CAM

Centre for Applied Microeconometrics

Department of Economics
University of Copenhagen

http://www.econ.ku.dk/CAM/

Working paper no. 2011-01

Nonparametric Identification and Estimation of Transformation Models

Pierre-Andre Chiappori, Ivana Komunjer,
and Dennis Kristensen

The activities of CAM were originally financed by a grant from
The Danish National Research Foundation

NONPARAMETRIC IDENTIFICATION AND ESTIMATION OF TRANSFORMATION MODELS

PIERRE-ANDRE CHIAPPORI, IVANA KOMUNJER, AND DENNIS KRISTENSEN

ABSTRACT. This paper derives sufficient conditions for nonparametric transformation models to be identified and develops estimators of the identified components. Our nonparametric identification result is global, and is derived under conditions that are substantially weaker than full independence. In particular, we show that a completeness assumption combined with conditional independence with respect to one of the regressors suffices for the model to be identified. The identification result is also constructive in the sense that it yields explicit expressions of the functions of interest. We show how natural estimators can be developed from these expressions, and analyze their theoretical properties. Importantly, it is demonstrated that the proposed estimator of the unknown transformation function converges at the parametric rate.

Keywords: Nonparametric identification; transformation models; kernel estimation; $\sqrt{n}$-consistency.

Date: January 27, 2011.

Affiliations and Contact Information: Chiappori: Department of Economics, Columbia University. Email: pc2167@columbia.edu. Komunjer: Department of Economics, University of California San Diego. Email: komunjer@ucsd.edu. Kristensen: Department of Economics, Columbia University. Email: dk2313@columbia.edu

Earlier versions of this paper were presented at the CAM workshop 2010 at University of Copenhagen, the $(EC)^2$ conference 2010 in Toulouse, the conference on "Revealed Preferences and Partial Identification" in Montreal, and seminars at MIT, Princeton, and the University of Chicago. We thank all the participants for useful comments. Errors are ours.

1. Introduction

A variety of structural econometric models comes in form of a transformation model containing unknown functions. One important class are duration models that have been widely applied to study duration data in labor economics (Keifer, 1988), IO (Mata and Portugal, 1994), insurance (Abbring, Chiappori, and Zavadil, 2007), and finance (Engle, 2000; Lo, MacKinlay, and Zhang, 2002), among others. Another class are hedonic models studied by Ekeland, Heckman, and Nesheim (2004) and Heckman, Matzkin, and Nesheim (2005). A yet different example are models of binary choice in which the underlying random utilities à la Hausman and Wise (1978) are additively separable in the stochastic term as well as the unobserved attributes of the alternatives. Further examples of nonseparable econometric models that fall in the transformation model framework can be found in a survey by Matzkin (2007).

The present paper focuses on the following two questions. First, under what conditions is the transformation model nonparametrically identified? And second, how can we estimate the identified components from data? Regarding the first question, our main result is to show that transformation models are nonparametrically globally identified under conditions that are significantly weaker than full independence. Our identification strategy is constructive in a sense that we obtain explicit expressions of the relevant components of the model in terms of primitives such as the joint distribution of the observables. This in turn allows us to develop simple nonparametric estimators of the identified components which we analyze. This analysis leads to the second main result of the paper, which is to show that our nonparametric estimator of the transformation function attains parametric convergence rate. This in turn implies that for the estimation of the regression function, we can treat the transformation function as known.

We now discuss how our identification result relates to the existing literature. It is well-known that in nonparametric linear models $Y = g(X) + \epsilon$, the unknown function g can be identified from $E(\epsilon|Z) = 0$ w.p.1 if the conditional distribution of the endogenous regressor X given the instrument Z is *complete* (see Darolles, Florens,

and Renault, 2002; Blundell and Powell, 2003; Newey and Powell, 2003; Hall and Horowitz, 2005; Severini and Tripathi, 2006; d'Haultfoeuille, 2011, amongst others). Recently, Fève and Florens (2010) extended the completeness condition to identify (φ, β) in a semi-parametric transformation model $\varphi(Y) = W + \beta' X + \epsilon$, in which φ is strictly monotonic and $E[\epsilon|X, Z] = 0$. Using the inverse problems techniques, they established identifiability of the model without imposing any independence assumptions.

In this paper, we show that a similar completeness condition—when combined with conditional independence—is sufficient for identification of T, g and $F_{\epsilon|X}$ in a nonparametric transformation model $Y = T(g(X) + \epsilon)$, where T is strictly monotonic. Specifically, we work in a framework in which X can be decomposed into an exogenous subvector X_1 such that $\epsilon \perp X_1 \mid X_{-1}$, and an endogenous subvector X_{-1} whose conditional distribution given Z is complete. Our main assumption is that $E(\epsilon|Z) = 0$ w.p.1.

Even though the nonparametric transformation model is nonlinear in g and $F_{\epsilon|X}$, we obtain identification results that are global. We note that by letting $\theta \equiv (T, g)$ we can write the model as a special case of a nonlinear nonparametric instrumental variable model $E[\rho(Y, X, \theta)|Z] = 0$ w.p.1 where $\rho(Y, X, \theta) \equiv T^{-1}(Y) - g(X)$. For such models, Chernozhukov, Imbens, and Newey (2007) propose an extension of the completeness condition that guarantees θ to be locally nonparametrically identified. It is worth pointing out that their results are local in nature, and that nothing is being said about the identifiability of $F_{\epsilon|X}$.

Our identification results are close in spirit to those obtained by Ridder (1990), Ekeland, Heckman, and Nesheim (2004), and Jacho-Chávez, Lewbel, and Linton (2010). Using the independence of ϵ and X, Ridder (1990) establishes the nonparametric identifiability of (λ, g, F_ϵ) in a Generalized Accelerated Failure-Time (GAFT) model $\lambda(\tau) = g(X) + \epsilon$, where τ is the duration, $\lambda' > 0$ and F_ϵ is the distribution of the unobserved heterogeneity term ϵ. Letting $T \equiv \ln \circ \lambda^{-1}$, this result is related to that of Ekeland, Heckman, and Nesheim (2004) who show that assuming $\epsilon \perp X$

is sufficient to establish nonparametric identifiability (up to unknown constants) of T, g and F_ϵ in a nonparametric transformation model of the kind studied here.[1] A similar result has been obtained by Jacho-Chávez, Lewbel, and Linton (2010).

We extend the identification results of Ridder (1990), Ekeland, Heckman, and Nesheim (2004), and Jacho-Chávez, Lewbel, and Linton (2010) in two important directions: first, we prove nonparametric identification of the function T even when the regressor X contains an endogenous component; and second, we show that if there exists nonparametric instrumental variables Z such that the conditional distribution of X_{-1} given Z is complete, then the conditional moment conditions $E(\epsilon|Z) = 0$ w.p.1 are sufficient as well as necessary to identify g nonparametrically.[2] It is worth pointing out that our identification strategy allows to nonparametrically identify the transformation T even if the completeness assumption fails; the latter is only used to identify g and $F_{\epsilon|X}$.

The results of this paper are also related to the literature on nonparametric identification under monotonicity assumptions surveyed in Matzkin (2007). For example, Matzkin (2003) provides conditions under which in models of the form $Y = m(X, \epsilon)$ with m strictly monotone, the independence assumption $\epsilon \perp X$ is sufficient to globally identify m and F_ϵ (see also Chesher, 2003, for additional local results). In a

[1] In the same paper, the authors derive an additional result that relaxes the independence assumption and replaces it with $E(\epsilon|X) = 0$ w.p.1. They show that the latter is sufficient to identify general parametric specifications for $T(y, \phi)$ and $g(x, \theta)$ where ϕ and θ are finite dimensional parameters. Once $T(y, \phi)$ and $g(x, \theta)$ are specified, the results derived by Komunjer (2008) can be used to further check whether global GMM identification of ϕ and θ holds.

