x,w)f_0(x,w).$$

$$.(K_{1xn}(\tilde{x},x-x^*)g_{2n}(x^*,w)+K_{2xn}(\tilde{x},x^*,w)g_{1n}(x-x^*))\mathrm{d}w\mathrm{d}x\mathrm{d}y\mathrm{d}x^*$$

$$\delta_{x^*}(\tilde{x}^*)=E(L_{x^*n}(\tilde{x}^*,y,x,w))=-\int_{\mathcal{X}^*}\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}}f^*(y|x^*,w)s_{0n}(y|x,w)f_0(x,w).$$

$$.(K_{1x^*n}(\tilde{x}^*, x - x^*)g_{2n}(x^*, w) + K_{2x^*n}(\tilde{x}^*, x^*, w)g_{1n}(x - x^*))\mathrm{d}w\mathrm{d}x\mathrm{d}y\mathrm{d}x^*$$

$$\delta_{xw}(\tilde{x}, \tilde{w}) = E(L_{xwn}(\tilde{x}, \tilde{w}, y, x, w)) = -\int_{\mathcal{X}^*}\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} f^*(y|x^*, w)s_{0n}(y|x, w)f_0(x, w).$$

$$.K_{2xwn}(\tilde{x}, \tilde{w}, x^*, w)g_{1n}(x - x^*))\mathrm{d}w\mathrm{d}x\mathrm{d}y\mathrm{d}x^*$$

where the notation for the argument indicates which random variable enters the correction term. The correction term now is

$$\alpha(x, w, x^*) = \delta_x(x) + \delta_{x^*}(x^*) + \delta_{xw}(x, w) \tag{105}$$

At the population densities we have that if $T_n = S_n = \infty$ and $\eta_n = \gamma_n = 0$, $E(\delta_x(x)) = E(\delta_{x^*}(x^*)) = E(\delta_{xw}(x, w)) = 0$ so that (105) is indeed the correction term.

We now prove that the semi-parametric MLE is asymptotically linear. We essentially follow the steps of Newey (1994). Define

$$\overline{m}_n(\theta_0, \hat{h}) = \frac{1}{n}\sum_{i=1}^{n} m_n(y_i, z_i, \hat{h}) \tag{106}$$

with $\hat{h} = (\hat{f}_x \ \hat{f}_{x^*} \ \hat{f}_{xw})'$ and m_n defined in (101). The population densities are denoted by $h_0 = (\hat{f}_{x0} \ \hat{f}_{x^*0} \ \hat{f}_{xw0})'$.

If the first two statements of assumption A1 in Theorem 1 hold, then $|s^*(y|x^*, w; \theta_0)|$, $f_{0n}(y|x, w)$, $|s_{0n}(y|x, w)|$ are also bounded from 0 and ∞ on their support and for all h. From their definition $|g_{1n}(x - x^*)|$ and $|g_{2n}(x^*, w)|$ are also bounded on on their support for all h, and the same is true for $|K_{1xn}|$,$|K_{2xn}|$, $|K_{1x^*n}|$, $|K_{2x^*n}|$, $|K_{2xwn}|$ and $|K_{1x,x^*n}|$, $|K_{2x,xn}|$, $|K_{2x,x^*n}|$, $|K_{2x,xwn}|$, $|K_{1x^*,x^*n}|$, $|K_{2x^*,xwn}|$. This in turn implies that $|L_{xn}|$, $|L_{x^*n}|$, $|L_{xwn}|$ and $|M_{xn}|$, $|M_{x^*n}|$, $|M_{xwn}|$ are bounded functions for all h, so that the same is true for $|N_{xn}|$, $|N_{x^*n}|$, $|N_{xwn}|$. Together this implies that the second-order Fréchet derivatives are bounded functions of their arguments for all h and certainly for h in some neighborhood of h_0. By Proposition 3, p. 177 in Luenberger (1969) this implies that (to economize we use the same notation for the integration variables and the arguments of h)