[2] The results also extend those of Hoderlein (2009) who considers identification and estimation of semiparametric endogenous binary choice models in which $T(X) = \beta' X$. As shown in Hoderlein (2009), the slope parameter β can then be identified as the mean ratio of derivatives of two functions of the instrument Z.

sense, our result shows that the independence condition can be substantially relaxed, if a certain form of separability between Y, X and ϵ holds, namely, if we have $T^{-1}(Y) = g(X) + \epsilon$.[3]

Our estimation strategy for T is closely related to the work of Horowitz (1996) who shows that in the special case where $g(X) = \beta' X$, the transformation function T is identified as an integral function over relevant derivatives of the cumulative distribution function (cdf) of Y given X. Horowitz (1996) then uses this result to develop $\sqrt{n}$-consistent, asymptotically normal, nonparametric estimators of T and F_ϵ when $g(X) = \beta' X$.[4] We obtain a similar expression of T in the general nonparametric case, which allows us to develop natural two-step estimator: First, we obtain nonparametric kernel estimators of the conditional cdf ; second, plugging this estimator into the expression of T as a functional of this cdf, an estimator of T is obtained. The resulting estimator of T involves integrating over (some transformation of) the conditional cdf, and this integration leads to parametric convergence rates of the estimator despite the fact that the first-step estimator converges with nonparametric rate akin to two-step semiparametric estimators (see, e.g., Newey and McFadden, 1994).

Once T has been estimated, estimation of the regression function $g(x)$ can be done by nonparametric IV with $\hat{T}^{-1}(Y)$ replacing the true but unknown dependent variable, $T^{-1}(Y)$. Specifically, we adjust the estimator of Blundell, Chen, and Kristensen (2007) to allow for pre-estimated dependent variables thereby yielding a feasible estimator of $g(x)$. Given the parametric convergence rate of $\hat{T}$, our nonparametric IV estimator of $g(x)$ converges with the same rate as if we knew T and as such we suffer no loss of efficiency from T being unknown.

In the context of semi-parametric transformation models, $\varphi(Y) = W + \beta' X + \epsilon$, Fève and Florens (2010) proposed a sequential approach to estimate φ and β

[3]See also the discussion on page 24 in Blundell and Powell (2003).

[4]Estimators of β have been available since Han (1987).

using regularization. Their estimator is based on the conditional moment restriction $E[\epsilon|X,Z]=0$ alone, however its rate of convergence is slower than $\sqrt{n}$.

In the special case where the transformation T is finitely parameterized, Linton, Sperlich, and van Keilegom (2008) construct a mean square distance from independence estimator for the transformation parameter. Jacho-Chávez, Lewbel, and Linton (2010) have developed alternative, fully nonparametric estimators of the transformation model considered, but these estimators of the transformation function do not obtain parametric rate, and do not allow for endogeneous regressors. Finally, it is worth pointing out that the general sieve estimation methods developed in Ai and Chen (2003) and Chernozhukov, Imbens, and Newey (2007) should in principle be applicable to the transformation model yielding consistent estimators for (T,g) simultaneously. However, a full theoretical analysis of these general estimators in the case of the transformation model has not been made and it is unclear whether the nonparametric components of the sieve estimators will attain parametric rate.

The remainder of the paper is organized as follows. Section 2 introduces the transformation model and recalls basic definitions. In Section 3, we derive necessary and sufficient conditions for the model to be nonparametrically identified. Our identification strategy is constructive in a sense that it leads to a natural estimator for T; identification of g and $F_{\epsilon|X}$ is derived once the transformation is known. In Section 4 we propose estimators of T, g and $F_{\epsilon|X}$ and analyze their asymptotic properties. The last section concludes. All of our proofs are relegated to an Appendix.

2. Model

We start by introducing the model and the assumptions. We consider a nonparametric transformation model of the form

$$Y = T\left(g(X)+\epsilon\right), \tag{1}$$

where Y belongs to $\mathcal{Y}\subseteq\mathbb{R}$, $X=(X_1,\ldots,X_{d_x})$ belongs to $\mathcal{X}\subseteq\mathbb{R}^{d_x}$, and ϵ is in $\mathcal{E}\subseteq\mathbb{R}$. The variables Y and X are observed, while ϵ remains latent. The

transformation T and the regression function g in (1) are unknown real functions; additional restrictions on T and g will be imposed below.

Hereafter we maintain the following assumptions.

Assumption A1. *Let $\mathcal{X}$ be the support of X.*[5] *Then, for a.e. $x \in \mathcal{X}$, the conditional distribution $F_{\epsilon|X}(\cdot, x)$ of ϵ given $X = x$ is absolutely continuous with a continuous density $f_{\epsilon|X}(\cdot, x)$.*

Assumption A1 states that for almost every realization $x \in \mathcal{X}$ of X, the conditional density of ϵ given $X = x$ exists and is continuous. Let $\mathcal{E}_x \subseteq \mathbb{R}$ denote the support of ϵ given $X = x$; then, $\int_{\mathcal{E}_x} f_{\epsilon|X}(t, x)dt = 1$ and $f_{\epsilon|X}(\cdot, x) > 0$ on $\mathcal{E}_x$. In particular, Assumption A1 implies that the random variable $\delta \equiv g(X) + \epsilon$ is continuously distributed with density $f_\delta(d) = \int_{\mathcal{X}} f_{\epsilon|X}(d - g(x), x)dF(x)$ where $F(\cdot)$ denotes the cdf of X. The following assumption ensures that the support of δ is a connected subset of $\mathbb{R}$ (i.e. an interval).

Assumption A2. *The support $\mathcal{D}$ of $g(X) + \epsilon$ is connected in $\mathbb{R}$.*

Put differently, Assumption A2 requires that the closure of the set $\{d \in \mathbb{R} : f_\delta(d) > 0\}$ be connected in $\mathbb{R}$. For example, this excludes the situations in which X is a scalar binary variable, and the supports $\mathcal{E}_0$ and $\mathcal{E}_1$ of ϵ given X are disjoint intervals. We are now ready to put further restrictions on the transformation $T : \mathcal{D} \to \mathbb{R}$ in (1).

Assumption A3. *T is continuously differentiable on $\mathcal{D}$, $T'(d) > 0$ for every $d \in \mathcal{D}$, and $0 \in \mathcal{Y} = T(\mathcal{D})$.*

We restrict our attention to the transformations T in (1) that are smooth and strictly increasing from $\mathcal{D}$ onto $\mathcal{Y}$. Without loss of generality, we assume that $0 \in T(\mathcal{D})$, i.e. 0 belongs to the support of Y. Assumptions A1, A2 and A3 guarantee that the conditional distribution $F_{Y|X}(\cdot, x)$ of Y given $X = x$ is absolutely continuous

[5]Following the usual convention, the support of a random variable is defined as the smallest closed set whose complement has probability zero.

with a continuous density $f_{Y|X}(\cdot, x)$. Moreover, the support $\mathcal{Y}$ of Y is a connected subset of $\mathbb{R}$.

We now further restrict the dependence between ϵ and X. For this, let X_1 denote the first component of X; whenever $d_x > 1$, we denote by X_{-1} the remaining subvector of X, i.e. $X_{-1} \equiv (X_2, \ldots, X_{d_x})$. The supports of X_1 and X_{-1} are denoted $\mathcal{X}_1$ and $\mathcal{X}_{-1}$, respectively. We make the following assumption:

Assumption A4. $\epsilon \perp X_1 \mid X_{-1}$.

Assumption A4 states that ϵ is independent of at least one component of X, given the remaining components of X; we may, with no loss of generality, assume that this conditionally exogenous component is X_1. Put in words, the property in A4 says that the variable X_1 is excluded from the conditional distribution of ϵ given X. This is why we call exclusion restriction the conditional independence assumption in A4.

Assumption A5. *The random variable X_1 is continuously distributed on $\mathcal{X}_1 \subseteq \mathbb{R}$.*

According to Assumption A5, the first component X_1 of each observable vector X is continuous. Note that except for the continuity of the random variable X_1, A5 does not restrict its support $\mathcal{X}_1$. In particular, $\mathcal{X}_1$ need not be equal to $\mathbb{R}$, and X_1 may well have bounded support. Perhaps more importantly, assumption A5 allows all the other components $X_2, \ldots, X_{d_x}$ to be either continuous or discrete with bounded or unbounded supports. We now further restrict the regression function $g : \mathcal{X} \to \mathbb{R}$ in (1).

Assumption A6. *For a.e. $x \in \mathcal{X}$, the partial derivative $\partial g(x)/\partial x_1$ exists.*

Similar to A5, Assumption A6 only restricts the behavior of the partial derivative of g with respect to x_1. Nothing is being said about the behavior of g with respect to the remaining components x_{-1}.

In addition to the restrictions on the joint distribution of ϵ and X_1 conditional on X_{-1} stated in Assumption A4, we now restrict the joint distribution of ϵ and X_{-1}.

For this, we shall assume that there exists a vector of instruments $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ with respect to which the distribution of X_{-1} is complete, and such that ϵ is mean independent of Z.

Assumption A7. *For a.e. $z \in \mathcal{Z}$, $E(\epsilon|Z = z) = 0$ and the conditional distribution of X_{-1} given $Z = z$ is complete: for every function $h : \mathcal{X}_{-1} \mapsto \mathbb{R}$ such that $E[h(X_{-1})]$ exists and is finite, $E[h(X_{-1}) \mid Z = z] = 0$ implies $h(x_{-1}) = 0$ for a.e. $x_{-1} \in \mathcal{X}_{-1}$.*

Recall from A4 that ϵ is assumed to be conditionally independent of X_1 given X_{-1}, i.e. the first component of X is conditionally exogenous. The other components are on the other hand allowed to be endogenous provided the completeness condition in A7 holds.[6]

3. Identification

Following the related literature (e.g., Koopmans and Reiersøl, 1950; Brown, 1983; Roehrig, 1988; Matzkin, 2003) we hereafter call *structure* a particular value of the triplet $(T, g, F_{\epsilon|X})$ in Equation (1), where $T : \mathcal{D} \mapsto \mathbb{R}$, $g : \mathcal{X} \mapsto \mathbb{R}$, and $F_{\epsilon|X} : \mathbb{R} \times \mathcal{X} \mapsto \mathbb{R}$. The model then simply corresponds to the set of all structures $(T, g, F_{\epsilon|X})$ that satisfy the restrictions given by Assumptions A1 through A7. Each structure in the model induces a conditional distribution $F_{Y|X}$ of the observables, and two structures $(\tilde{T}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ and $(T, g, F_{\epsilon|X})$ are observationally equivalent if they generate the same $F_{Y|X}$.

[6]Further discussion of the completeness condition can be found in Darolles, Florens, and Renault (2002), Blundell and Powell (2003), Newey and Powell (2003), Hall and Horowitz (2005), Severini and Tripathi (2006), and d'Haultfoeuille (2011), among others. For example, it is equivalent to requiring that for every function $h : \mathcal{X}_{-1} \to \mathbb{R}$ such that $E[h(X_{-1})] = 0$ and $\text{var}[h(X_{-1})] > 0$, there exists a function $k : \mathcal{Z} \to \mathbb{R}$ such that $E[h(X_{-1})k(Z)] \neq 0$ (see Lemma 2.1. in Severini and Tripathi, 2006).

We now address the identification problem, namely: If $(T, g, F_{\epsilon|X})$ is a structure that generates $F_{Y|X}$, is it possible to find an alternative structure that is different from but observationally equivalent to $(T, g, F_{\epsilon|X})$? More formally, the structure $(T, g, F_{\epsilon|X})$ is globally identified if any observationally equivalent structure $(\tilde{T}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ satisfies: for every $t \in \mathbb{R}$, every $y \in \mathcal{Y}$, and a.e. $x \in \mathcal{X}$,

$$\tilde{\Theta}(y) = \Theta(y), \quad \tilde{g}(x) = g(x), \quad \text{and} \quad \tilde{F}_{\tilde{\epsilon}|X}(t, x) = F_{\epsilon|X}(t, x),$$

where we have let $\Theta : \mathcal{Y} \to \mathbb{R}$ denote the inverse mapping T^{-1},

$$\Theta(y) = T^{-1}(y). \tag{2}$$

The conditional independence property in Assumption A4 has strong implications which we now derive. In what follows, let $\Phi(y, x)$ denote the conditional cdf of Y given X,

$$\Phi(y, x) \equiv F_{Y|X}(y, x) = P(Y \leq y | X = x).$$

Under Assumption A3, Θ is continuously differentiable and strictly increasing on $\mathcal{Y}$. Note that in addition $\Theta(\mathcal{Y}) = \mathcal{D}$. Equation (1) is equivalent to $\epsilon = \Theta(Y) - g(X)$, so by $\Theta' > 0$ and the conditional independence of ϵ and X_1 given X_{-1},

$$\Phi(y, x) = P(\epsilon \leq \Theta(y) - g(x) | X = x) = F_{\epsilon|X}(\Theta(y) - g(x), x_{-1}), \tag{3}$$

for all $(y, x) \in \mathcal{Y} \times \mathcal{X}$. The identification problem can then be restated as follows: Given Φ, to what extent is it possible to recover the functions Θ, g and $F_{\epsilon|X}$ which for every $y \in \mathcal{Y}$ and a.e. $x \in \mathcal{X}$ satisfy Equation (3)?

For one thing, it is clear from Equation (1) that some normalization of the model is needed; indeed, for any $\lambda > 0$ and $\mu \in \mathbb{R}$, the transformation model (1) is equivalent to $Y = \tilde{T}(\lambda g(X) + \mu + \lambda\epsilon)$ where $\tilde{T}$ is defined by $\tilde{T}(t) \equiv T((t - \mu)/\lambda)$. We therefore impose that any structure $(T, g, F_{\epsilon|X})$ in (1) satisfies the normalization condition:

$$T(0) = 0 \text{ and } E(\epsilon) = 0, \; E[g(X)] = 1. \tag{4}$$

The main identification result is as follows:

Theorem 1. *Let Assumptions A1 through A6 and the normalization condition* (4) *hold. Assume in addition that the set* $A \equiv \{x \in \mathcal{X} : \partial\Phi(y,x)/\partial x_1 \neq 0$ *for every* $y \in \mathcal{Y}\}$ *is nonempty. Then:*

(i) T *is globally identified;*

(ii) $(g, F_{\epsilon|X})$ *are globally identified if and only if Assumption A7 holds.*

The first part of Theorem 1 shows that under Assumptions A1-A6 and the additional condition on the set A, the transformation T is globally identified. The requirement that A has nonempty interior can be thought of as a generalized rank condition saying that X_1 has a causal impact on Y. The intuition behind this condition appears by taking derivatives w.r.t. x_1 in Equation (3),

$$\frac{\partial\Phi(y,x)}{\partial x_1} = -f_{\epsilon|X}\left(\Theta\left(y\right) - g(x), x_{-1}\right)\frac{\partial g(x)}{\partial x_1}.$$

Thus, the requirement has two parts: First, we need that for some $x \in \mathcal{X}$, $\partial g(x)/\partial x_1 \neq 0$; this requirement excludes the situations in which g is a constant function. Second, we need that for the same value x, $\{t \in \mathbb{R} : t = \Theta(y) - g(x), y \in \mathcal{Y}\} \subseteq \mathcal{E}_x$; this assumption ensures that $f_{\epsilon|X}\left(\Theta\left(y\right) - g(x), x_{-1}\right) > 0$ for every $y \in \mathcal{Y}$, and is akin to Assumption 5a in Horowitz (1996). A simple primitive condition for the second requirement is that $\mathcal{E}_x = \mathbb{R}$, for example. Rather than imposing specific sufficient conditions on $f_{\epsilon|X}$ and g, we maintain the high-level condition that A is nonempty.

The second part of Theorem 1 states that the completeness condition A7 is both sufficient *and necessary* to nonparametrically identify the regression function g and the distribution $F_{\epsilon|X}$. Note that the result is global even though the model (1) is nonlinear in g and $F_{\epsilon|X}$.

While necessary to identify $(g, F_{\epsilon|X})$, the completeness assumption A7 is not used to identify the transformation T. In fact, the proof of Theorem 1 shows that under A1-A6 and the normalization condition (4), the inverse transformation $\Theta = T^{-1}$ can

be written as:

$$\Theta(y) = \frac{\int_0^y \Phi_y(u,x)[\Phi_1(u,\bar{x})]^{-1}du}{\int_{\mathbb{R}} f_Y(y) \int_0^y \Phi_y(u,x)[\Phi_1(u,x)]^{-1}dudy}, \quad x \in A. \tag{5}$$

Here, $f_Y(y)$ denotes the unconditional density of Y, and we use subscripts to denote partial derivatives.[7] Key to the identification of T is the conditional independence assumption (A4) which in particular guarantees that the right hand side in Equation (5) is not a function of x. Hence, evaluating this quantity at any x for which $\Phi_1(y,x)$ never vanishes allows to recover Θ. The expression in (5) also makes clear why Theorem 1 needs to assume the set A of such x's to be nonempty.