$$\sup_{(\tilde{x}, \tilde{x}^*, \tilde{w})\in\mathcal{X}\times\mathcal{X}^*\times\mathcal{W}} |m_n(y, x, w, h(\tilde{x}, \tilde{x}^*, \tilde{w})) - \tag{107}$$

$$-m_n(y,x,w,h_0(\tilde{x},\tilde{x}^*,\tilde{w})) - \int_{\mathcal{X}} L_{xn}(\tilde{x},y,x,w)(f_x(\tilde{x})-f_{x0}(\tilde{x}))\mathrm{d}\tilde{x}-$$

$$-\int_{\mathcal{X}^*} L_{x^*n}(\tilde{x}^*,y,x,w)(f_{x^*}(\tilde{x}^*)-f_{x^*0}(\tilde{x}^*))\mathrm{d}\tilde{x}^*-$$

$$-\int_{\mathcal{X}}\int_{\mathcal{W}} L_{xwn}(\tilde{x},\tilde{w},y,x,w)(f_{xw}(\tilde{x},\tilde{w})-f_{xw0}(\tilde{x},\tilde{w}))\mathrm{d}\tilde{x}\mathrm{d}\tilde{w}| \leq$$

$$\leq b_3(y,x,w) \sup_{|s|\leq S_n, |t|\leq T_n} |\phi(s,t)-\phi_0(s,t)|^2$$

where $\phi = (\phi_x \ \phi_{x^*} \ \phi_{xw})'$ and ϕ_0 the same characteristic functions for the population distributions. The function $b_3(y,x,w)$ is bounded and hence $E(b_3(y,x,w)) < \infty$. Hence, if we estimate f_x, f_{x^*}, f_{xw} by their empirical distributions, that assign $\frac{1}{n}$ to observations $x_i, w_i, i=1,\ldots,n$ and $\frac{1}{n_1}$ to $x_i^*, w_i, i=1,\ldots,n_1$

$$\left|\sqrt{n}\overline{m}_n(\theta_0,\hat{h}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(m_n(y_i,x_i,w_i,h_0) + \int_{\mathcal{X}} L_{xn}(\tilde{x},y_i,x_i,w_i)(\hat{f}_x(\tilde{x})-f_{x0}(\tilde{x}))\mathrm{d}\tilde{x}+ \right.\right. \tag{108}$$

$$+\int_{\mathcal{X}^*} L_{x^*n}(\tilde{x}^*,y_i,x_i,w_i)(\hat{f}_{x^*}(\tilde{x}^*)-f_{x^*0}(\tilde{x}^*))\mathrm{d}\tilde{x}^*+$$

$$\left.\left.+\int_{\mathcal{X}}\int_{\mathcal{W}} L_{xwn}(\tilde{x},\tilde{w},y_i,x_i,w_i)(\hat{f}_{xw}(\tilde{x},\tilde{w})-f_{xw0}(\tilde{x},\tilde{w}))\mathrm{d}\tilde{x}\mathrm{d}\tilde{w}\right)\right| \leq$$

$$\leq \frac{1}{n}\sum_{i=1}^n b_3(y_i,z_i)\sqrt{n} \sup_{|s|\leq S_n, |t|\leq T_n} |\hat{\phi}(s,t)-\phi_0(s,t)|^2$$

The right-hand side is $o(1)$ almost surely if the S_n, T_n, α_n satisfy the conditions in Lemmas 1 and 2. Note that in the notation we do not distinguish between the sample sizes in the two samples. It is obvious that no assumption on the relative rate at which these samples sizes increase is needed. It suffices that both go to ∞.

The next step is the stochastic equicontinuity of this asymptotically equivalent expression which can be checked for each term. For instance we need to show

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\int_{\mathcal{X}} L_{xn}(\tilde{x},y_i,x_i,w_i)(\hat{f}_x(\tilde{x})-f_{x0}(\tilde{x}))\mathrm{d}\tilde{x}- \right.\right. \tag{109}$$

$$\left.\left.-\int_{\mathcal{Y}}\int_{\mathcal{X}}\int_{\mathcal{W}}\int_{\mathcal{X}} L_{xn}(\tilde{x},y,x,w)(\hat{f}_x(\tilde{x})-f_{x0}(\tilde{x}))\mathrm{d}\tilde{x}f_0(y,x,w)\mathrm{d}y\mathrm{d}x\mathrm{d}w\right)\right| = o_p(1)$$

From the definition of L_{xn} we must consider terms as

$$\left|\frac{1}{2\pi}\int_{-T_n}^{T_n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_{\mathcal{X}^*}\Big(h(y_i,x_i,w_i,x^*)g_{2n}(x^*,w_i)e^{-it(x_i-x^*)}- \right. \tag{110}$$

$$-\int_{\mathcal{Y}}\int_{\mathcal{X}}\int_{\mathcal{W}}h(y,x,w,x^*)g_{2n}(x^*,w)e^{-it(x-x^*)}f_0(y,x,w)\mathrm{d}y\mathrm{d}x\mathrm{d}w\Big)\,\mathrm{d}x^*.$$

$$\left.\cdot\frac{\hat{\phi}_x(t)-\phi_{x0}(t)}{\phi_{x^*}(t,\eta_n)}\mathrm{d}t\right|$$

with

$$h(y,x,w,x^*)=\frac{f^*(y|x^*,w)}{f_{0n}(y|x,w)}(s^*(y|x^*,w)-s_{0n}(y|x,w))$$

which is bounded by

$$\frac{1}{2\pi}\sup_{|t|\leq T_n}T_n\left|\frac{\hat{\phi}_x(t)-\phi_{x0}(t)}{\phi_{x^*}(t,\eta_n)}\right|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\int_{\mathcal{X}^*}h(y_i,x_i,w_i,x^*)g_{2n}(x^*,w_i)\frac{1}{T_n}\int_{-T_n}^{T_n}e^{-it(x_i-x^*)}\mathrm{d}t\mathrm{d}x^*-\right.\right. \tag{111}$$

$$\left.\left.-\int_{\mathcal{Y}}\int_{\mathcal{X}}\int_{\mathcal{W}}\int_{\mathcal{X}^*}h(y,x,w,x^*)g_{2n}(x^*,w)\frac{1}{T_n}\int_{-T_n}^{T_n}e^{-it(x-x^*)}\mathrm{d}t\mathrm{d}x^*f_0(y,x,w)\mathrm{d}y\mathrm{d}x\mathrm{d}w\right)\right|$$

The second factor is a normalized sum of i.i.d. mean 0 random variables. This sum is $O_p(1)$ if the second moment of these random variables is finite. We have

$$\left|\int_{\mathcal{X}^*}h(y,x,w,x^*)g_{2n}(x^*,w)\frac{1}{T_n}\int_{-T_n}^{T_n}e^{-it(x-x^*)}\mathrm{d}t\mathrm{d}x^*\right|\leq 2\int_{\mathcal{X}^*}|h(y,x,w,x^*)g_{2n}(x^*,w)|\mathrm{d}x^*\leq \tag{112}$$

$$\leq C\int_{\mathcal{X}^*}|g_{2n}(x^*,w)|\mathrm{d}x^*$$

if h is bounded for which (109) is sufficient. By dominated convergence the second factor converges to $f_0(w)$ which is also bounded by (109). We conclude that the equicontinuity condition is satisfied if

$$\sup_{|t|\leq T_n}T_n\left|\frac{\hat{\phi}_x(t)-\phi_{x0}(t)}{\phi_{x^*}(t,\eta_n)}\right|=o_p(1) \tag{113}$$

Because the left-hand side is $o\left(\frac{T_n\alpha_n}{\eta_n}\right)$ almost surely, this is true under the same conditions that ensure uniform consistency of $\hat{g}_1(x-x^*)$ (see Lemma 6). The rest of the proof of stochastic equicontinuity is analogous, except that for some terms the conditions coincide with those of Lemma 7 that ensure uniform consistency of $\hat{g}_2(x^*, w)$. The correction term for the estimation of f_{x^*} involves an independent sample of size n_1. From the proof it is clear that stochastic equicontinuity holds of both sample sizes go to ∞ and that no assumption on the relative rate is needed.