It is worth pointing out that the case of several conditionally exogenous variables is a particular version of the setting above. Indeed, assume that the disturbance ϵ in the model (1) is known to be conditionally independent of X_i $(1 \leq i \leq I)$ given the remaining components of X. Since $E(\epsilon) = 0$, it then holds that w.p.1 $E(\epsilon|X_i) = 0$. Hence, it suffices to include X_i in the vector of instruments Z.

As Equation (5) shows, our identification strategy is *constructive* in a sense that it leads for a closed form expression of $\Theta = T^{-1}$ as a function of the observables. In the next section, we develop nonparametric estimators of Θ and g that builds on this expression, and examine their properties.

4. Estimation

We use the identification strategy of the previous section to derive explicit estimators of $(T, g, F_{\epsilon|X})$.

Suppose we have a random sample (Y_i, X_i, Z_i) $(i = 1, \ldots, n)$ drawn from the transformation model in Equation (1) and that Assumptions A1 to A7 hold. We study the estimation of each of the identified components of the model in turn: First, we propose an estimator of the inverse transformation function $\Theta = T^{-1}$ under the normalization (4). Next, given this estimator, we proceed to estimate the regression function g and the conditional cdf of the error term $F_{\epsilon|X}$.

[7]Specifically, $g_1(x) \equiv \frac{\partial g(x)}{\partial x_1}$, $\Phi_y(y,x) \equiv \frac{\partial \Phi(y,x)}{\partial y}$ and $\Phi_1(y,x) \equiv \frac{\partial \Phi(y,x)}{\partial x_1}$.

To develop our estimator of Θ, we rewrite the expression in Equation (5) as:

$$\Theta(y) = \frac{S(y,x)}{E[S(Y,x)]}, \tag{6}$$

where we have defined

$$S(y,x) \equiv \int_0^y \frac{\Phi_y(u,x)}{\Phi_1(u,x)} du \quad \text{and} \quad E[S(Y,x)] = \int_{\mathbb{R}} S(y,x) f_Y(y) dy. \tag{7}$$

The estimation method that we propose is straightforward in principle: We first obtain a nonparametric estimator of the conditional cdf $\Phi(y,x)$. We then plug this estimator into Equation (7) to obtain an estimator of S and its moment which in turn are substituted into Equation (6). This yields a nonparametric estimator of $\Theta(y)$.

To be more specific, observe that the conditional cdf can be written as

$$\Phi(y,x) = \frac{p(y,x)}{f(x)}, \quad p(y,x) \equiv \int_{-\infty}^y f_{Y,X}(u,x) du, \quad f(x) \equiv \int_{\mathbb{R}} f_{Y,X}(u,x) du,$$

where $f_{Y,X}(y,x)$ is the joint pdf of (Y,X). Thus, a natural kernel-based estimator of $\Phi(y,x)$ is

$$\hat{\Phi}(y,x) = \frac{\hat{p}(y,x)}{\hat{f}(x)}, \tag{8}$$

where

$$\hat{p}(y,x) = \frac{1}{n}\sum_{i=1}^n \mathcal{K}_{h_y}(Y_i - y)\mathbf{K}_{h_x}(X_i - x), \quad \hat{f}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_{h_x}(X_i - x),$$

with $\mathcal{K}_{h_y}(y) = \mathcal{K}(y/h_y)/h_y$, $\mathbf{K}_{h_x}(x) = \mathbf{K}(x/h_x)/h_x^{d_x}$ and $h_x, h_y > 0$ being univariate bandwidths. The functions $\mathcal{K}(y)$ and $\mathbf{K}(x)$ are given as $\mathcal{K}(y) = \int_{-\infty}^y K(u) du$ and $\mathbf{K}(x) = \prod_{i=1}^{d_x} K(x_i)$ with $K : \mathbb{R} \mapsto \mathbb{R}$ being a univariate kernel. Note that we could allow for individual bandwidths for each variable in X_i but to keep the notation simple we here use a common bandwidth across all regressors. Also note that we could replace $\mathcal{K}_{h_y}(Y_i - y)$ with the indicator function $\mathbb{I}\{Y_i \leqslant y\}$ if we were only interested in estimating $\Phi(y,x)$ itself, but since we also need to estimate its derivatives we here employ the above estimator since it is differentiable.

With this estimator of the conditional cdf at hand, we propose to estimate $\Theta(y)$ by

$$\hat{\Theta}(y) = \int_{\mathcal{X}} w(x) \frac{\hat{S}(y,x)}{\hat{E}[\hat{S}(Y,x)]} dx,$$

where $w(x)$ is a weighting function with compact support $\mathcal{X}_0 \subseteq \mathcal{X}$ satisfying

$$\int_{\mathcal{X}_0} w(x)\, dx = 1,$$

and

$$\hat{S}(y,x) \equiv \int_0^y \frac{\hat{\Phi}_y(u,x)}{\hat{\Phi}_1(u,x)} du \quad \text{and} \quad \hat{E}[\hat{S}(Y,x)] \equiv \frac{1}{n}\sum_{i=1}^n \hat{S}(Y_i,x).$$

The weighting function w serves to purposes: First, it is used to control for the usual denominator problem present in many semiparametric estimators where we divide by a nonparametric density estimator. In particular, we will require that $\inf_{x\in\mathcal{X}_0} f(x) > 0$ and $\inf_{y\in\mathcal{Y}_0, x\in\mathcal{X}_0} \Phi_1(y,x) > 0$. Second, it allows us to improve on efficiency of the estimator by reweighting $\hat{S}(y,x)/\hat{E}[\hat{S}(Y,x)]$ as a function of x. There is a tension between these two purposes since in order to obtain full efficiency, we expect that w needs to have full support which is ruled about by compact support assumption. It should be possible to weaken this restriction though and allow the support $\mathcal{X}_0$ to grow with sample size. This will however lead to more complicated conditions and proofs and so we maintain the compact support assumption for simplicity.

We note that the proposed estimator is similar to the estimator of Horowitz (1996) who considers the semiparametric model where the regression function is restricted to $g(x) = \beta' x$. Assuming for simplicity that β is known such that $V \equiv \beta' X$ is observed, the estimator of Horowitz (1996) can be written as:

$$\tilde{\Theta}(y) = -\int_{\mathcal{V}} \omega(v) \int_0^y \frac{\hat{G}_y(u,v)}{\hat{G}_1(u,v)} du dv,$$

where $\hat{G}(y,v)$ is a kernel estimator of $G(y,v) = P(Y \le y | V = v) = F_{Y|V}(y,v)$, $\mathcal{V}$ is the support of V, and $\omega(v)$ is a weighting function with compact support. As shown in Horowitz (1996), due to the double-integration, $\tilde{\Theta}(y)$ is $\sqrt{n}$-consistent despite the fact that it relies on first-step nonparametric estimators that converge with slower

rate. Similarly, in our case we also integrate over both y and x, and as such we expect that our proposed estimator $\hat{\Theta}(y)$ will be $\sqrt{n}$-consistent.

Once $\hat{\Theta}(y)$ has been obtained, the regression function and the conditional cdf of the error term can be estimated using nonparametric IV techniques: First, suppose that $\Theta(y)$ is known. Then, $\Theta(Y) = g(X) + \epsilon$ with $E[\epsilon|X_1, Z] = 0$ so the estimation of g is a standard nonparametric IV regression problem. We can thus import techniques from that part of the literature such as the kernel estimator of Hall and Horowitz (2005) or the sieve estimator of Blundell, Chen, and Kristensen (2007). In this paper, we focus on the sieve estimator for $g(x)$ proposed in Blundell, Chen, and Kristensen (2007), which takes the following form when $\Theta(y)$ is known:

$$\tilde{g} = \arg\min_{g_n \in \mathcal{G}_n} \sum_{i=1}^{n} \{\tilde{h}(X_{1,i}, Z_i) - \hat{M}(X_{1,i}, Z_i|g_n)\}^2, \tag{9}$$

where $\tilde{h}(x_1, z)$ and $\hat{M}(x_1, z|g_n)$ are first-step nonparametric estimators (such as a kernel regression or a series estimators) of

(10)

$$h(x_1, z) \equiv E[\Theta(Y)|X_1 = x_1, Z = z], \text{ and } M(x_1, z|g_n) \equiv E[g_n(X)|X_1 = x_1, Z = z],$$

and $\mathcal{G}_n$ is a sieve space. We have here left out the weighting function used in Blundell, Chen, and Kristensen (2007) since this is only used to obtain efficiency of the parametric component of their model. With $\Theta(y)$ unknown, we propose to modify the above nonparametric sieve IV estimator with the true but unknown dependent variable, $\Theta(Y)$, being replaced by generated ones, $\hat{\Theta}(Y)$. This leads to the following feasible version of the above sieve estimator:

$$\hat{g} = \arg\min_{g_n \in \mathcal{G}_n} \sum_{i=1}^{n} \{\hat{h}(X_{1,i}, Z_i) - \hat{M}(X_{1,i}, Z_i|g_n)\}^2, \tag{11}$$

where $\hat{h}(x_1, z)$ is an estimator of $E[\hat{\Theta}(Y)|X_1 = x_1, Z = z]$; that is, the unknown function Θ is replaced by its estimator $\hat{\Theta}$.