By (109) the correction term for the fact that f_x is estimated is

$$\sqrt{n}\int_{\mathcal{X}}\int_{\mathcal{Y}}\int_{\mathcal{X}}\int_{\mathcal{W}} L_{xn}(\tilde{x}, y, x, w) f_0(y, x, w)\mathrm{d}y\mathrm{d}x\mathrm{d}w(\hat{f}_x(\tilde{x}) - f_{x0}(\tilde{x}))\mathrm{d}\tilde{x} = \tag{114}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_x(x_i) - \sqrt{n}\int_{\mathcal{Y}}\int_{\mathcal{X}}\int_{\mathcal{W}}\int_{\mathcal{X}} L_{xn}(\tilde{x}, y, x, w) f_{x0}(\tilde{x})\mathrm{d}\tilde{x} f_0(y, x, w)\mathrm{d}y\mathrm{d}x\mathrm{d}w$$

Now consider

$$\sqrt{n}\int_{\mathcal{X}} L_{xn}(\tilde{x}, y, x, w) f_{x0}(\tilde{x})\mathrm{d}\tilde{x} = \sqrt{n}\int_{\mathcal{X}^*} h(y, x, w, x^*)\left(g_{2n}(x^*, w)\int_{\mathcal{X}} K_{1xn}(\tilde{x}, x - x^*) f_{x0}(\tilde{x})\mathrm{d}\tilde{x} + \right.$$
$$\left. + g_{1n}(x - x^*)\int_{\mathcal{X}} K_{2xn}(\tilde{x}, x^*, w) f_{x0}(\tilde{x})\mathrm{d}\tilde{x}\right)\mathrm{d}x^* \tag{115}$$

Note

$$\int_{\mathcal{X}} K_{1xn}(\tilde{x}, x - x^*) f_{x0}(\tilde{x})\mathrm{d}\tilde{x} = g_{1n}(x - x^*)$$

and

$$\int_{\mathcal{X}} K_{2xn}(\tilde{x}, x^*, w) f_{x0}(\tilde{x})\mathrm{d}\tilde{x} = -\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^* - itw} I\left(|\phi_x(s)| > \frac{1}{2}\gamma_n\right)\frac{\phi_{xw}(s, t)\phi_{x^*}(s)\phi_x(s)}{\phi_x(s, \gamma_n)}\mathrm{d}s\mathrm{d}t$$

If we choose S_n as in Lemma 7, then the indicator function is 1 and $\phi_x(s, \gamma_n) = \phi_x(s)$ for $|s| \le S_n$. Hence

$$\int_{\mathcal{X}} K_{2xn}(\tilde{x}, x^*, w) f_{x0}(\tilde{x})\mathrm{d}\tilde{x} = -g_{2n}(x^*, w)$$

We conclude that for this choice of S_n (115) is 0 for all n. If we choose T_n as in Lemma 6 the corresponding expressions in the other correction terms also vanish. Note that the correction term for the estimation of f_{x^*} is a sum over the independent sample of size n_2.

We summarize the discussion in

Lemma 8 *If the assumptions of Theorem 1 hold and in addition*

(A4) $\lim_{n\to\infty} \frac{n}{n_1} = \lambda$, $0 < \lambda < \infty$, *and* $E(m(y,x,w,\theta_0,h_0)m(y,x,w,\theta_0,h_0)') < \infty$.