Finally, given $\hat{\Theta}(y)$ and $\hat{g}(x)$, we can compute the corresponding residuals, $\hat{\epsilon}_i = \hat{\Theta}(Y_i) - \hat{g}(X_i)$, $i = 1, ..., n$. Standard nonparametric estimators of conditional cdf's,

such as the kernel one presented above, can now be employed with the residuals replacing the actual, unobserved errors,

$$\hat{F}_{\epsilon|X}(t, x_{-1}) = \frac{\sum_{i=1}^{n} \mathcal{K}_{h_y}(\hat{\epsilon}_i - t) \mathbf{K}_{h_{x_{-1}}}(X_{-1,i} - x_{-1})}{\sum_{i=1}^{n} \mathbf{K}_{h_{x_{-1}}}(X_{-1,i} - x_{-1})},$$

We now proceed to analyze the asymptotic properties of the estimators. In order to do so, we introduce additional assumptions on the model and the kernel function used in the estimation. The kernel K used to define our estimator of $\Theta(y)$ is assumed to belong to the following class of kernel function:

Assumption A8. *The univariate kernel K is differentiable, and there exists constants $C, \eta > 0$ such that*

$$\left|K^{(i)}(z)\right| \leq C|z|^{-\eta}, \quad \left|K^{(i)}(z) - K^{(i)}(z')\right| \leq C|z - z'|, \quad i = 0, 1,$$

where $K^{(i)}(z)$ denotes the ith derivative. Furthermore, $\int_{\mathbb{R}} K(z)\,dz = 1$, $\int_{\mathbb{R}} z^j K(z)\,dz = 0$, $1 \leq j \leq m-1$, and $\int_{\mathbb{R}} |z|^m K(z)\,dz < \infty$.

The above class is fairly general and accommodate kernels with both bounded and unbounded support. We do however require the kernel K to be differentiable which rules out uniform and Epanechnikov kernels. This is however only used for technical reasons, and we expect the following results to also hold for non-differentiable kernels. We allow for both standard second-order kernels ($m = 2$) such as the Gaussian one, and higher-order kernel ($m > 2$). The use of higher-order kernels in conjunction with smoothness conditions on the densities in the model allow us to control for the smoothing bias induced by the use of kernels. In general, the kernel has to be of higher order, in order for $\hat{\Theta}(y)$ to be $\sqrt{n}$-consistent.

The smoothness conditions that we will impose on the density of data are as follows:

Assumption A9. *The joint density, $f_{Y,X}(y, x)$ is bounded, m times differentiable w.r.t. (y, x) with bounded derivatives; its mth order partial derivatives are uniformly*

continuous. Furthermore, $\sup_{x\in\mathcal{X},y\in\mathcal{Y}} \|(x,y)\|^b f_{Y,X}(y,x) < \infty$ *for some constant* $b > 0$.

Note that the number of derivatives, $m \geq 2$, is assumed to match up with the order of the kernel K. The requirement that $\sup_{x,y} \|(x,y)\|^b f_{Y,X}(y,x) < \infty$ is implied by $E[|Y|^b] < \infty$ and $E[\|X\|^b] < \infty$.

As noted earlier the weighting function is used to control the denominator problem of our estimator. More specifically, with $\mathcal{X}_0$ denoting the support of w, we require that:

Assumption A10. *The following bounds hold:* $\inf_{y\in\mathcal{Y},x\in\mathcal{X}_0} \Phi_1(y,x) > 0$, $\inf_{x\in\mathcal{X}_0} f(x) > 0$ *and* $\sup_{y\in\mathcal{Y}} |\Theta(y)| < \infty$.

The lower bound condition on $\Phi_1(y,x)$ is related to the set A introduced in Theorem 1 and further restricts the behavior of $\Phi_1(y,x)$. In particular, Assumption A10 implies that $\mathcal{X}_0 \subseteq A$ and so A has non-empty interior. If in fact A has empty interior, we cannot obtain $\sqrt{n}$-consistency.

The lower bounds imposed on $\Phi_1(y,x)$ tand $f(x)$ allows us to control the estimation error $\hat{S}(y,x) - S(y,x)$ uniformly over $(y,x) \in \mathcal{Y}\times\mathcal{X}_0$. The above condition implicitly restricts the support of the weighting function to be compact, and Y to have bounded support. We conjecture that the assumption could be weakened to $\inf_{y\in\mathcal{Y}_0,x\in\mathcal{X}_0} \Phi_1(y,x) > 0$ for some (possibly bounded) interval $\mathcal{Y}_0 \subseteq \mathcal{Y}$, thereby allowing for unbounded support of Y. However, this would come at the price of having to introduce trimming in the definition of our estimator $\hat{E}[\hat{S}(Y,x)]$. To avoid more complicated estimators and proofs, we therefore maintain the above stronger assumption.

Finally, we impose the following restrictions on the the rate with which the bandwidth sequences $h_x = h_{x,n}$ and $h_y = h_{y,n}$ are allowed to shrink towards zero as $n \to \infty$:

Assumption A11. $\sqrt{n}h_x^{2m} \to 0$, $\sqrt{n}h_y^{2m} \to 0$, $\sqrt{n}h_x^{d_x+2}/\log(n) \to \infty$, $\sqrt{n}h_y h_x^{d_x+1}/\log(n) \to \infty$.

Assumption A11 puts restrictions on the two bandwidths sequences ensuring that the squared estimation error of the kernel estimators $\hat{p}(y,x)$ and $\hat{f}(x)$ and their relevant derivatives all are of order $o_P(1/\sqrt{n})$ uniformly over $y \in \mathcal{Y}_0$ and $x \in \mathcal{X}_0$. As is standard for kernel estimators, there is a curse-of-dimensionality which appears in the last two restrictions on h_x: When the dimension of X, $d_x \geq 1$, is large, we in general need to use higher-order kernels in order for all four conditions to hold simultaneously. For example, if $h_x \propto n^{-r_x}$ and $h_y \propto n^{-r_y}$ then Assumption A11 holds whenever

$$m > \frac{d_x + 2}{2} \quad \text{and} \quad \frac{1}{4m} < r_x, r_y < \frac{1}{2(d_x+2)}.$$

To state the asymptotic distribution of the estimator, we collect data in $U_i = (Y_i, X_i)$ and introduce the function $\delta^w(U_i|y)$ given by

$$\delta^w(U_i|y) \equiv \psi_i^{\bar{w}_1}(U_i|y) - \varphi^{\bar{w}_2}(U_i), \tag{12}$$

with $\bar{w}_1(x) \equiv w(x)/E[S(Y,x)]$, $\bar{w}_2(x) \equiv w(x)/E[S(Y,x)]^2$ and

$$\begin{aligned} \psi^{\bar{w}}(U_i|y) \equiv\ & \bar{w}(X_i)\int_{Y_i}^{y} D_{p,0}(u,X_i)\,du + \mathbb{I}\{Y_i \leq y\}\,\bar{w}(X_i)\,D_{p,y}(Y_i,X_i) \\ & + \int_{Y_i}^{y} \frac{\partial\left[\bar{w}(X_i)\,D_{p,1}(u,X_i)\right]}{\partial x_1} du, \end{aligned} \tag{13}$$

$$\varphi^{\bar{w}}(U_i) \equiv \bar{\psi}^{\bar{w}}(U_i) + \int_{\mathcal{X}} \bar{w}(x)\left\{S(Y_i,x) - E[S(Y,x)]\right\} dx, \quad \bar{\psi}^{\bar{w}}(U_i) \equiv E\left[\psi^{\bar{w}}(U_i|Y_j)\,|U_i\right]. \tag{14}$$

The functions $D_{p,0}(y,x)$, $D_{p,y}(y,x)$ and $D_{p,1}(y,x)$ are defined in Equation (27) in Appendix. Under the above conditions, we then have the following asymptotic distribution of the proposed estimator:

Theorem 2. *Let Assumptions A1 through A11 and the normalization condition* (4) *hold. Then, the following functional weak convergence result holds for any compact set* $[y_1, y_2] \subseteq \mathcal{Y}$*:*

$$\sqrt{n}(\hat{\Theta}(y) - \Theta(y)) \to^d W(y), \quad y_1 \leq y \leq y_2,$$

where $y \to W(y)$ *is a zero-mean Gaussian process with covariance kernel* $\Omega(y_1, y_2) = E[\delta^w(U_i|y_1)\,\delta^w(U_i|y_2)]$.