then

$$\left| \sqrt{n}\overline{m}_n(\theta_0,\hat{h}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m(y_i,x_i,w_i,h_0) + \delta_x(x_i) + \delta_{xw}(x_i,w_i)\right) - \frac{\sqrt{n}}{n_1}\sum_{i=1}^{n_1}\delta_{x^*}(x_i^*) \right| = o_p(1)$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m(y_i,x_i,,w_i,h_0) + \delta_x(x_i) + \delta_{xw}(x_i,w_i)\right) + \frac{\sqrt{n}}{n_1}\sum_{i=1}^{n_1}\delta_{x^*}(x_i^*) \xrightarrow{d} N(0,\Omega)$$

$$\Omega = E\left[\left(m(y,x,w,h_0) + \delta_x(x) + \delta_{xw}(x,w)\right)\left(m(y,x,w,h_0) + \delta_x(x) + \delta_{xw}(x,w)\right)'\right] + \lambda E\left[\delta_{x^*}(x^*)\delta_{x^*}(x^*)'\right]$$

and

$$\delta_x(\tilde{x}) = -\int_{\mathcal{X}^*}\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} f^*(y|x^*,w)s_{0n}(y|x,w)f_0(x,w).$$

$$.(K_{1xn}(\tilde{x},x-x^*)g_{2n}(x^*,w) + K_{2xn}(\tilde{x},x^*,w)g_{1n}(x-x^*))\,dw\,dx\,dy\,dx^*$$

$$K_{1xn}(\tilde{x},x-x^*) = \frac{1}{2\pi}\int_{-T_n}^{T_n}\frac{e^{-it(x-x^*)+it\tilde{x}}}{\phi_{x^*}(t,\eta_n)}\,dt$$

$$K_{2xn}(\tilde{x},x^*,w) = -\frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}} I\left(|\phi_x(s)| > \frac{1}{2}\gamma_n\right)\frac{\phi_{xw}(s,t)\phi_{x^*}(s)}{\phi_x(s,\gamma_n)^2}\,ds\,dt$$

and

$$\delta_{x^*}(\tilde{x}^*) = -\int_{\mathcal{X}^*}\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} f^*(y|x^*,w)s_{0n}(y|x,w)f_0(x,w).$$

$$.(K_{1x^*n}(\tilde{x}^*,x-x^*)g_{2n}(x^*,w) + K_{2x^*n}(\tilde{x}^*,x^*,w)g_{1n}(x-x^*))\,dw\,dx\,dy\,dx^*$$

$$K_{1x^*n}(\tilde{x},x-x^*) = -\frac{1}{2\pi}\int_{-T_n}^{T_n} I\left(|\phi_{x^*}(t)| > \frac{1}{2}\eta_n\right) e^{-it(x-x^*)+it\tilde{x}}\frac{\phi_x(t)}{\phi_{x^*}(t,\eta_n)^2}\,dt$$

$$K_{2x^*n}(\tilde{x},x^*,w) = \frac{1}{(2\pi)^2}\int_{-S_n}^{S_n}\int_{-T_n}^{T_n} e^{-isx^*-itw+is\tilde{x}}\frac{\phi_{xw}(s,t)}{\phi_x(t,\gamma_n)}\,ds\,dt$$

and

$$\delta_{xw}(\tilde{x},\tilde{w}) = -\int_{\mathcal{X}^*}\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{W}} f^*(y|x^*,w)s_{0n}(y|x,w)f_0(x,w).$$

$$.K_{2xwn}(\tilde{x},\tilde{w},x^*,w)g_{1n}(x-x^*))\,dw\,dx\,dy\,dx^*$$

$$K_{1xwn}(\tilde{x}, \tilde{w}, x - x^*) \equiv 0$$

$$K_{2xwn}(\tilde{x}, \tilde{w}, x^*, w) = \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^* - itw + is\tilde{x} + it\tilde{w}} \frac{\phi_{x^*}(s)}{\phi_x(s, \gamma_n)} ds\, dt$$

and

$$g_{1n}(x - x^*) = \frac{1}{2\pi} \int_{-T_n}^{T_n} e^{-it(x - x^*)} \frac{\phi_x(t)}{\phi_{x^*}(t, \eta_n)} dt$$

$$g_{2n}(x^*, w) = \frac{1}{(2\pi)^2} \int_{-S_n}^{S_n} \int_{-T_n}^{T_n} e^{-isx^* - itw} \frac{\phi_{xw}(s, t)\phi_{x^*}(s)}{\phi_x(s, \gamma_n)} ds\, dt$$

We have left the variance in a form that can be easily estimated. Some simplifications occur is we let $n, n_1 \to \infty$, but the resulting expressions are not so easily estimated.