As can be seen from the above expression, the function $\delta^w(U_i|y) = \delta(U_i|y, w, \Phi)$ is a known functional of w and Φ, and so the asymptotic covariance kernel can be consistently estimated by

$$\hat{\Omega}(y_1, y_2) = \frac{1}{n}\sum_{i=1}^{n} \delta(U_i|y_1, w, \hat{\Phi})\delta(U_i|y_2|w, \hat{\Phi}),$$

where $\hat{\Phi}$ is the kernel estimator given in Equation (8).

In principle, efficiency of the estimator can be obtained by minimizing the asymptotic variance $E[\delta_i^w(y)^2]$ as a functional of w. Given the complex expression of the influence function $\delta_i^w(y)$, this is a quite complicated problem though and so we leave the derivation of the optimal weighting function for future research.

With Theorem 2 in hand, we are now able to develop the asymptotic properties of the regression estimator proposed in Equation (11). To this end, we first extend the conditions of Blundell, Chen, and Kristensen (2007) to a multivariate setting to ensure that the infeasible estimator $\tilde{g}$ in Equation (9) is well-behaved; these are straightforward extensions and also rather technical and so have been relegated to the Appendix. Next, we impose the following assumption:

Assumption A12. *The support* $\mathcal{Y}$ *of* Y *is compact.*

This condition is a slight strengthening of Assumption A12 with the latter implying that $\mathcal{Y}$ is bounded. Sufficient conditions for the compact support assumption is that g is bounded and ϵ has compact support. When the function g in the model (1) is bounded, then the completeness condition A7 can be replaced by a *bounded* completeness condition: for every bounded function $m : \mathcal{X}_{-1} \to \mathbb{R}$, $E[m(X_{-1})|Z] = 0$ w.p.1 implies $m(X_{-1}) = 0$ w.p.1. The bounded completeness condition is weaker than the completeness condition (see, e.g., Blundell, Chen, and Kristensen, 2007, for a discussion).

The compactness of $\mathcal{Y}$ together with Theorem 2 implies that $\hat{\Theta}(y)$ converges uniformly over its support, $\sup_{y\in\mathcal{Y}} |\hat{\Theta}(y) - \Theta(y)| = O_P(1/\sqrt{n})$. This in turn allows us to show that the feasible estimator $\hat{g}$ is asymptotically equivalent to $\tilde{g}$.

Theorem 3. *Let Assumptions A1 through A12 and the normalization condition* (4) *hold. Assume in addition that Assumptions A13 through A17 in the Appendix hold. Then, the feasible sieve IV estimator $\hat{g}$ satisfies*

$$\|\hat{g} - g\|_X = \sqrt{\int_{\mathcal{X}} [\hat{g}(x) - g(x)]^2 f_X(x)\, dx} = O_p\left(k_n^{-r/d_x} + \tau_n\sqrt{k_n/n}\right),$$

where $d_x = \dim(X)$, $k_n = \dim(\mathcal{G}_n)$, $r \geq 1$ is the degree of smoothness of g and τ_n is the sieve measure of ill-posedness:

$$\tau_n \equiv \sup_{g_n\in\mathcal{G}_n : g_n \neq 0} \frac{\sqrt{E\{g_n(X)\}^2}}{\sqrt{E\{E[g_n(X)|X_1, Z]\}^2}}. \tag{15}$$

The convergence rate depends on the sieve-measure of ill-posedness τ_n which in turn depends on the decay rate of the singular values, which we denote $\{\mu_k\}$, of the conditional mean operator $g \mapsto M(x_1, z|g)$ defined in Equation (10); see Section 4 in Blundell, Chen, and Kristensen (2007) for a further discussion. If for example, the singular values satisfy $\mu_k \asymp k^{-s/d_x}$, for some $s > 0$ then $\tau_n \leq const \times k_n^{s/d_x}$ and we obtain $\|\hat{g} - g\|_X = O_p\left(n^{-r/[2(r+s)+d_x]}\right)$.

The convergence rate stated in Theorem 3 is identical to the one for the infeasible estimator, $\tilde{g}$, that assumes knowledge of T; thus, there is no (asymptotic) loss from not knowing T in the estimation of g. This is due to the fact that $\hat{T}$ converges with faster rate than $\tilde{g}$, and so it does not influence the feasible estimator $\hat{g}$. The above result only gives the rate of convergence of the estimator. We conjecture that the general results of Belloni, Chen, Chernozhukov, and Liao (2010) could be applied to our problem to develop distributional results.

We conjecture that Theorem 3 remains true without the assumption of bounded, compact support of Y. In particular, by inspection of the proof of Theorem 3, we see that all that is needed for the result to hold is that $||\hat{\Theta}(\cdot) - \Theta(\cdot)||_Y =$

$o_P\left(n^{-r/[2(r+s)+1]}\right)$, where $\|\cdot\|_Y$ is the L_2-norm, $\|\Theta\|_Y^2 = \int_{\mathcal{Y}} \Theta^2(y) f_Y(y)\, dy$. We expect this to hold in great generality.

Finally, we note that with $\hat{g}$ and $\hat{\Theta}$ converging uniformly, the estimator $\hat{F}_{\epsilon|X}(t, x_{-1})$ is clearly also consistent. However, the derivation of the asymptotic distribution of $\hat{F}_{\epsilon|X}(t, x_{-1})$ remains an open problem.

5. Conclusion

We conclude by discussing possible extensions of our identification result. Assume that instead of relying on the conditional independence between ϵ and X_1 given X_{-1}, we use the fact that there exists an instrument V, such that ϵ and X_1 are conditionally independent given (X_{-1}, V), i.e. $\epsilon \perp X_1 \mid (X_{-1}, V)$. This would amount to considering the conditional distribution $F_{Y|X,V}$ of Y given (X, V) which now satisfies:

$$F_{Y|X,V}(y, x, v) \equiv \Phi(y, x, v) = F_{\epsilon|X,V}\left(\Theta(y) - g(x), x_{-1}, v\right)$$

Redefining X to be (X, V), the above expression falls exactly in the framework obtained in (16), with an additional restriction on the function g which now no longer depends on the components of X corresponding to V. When the conditional distribution of the redefined vector X_{-1} given Z is complete, we know that g is identifiable. This identification result holds even without restricting the way that g depends on V; a fortiori, the identification result remains true when g is restricted.

Appendix A. Sieve IV Assumptions

We here state the additional regularity conditions used to establish Theorem 3. First, we need som additional notation: The first-step conditional mean estimators: $\tilde{h}(x_1, z)$ and $\hat{M}(x_1, z|g_n)$ are assumed to take the form

$$\begin{aligned}\tilde{h}(x_1, z) &= p^{J_n}(x_1, z)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_{1,i}, Z_i)\Theta(Y_i), \\ \hat{M}(x_1, z|g_n) &= p^{J_n}(x_1, z)'(P'P)^- \sum_{i=1}^n p^{J_n}(X_{1,i}, Z_i) g_n(X_i),\end{aligned}$$

where $p^{J_n}(z_1, z) = (p_1(x_1, z), ..., p_{J_n}(x_1, z))'$ is a sieve basis of dimension $J_n \geq 1$, and $P = (p^{J_n}(X_{1,1}, Z_1), ..., p^{J_n}(X_{1,n}, Z_n))'$. Also let $\Lambda^r_c(\mathcal{X}) \equiv \{g \in \Lambda^r(\mathcal{X}) : ||g||_{\Lambda^r} \leq c\}$ be a Hölder ball (of radius c) of fucntions with smoothness r as introduced in Blundell, Chen, and Kristensen (2007). We are then ready to state the regularity conditions

Assumption A13. *(i) $g \in \mathcal{G} \equiv \Lambda^r_c(\mathcal{X})$ for some $r > 1/2$; (ii) $E[||X||^{2a}] < \infty$ for some $a > r$.*

Assumption A14. *The functions $h(x_1, z) \equiv E[\Theta(Y)|X_1 = x_1, Z = z]$ and $M(x_1, z|g_n) \equiv E[g_n(X)|X_1 = x_1, Z = z]$ belong to $\mathcal{H} \equiv \Lambda^{r_m}_c(\mathcal{X}_1 \times \mathcal{Z})$, $r_m > 1/2$, for any $g_n \in \mathcal{G}_n$.*

Assumption A15. *(i) the smallest and the largest eigenvalues of $E[p^{J_n}(X_1, Z)p^{J_n}(X_1, Z)']$ are bounded and bounded away from zero for each J_{2n}; (ii) $p^{J_n}(x_1, z)$ is either a cosine series or a B-spline basis of order γ_b, with $\gamma_b > r_m > 1/2$; (iii) the density of (X_1, Z) is continuous, bounded and bounded away from zero over its support $\mathcal{X}_1 \times \mathcal{Z}$, which is a compact interval with non-empty interior.*

Assumption A16. *There is a $g_n \in \mathcal{G}_n$ such that $\tau_n^2 \times E[E[g(X) - g_n(X)|X_1, Z]^2] \leq const \times ||g - g_n||^2_X$.*

Assumption A17. *(i) $k_n \to \infty$, $J_n/n \to 0$; (ii) $nJ_n^{-2r_m/(1+d_z)-1} \to 0$ and $\lim_{n\to\infty}(J_n/k_n) = c_0 > 1$;*

Appendix B. Proofs

Proof of Theorem 1. Consider a structure $(T, g, F_{\epsilon|X})$ that satisfies assumptions A1-A6, and generates $\Phi(y, x)$ in the sense of:

$$\Phi(y, x) = F_{\epsilon|X}(\Theta(y) - g(x), x_{-1}). \tag{16}$$

To establish the results of Theorem 1 we proceed in three steps. The first step establishes the identification of Θ. The second step shows that the completeness assumption A7 is sufficient to identify g and $F_{\epsilon|X}$. The third and final step shows that the completeness condition is also necessary.

Step 1: Identification of Θ. Under assumptions A1, A3, A5 and A6, the partial derivatives $\partial\Phi(y, x)/\partial y$, $\partial\Phi(y, x)/\partial x_1$ exist. Differentiating Equation (16) in y and x_1 gives:

$$\frac{\partial\Phi(y, x)}{\partial y} = \Theta'(y) f_{\epsilon|X}(\Theta(y) - g(x), x_{-1}) \tag{17}$$

$$\frac{\partial\Phi(y, x)}{\partial x_1} = -\frac{\partial g(x)}{\partial x_1} f_{\epsilon|X}(\Theta(y) - g(x), x_{-1}) \tag{18}$$

where Θ' is the derivative of Θ, and $f_{\epsilon|X}(t, x_{-1})$ denotes the pdf of ϵ given $X_{-1} = x_{-1}$.

Take any point $\bar{x} \in A$ with A defined in Theorem 1. Then for every $y \in \mathcal{Y}$, we have:

$$-\frac{\Theta'(y)}{\partial g(\bar{x})/\partial x_1} = s(y, \bar{x}) \quad \text{where} \quad s(y, \bar{x}) \equiv \frac{\partial\Phi(y, \bar{x})/\partial y}{\partial\Phi(y, \bar{x})/\partial x_1}. \tag{19}$$

Note that $s(y, \bar{x})$ is nonzero and keeps a constant sign for all $y \in \mathcal{Y}$. Note in addition that under Assumptions A2 and A3, $\mathcal{Y}$ is a connected subset of $\mathbb{R}$ (i.e. an interval) that contains 0. Then, integrating (19) from 0 to any $y \in \mathcal{Y}$ and using the fact that $\Theta(0) = 0$ we have:

$$\Theta(y) = -\frac{\partial g(\bar{x})}{\partial x_1} S(y, \bar{x}) \quad \text{where} \quad S(y, \bar{x}) \equiv \int_0^y s(t, \bar{x}) dt.$$

Multiplying the above equation by the pdf $f_Y(\cdot)$ of Y and then integrating w.r.t. y, we get:

$$1 = E[\Theta(Y)] = -\frac{\partial g(\bar{x})}{\partial x_1}\int_{\mathbb{R}} S(y,\bar{x}) f_Y(y)dy = -\frac{\partial g(\bar{x})}{\partial x_1} E[S(Y,\bar{x})],$$

where we have used the fact that $E[\Theta(Y)] = E[g(X)] + E[\epsilon] = 1$. Since $\bar{x} \in A$, $\partial g(\bar{x})/\partial x_1 \neq 0$ and is finite; hence, $E[S(Y,\bar{x})] \neq 0$ and is finite as well, so we can write:

$$\frac{\partial g(\bar{x})}{\partial x_1} = -\frac{1}{E[S(Y,\bar{x})]}. \tag{20}$$

Combining (19) and (20) then gives for every $y \in \mathcal{Y}$:

$$\Theta(y) = \frac{S(y,\bar{x})}{E[S(Y,\bar{x})]}, \tag{21}$$

and the right-hand side of (21) does not depend on $\bar{x}$. Hence, Θ is identified.

STEP 2: IDENTIFICATION OF g AND $F_{\epsilon|X}$. Now take any $x \in \mathcal{X}$ such that $\partial g(x)/\partial x_1 \neq 0$. For any such x, there exists a $y_x \in \mathcal{Y}$ such that $\Theta(y_x) - g(x) \in \mathcal{E}_x$, i.e. such that $f_{\epsilon|X}(\Theta(y_x) - g(x), x_{-1}) > 0$. Taking again ratios in (17)-(18), it follows that

$$\partial g(x)/\partial x_1 = -\frac{\Theta'(y_x)}{s(y_x,x)} \quad \text{where} \quad s(y_x,x) = \frac{\partial\Phi(y_x,x)/\partial y}{\partial\Phi(y_x,x)/\partial x_1},$$

and with Θ as in (21). Now let $\Gamma : \mathcal{X} \to \mathbb{R}$ be defined as:

$$\Gamma(x) \equiv \begin{cases} -\frac{\Theta'(y_x)}{s(y_x,x)}, & \text{if } x \in \{x \in \mathcal{X} : \frac{\partial g(x)}{\partial x_1} \neq 0\}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, we have that $\partial g(x)/\partial x_1 = \Gamma(x)$ for a.e. $x \in \mathcal{X}$. A particular solution $\bar{g} : \mathcal{X} \to \mathbb{R}$ to this partial differential equation is

$$\bar{g}(x_1, x_2, \ldots, x_{d_x}) = \int_c^{x_1} \Gamma(u, x_2, \ldots, x_{d_x})du \tag{22}$$

where $c \in \mathcal{X}_1$. Obviously, any solution to $\partial g(x)/\partial x_1 = \Gamma(x)$ must have the same partial in x_1 as $\bar{g}$ in (22); it must therefore be of the form:

$$g(x) = \bar{g}(x) + \beta(x_{-1}) \tag{23}$$

for some function $\beta : \mathcal{X}_{-1} \to \mathbb{R}$. Now let g be an arbitrary solution, and consider $E(\epsilon|Z)$ where $\epsilon = \Theta(Y) - g(X)$ with Θ as in (21) and g as in (23). Letting $F_{Y|Z}$ and $F_{X|Z}$ denote the conditional distributions of Y given Z and of X given Z, respectively, we have:

$$\begin{aligned} E(\epsilon|Z=z) &= \int_{\mathbb{R}} \Theta(y) dF_{Y|Z}(y,z) - \int_{\mathcal{X}} g(x) dF_{X|Z}(x,z) \\ &= \int_{\mathbb{R}} \Theta(y) dF_{Y|Z}(y,z) - \int_{\mathcal{X}} [\bar{g}(x) + \beta(x_{-1})] dF_{X|Z}(x,z) \end{aligned} \tag{24}$$

Now, consider a structure $(\tilde{T}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ that is observationally equivalent to $(T, g, F_{\epsilon|X})$ and has the same properties as $(T, g, F_{\epsilon|X})$. It follows from (24) that for a.e. $z \in \mathcal{Z}$:

$$E(\epsilon|Z=z) = 0 = E(\tilde{\epsilon}|Z=z) \Rightarrow E\left[\beta(X_{-1}) - \tilde{\beta}(X_{-1})|Z=z\right] = 0,$$

where $\tilde{\epsilon} = \tilde{\Theta}(Y) - \tilde{g}(X)$. Then, the completeness assumption A7 implies $\beta(x_{-1}) = \tilde{\beta}(x_{-1})$ for a.e. $x_{-1} \in \mathcal{X}_{-1}$. Combined with Equation (23), this implies that

$$g(x) = \tilde{g}(x), \quad \text{for a.e. } x \in \mathcal{X}.$$

Thus g is identified.