Finally we have

Theorem 2 *If assumptions (A1)-(A4) are satisfied, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V) \tag{116}$$

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

$$M = E\left(\frac{\partial m(y, x, w, h_0)}{\partial \theta'}\right)$$

References

[1] Amemiya, Y. and Fuller, W.A., 1988, "Estimation for the nonlinear functional relationship," the Annals of Statistics, 16, pp. 147-160.

[2] Angrist, J., and A. Krueger, 1999, "Empirical strategies in labor economics," in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. 3A (North-Holland, Amsterdam) pp. 1277-1366.

[3] Bane, M.J. and D. Ellwood, 1994, "Understanding welfare dynamics," in Welfare Realities: From Rhetoric to Reform, eds. M.J. Bane and D. Ellwood. Cambridge: Harvard University Press.

[4] Barron, J.M., M.C. Berger and D.A. Black, 1997, On the Job Training (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

[5] Bekker, P.A., 1986, "Comment on identification in the linear errors in variables model," Econometrica, Vol. 54, No. 1, pp. 215-217.

[6] Blank, Rebecca, 1989, "Analyzing the length of welfare spells," Journal of Public Economics, 39(3), pp. 245-73.

[7] Blank, R. and P, Ruggles, 1994, "Short-term recidivism among public assistance recipients," American Economic Review 84 (May), pp. 49-53

[8] Blank, R. and P, Ruggles, 1996, "When do women use Aid to Families with Dependent Children and Food Stamps?" Journal of Human Resources 31 (Winter), pp. 57-89.

[9] Bound, J. Brown, C., Duncan, G. J., and Rodgers, W.L., 1989, "Evidence the validity of cross-sectional and longitudinal labor marker data," unpublished manuscript, University of Michigan.

[10] Bound, J. Brown, C., Duncan, G. J., and Rodgers, W.L., 1994, "Evidence the validity of cross-sectional and longitudinal labor marker data," Journal of Labor Economics 12, pp. 345-368.

[11] Bound, J., Griliches, Z. and Hall, B. 1986, "Wages, schooling and IQ of brothers and sisters: do the family factors differ?," International Economic Review, 27, pp. 77-105.

[12] Bound, J. C. Brown, G.J. Duncan and N. Mathiowetz, 2001, "Measurement error in survey data," in J.J. Heckman and E. Leamer eds., Handbook of Econometrics Vol 5.

[13] Bound, J. C. Brown, G.J. Duncan and W.L. Rodgers, 1989, "Measurement error in cross-sectional and longitudinal labor marker surveys: results from two validation studies," NBER Working Paper 2884.

[14] Bound, J., and Krueger, A.B., 1991, "The extent of measurement error in longitudinal earnings data: do two wrongs make a right?" Journal of Labor Economics, 9, pp. 1-24.

[15] Card, D., A. Hildreth, and L. Shore-Sheppard, 2001, "The measurement of Medicaid coverage in the SIPP: evidence from California, 1990-1996," NBER.

[16] Carroll, .J. and L.A. Stefanski, 1990, "Approximate quasi-likelihood estimation in models with surrogate predictors," Journal of the American Statistical Association 85, pp. 652-663.

[17] Carroll, R.J., D. Ruppert, and L.A. Stefanski, 1995, Measurement Error in Nonlinear Models. Chapman & Hall, New York.

[18] Carroll, R.J. and M.P. Wand, 1991, "Semiparametric estimation in logistic measurement error models," Journal of the Royal Statistical Society B 53, pp. 573-585.

[19] Chen, X, H. Hong, and E. Tamer, 2003, "Measurement Error Models with Auxiliary Data," unpublished manuscript.

[20] Code, J., 1992, "Using administrative record information to evaluate the quality of the income data collected in the survey of income and program participation," Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys (Statistics Canada, Ottawa) pp. 295-306.

[21] Cohen, S., and B. Carlson, 1994, "A comparison of household and medical provider reported expenditures in the 1987 NMES," Journal of Official Statistics 10, pp. 3-29.

[22] Csörgö, S., 1980, Empirical Characteristic Functions, Carleton mathematical lecture notes; no. 26.

[23] Dibbs, R., A. Hale, R. Loverock and S. Michaud, 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).

[24] Diggle, P.J. and Hall, P., 1993, "A Fourier approach to nonparametric deconvolution of a density estimate," Journal of Royal Statistical Society, series B 55, pp. 523-531.

[25] Ellwood, David, 1986, "Targeting 'Would Be' long-term recipients of AFDC," Princeton, NJ: Mathematica Policy Research.

[26] Feuerverger, A. and Mureika, R. A., 1977, The empirical characteristic function and its applications, Annals of Statistics, 5 pp. 88-97.

[27] Fitzgerald, J, 1995, "Local labor markets and local area effects on welfare duration," Journal of Policy Analysis and Management 14 (Winter), pp. 43-67.

[28] Fitzgerald, John, 1991, "Welfare durations and the marriage marker: evidence from the Survey of Income and Program Participation," Journal Human Resource, 26(3), pp. 545-61.

[29] Gini, C. 1921, "Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali," Metroeconomica, 1, pp. 63-82.

[30] Grondin, C., and S. Michaud, 1994, "Data quality of income data using computer-assisted interview: the experience of the Canadian survey of labour and income dynamics," Proceedings of the Survey Research Methods Section (American Statistical Association, Alexandria, VA) pp. 830-835.

[31] Hausman, J., Ichimura, H., Newey, W., and Powell, J., 1991, "Identification and estimation of polynomial errors-in-variables models," Journal of Econometrics, 50, pp. 273-295.

[32] Hausman, J.A., W.K. Newey, and J.L. Powell, 1995, "Nonlinear errors in variables: estimation of some Engel curves," Journal of Econometrics 65, pp. 205-233.

[33] Horowitz, Joel L. 1998, Semiparametric methods in econometrics. Springer.

[34] Horowitz, Joel L. and M. Markatou, 1996, "Semiparametric estimation of regression models for panel data," Review of Economic Studies 63, pp. 145-168.

[35] Hotz, V. J., R. George, J. Balzekas and F. Margolin, 1998, "Administrative data foe policy-relevant research: assessment of current utility and recommendations for development". Northwestern University/ University of Chicago Joint Center for Poverty Research.

[36] Hoynes, H., 2000, "Local labor markets and welfare spells: do demand conditions matter?" Review of Economics and Statistics 82 (August), pp. 351-368.

[37] Hoynes, H. and T. MaCurdy, 1994, "Has the decline in benefits shortened welfare spells?" American Economic Review 84 (May), pp. 43-48.

[38] Horowitz, J.L. and M. Markatou, 1996, "Semiparametric estimation of regression models for panel data" Review of Economic Studies 63, pp. 145-168.

[39] Hsiao, C., 1989, "Consistent estimation for some nonlinear errors-in-variables models," Journal of Econometrics, 41, pp. 159-185.

[40] Hsiao, C., 1991, "Identification and estimation of dichotomous latent variables models using panel data," Review of Economic Studies 58, pp. 717-731.

[41] Hsiao, C. and Q.K. Wang, 2000, "Estimation of structural nonlinear errors-in-variables models by simulated least-squares method," International Economic Review, Vol. 41, No. 2, pp. 523-542.

[42] Hu, Y, 2002, "Identification and estimation of the regression model with nonclassical measurement error using additional information," 2002, unpublished manuscript.

[43] Johnson, A., and M.E. Sanchez, 1993, "Household and medical provider reports on medical conditions: national medical expenditure survey, 1987," Journal of Economic and Social Measurement 19, pp. 199-223.

[44] Kane, T.J., C.E. Rouse and D. Staiger, 1999, "Estimating the returns to schooling when schooling is misreported," Working Paper 7235 (NBER).

[45] Lee, L.-F., and J.H. Sepanski, 1995, "Estimation of linear and nonlinear errors-in-variables models using validation data," Journal of the American Statistical Association, 90 (429).

[46] Lewbel, A., 1997, "Constructing instruments for regressions with measurement error when non additional data are available, with an application to Patents and R&D," Econometrica, 65(5), 1201-1213.

[47] Lewbel, A. 1998, "Semiparametric latent variable model estimation with endogenous or mismeasured regressors," Econometrica, 66, pp. 105-121.

[48] Li, T., and Q. Vuong, 1998, "Nonparametric estimation of the measurement error model using multiple indicators," Journal of Multivariate Analysis, 65, pp. 139-165.

[49] Li, T., 2002, "Robust and consistent estimation of nonlinear errors-in-variables models," Journal of Econometrics, 110, pp. 1-26.

[50] Lukacs, E, 1970, Characteristic Functions, 2nd ed., Griffin, London.

[51] Luenberger, D. G., 1969, Optimization by Vector Space Methods, Wiley, New York.

[52] Mathiowetz, N., and G. Duncan, 1988, "Out of work, out of mind: response errors in retrospective reports of unemployment," Journal of Business and Economic Statistics 6, pp. 221-229.

[53] Meyer, B. D., 1990, "Unemployment insurance and unemployment spells," Econometrica, Vol 58, No. 4, pp. 757-782.

[54] Moffitt, R. 1992. "Incentive effects of the U.S. welfare system: a review." Journal of Economic Literature 30(March), pp. 1-61.

[55] Moffitt, R. 2002, "The Temporary Assistance for Needy Families Program," NBER working paper 8749.

[56] Newey, W.K., 1994, "The asymptotic variance of semiparametric estimators," Econometrica 62, pp. 1349-1382.

[57] Newey, W.K., 2001, "Flexible simulated moment estimation of nonlinear errors-in-variables models," Review of Economics and Statistics, 83(4), pp. 616-627.

[58] O'Neill, June et al, 1984, "An analysis of time on welfare," Washington, DC: The Urban Institute.

[59] Pepe, M.S. and T.R. Fleming, 1991, "A general nonparametric method for dealing with errors in missing or surrogate covariate data," Journal of the American Statistical Association 86, pp. 108-113.

[60] Pollard, D., 1984, Convergence of Stochastic Processes, Springer, New York.

[61] Ridder, G. 1984, "The distribution of single-spell duration data," in G. R. Neumann and N. Westergard-Nielsen (eds), Studies in Labor Marker Analysis, Springer-Verlag, Berlin.

[62] Ridder, G. and R. Moffitt, 2003, "The econometrics of data combination," Submitted chapter for Handbook of Econometrics.

[63] Rodgers, W., and C. Brown and G. Duncan, 1993, "Errors in survey reports of earnings, hours worked, and hourly wages," Journal of the American Statistical Association 88, pp. 1208-1218.

[64] Schennach, S., 2000, "Estimation of nonlinear models with measurement error," Ph.D. thesis, MIT.

[65] Sepanski, J.H. and R.J. Carroll, 1993, "Semiparametric quasilikelihood and variance function estimation in measurement error models," Journal of Econometrics, 58, pp. 223-256.

[66] Serfling, R. J., 1980, Approximation Theorems of Mathematical Statistics, Wiley, New York.

[67] Van der Vaart, A.W., 1998, Asymptotic Statistics, Cambridge University Press.

[68] Wang, L., 1998, "Estimation of censored linear errors-in-variables models," Journal of Econometrics 84, pp. 383-400.

[69] Wansbeek, T. and E. Meijer, 2000, Measurement Error and Latent Variables in Econometrics, North Holland.