Since Θ and g are identified, we have $F_{\epsilon|X}(\Theta(y) - g(x), x_{-1}) = \tilde{F}_{\tilde{\epsilon}|X}(\Theta(y) - g(x), x_{-1})$ for every $y \in \mathcal{Y}$ and a.e. $x \in \mathcal{X}$. Now take any $x \in \mathcal{X}$; then the previous equality holds for any $t = \Theta(y) - g(x) \in \mathcal{E}_x$. By continuity, the equality continues to hold outside the support $\mathcal{E}_x$, i.e. $F_{\epsilon|X}(t, x_{-1}) = \tilde{F}_{\tilde{\epsilon}|X}(t, x_{-1})$ for every $t \in \mathbb{R}$. This establishes the identification of $F_{\epsilon|X}$ and completes the proof of sufficiency.

Step 3. Necessity. Finally, assume that the completeness condition is violated, in the sense that there exists some function $h : \mathcal{X}_{-1} \to \mathbb{R}$ that (i) does not vanish a.e., but (ii) is such that $E[h(X_{-1}) \mid Z = z] = 0$ for a.e. $z \in \mathcal{Z}$. Let $(T, g, F_{\epsilon|X})$ be a structure generating Φ, that satisfies Assumptions A1-A6 and the normalization condition (4). Define $(\tilde{T}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ by

$$\tilde{\Theta}(y) \equiv \Theta(y), \ \tilde{g}(x) \equiv g(x) + h(x_{-1}), \text{ and } \tilde{F}_{\tilde{\epsilon}|X}(t,x) \equiv F_{\tilde{\epsilon}|X}(t + h(x_{-1}), x_{-1}),$$

for every $y \in \mathcal{Y}$, every $t \in \mathbb{R}$, and a.e. $x \in \mathcal{X}$. Then, the structure $(\tilde{T}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ satisfies the normalization condition (4), as well as assumptions A1-A6. Note that assumption A6 only requires $\tilde{g}$ to be smooth with respect to the first component x_1; hence, it is satisfied even if the function $h(x_{-1})$ is discontinuous. Since the structure $(\tilde{T}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ is observationally equivalent to $(T, g, F_{\epsilon|X})$, $(T, g, F_{\epsilon|X})$ is not identified. □

Proof of Theorem 2. Write

$$\begin{aligned}
\hat{\Theta}(y) - \Theta(y) &= \int_{\mathcal{X}_0} w(x) \left\{ \frac{\hat{S}(y,x)}{\hat{E}[\hat{S}(Y,x)]} - \frac{S(y,x)}{E[S(Y,x)]} \right\} dx \\
&= \int_{\mathcal{X}_0} \frac{w(x)}{E[S(Y,x)]} \{\hat{S}(y,x) - S(y,x)\} dx \\
&\quad - \int_{\mathcal{X}_0} \frac{w(x)}{E[S(Y,x)]^2} \{\hat{E}[\hat{S}(Y,x)] - E[S(Y,x)]\} dx \\
&\quad + O(||\hat{S} - S||_\infty^2) + O(||\hat{E}[\hat{S}] - E[S]||_\infty^2),
\end{aligned}$$

where $\|\cdot\|_\infty$ here and in the following denotes the supremum norm over the set $\mathcal{Y} \times \mathcal{X}_0$; that is, $\|S\|_\infty = \sup_{(y,x) \in \mathcal{Y} \times \mathcal{X}_0} \|S(y,x)\|$. Applying in turn Lemmas 1 and 2,

$$\begin{aligned}
&\int_{\mathcal{X}_0} \frac{w(x)}{E[S(Y,x)]} \{\hat{S}(y,x) - S(y,x)\} dx \\
= &\int_{\mathcal{X}_0} \frac{w(x)}{E[S(Y,x)]} \{\nabla_p S(y,x)[\hat{p} - p] + \nabla_f S(y,x)[\hat{f} - f]\} dx + o_P\left(1/\sqrt{n}\right) \\
= &\frac{1}{n} \sum_{i=1}^{n} \psi^{\bar{w}_2}(U_i|y) + o_P\left(1/\sqrt{n}\right),
\end{aligned}$$

with $\psi^{\bar{w}}(U_i|y)$ defined in Equation (13) and $\bar{w}_1(x) \equiv w(x)/E[S(Y,x)]$. Next, from Lemma 3, we obtain

$$\int_{\mathcal{X}_0} \frac{w(x)}{E[S(Y,x)]^2} \left\{ \hat{E}\left[\hat{S}(Y,x)\right] - E[S(Y,x)] \right\} dx = \frac{1}{n} \sum_{i=1}^{n} \varphi^{\bar{w}_2}(U_i) + o_P\left(1/\sqrt{n}\right)$$

where $\varphi^{\bar{w}}(U_i)$ is defined in Equation (14) and $\bar{w}_2(x) \equiv w(x)/E[S(Y,x)]^2$. Finally, by Lemmas 1 and 4, $||\hat{S} - S||_\infty^2 = o_P(1/\sqrt{n})$ and $||\hat{E}[\hat{S}] - E[S]||_\infty^2 = o_P(1/\sqrt{n})$. In

total, uniformly over $\mathcal{Y}$,

$$\sqrt{n}(\hat{\Theta}(y) - \Theta(y)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta^w(U_i|y) + o_P(1),$$

where $\delta^w(U_i|y)$ is defined in Equation (12). Pointwise weak convergence now follows by the CLT for i.i.d. sequences. This extends to weak functional convergence over any compact set $[y_1, y_2] \subseteq \mathcal{Y}$ if we can show stochastic equicontinuity. However, this follows from, for example, der Vaart and Wellner (1996) since $y \mapsto \delta^w(U_i|y)$ is continuous almost surely and has an L_2-envelope, $|\delta^w(U_i|y)| \leq \bar{\delta}^w(U_i)$, $y \in [y_1, y_2]$, with $E[\bar{\delta}^w(U_i)^2] < \infty$. The envelope takes the form $\bar{\delta}^w(U_i) := \bar{\psi}^{\bar{w}_1}(U_i) + \varphi^{\bar{w}_2}(U_i)$ where

$$\begin{aligned} \psi^{\bar{w}}(U_i) \equiv\ & \bar{w}(X_i) \int_{Y_i}^{y_2} |D_{p,0}(u, X_i)|\, du + \bar{w}(X_i) |D_{p,y}(Y_i, X_i)| \\ & + \int_{Y_i}^{y_2} \left| \frac{\partial [\bar{w}(X_i) D_{p,1}(u, X_i)]}{\partial x_1} \right| du. \end{aligned}$$

□

Proof of Theorem 3. We first extend Theorem 2 of Blundell, Chen, and Kristensen (2007) to allow for multiple regressors and IVs. To this end, we establish multivariate versions of Claims 1-2 as stated in the proof of Theorem 2 in Blundell, Chen, and Kristensen (2007). We do this without proof since these are standard results for sieve estimators:

Claim 1: For any $g \in \mathcal{G}$, there is a $g_n \in \mathcal{G}_n$ satisfying $\|g - g_n\|_X \leq \text{const.} \times k_n^{-r/d_x}$. Similarly, for any $h \in \mathcal{H}$, there is a $h_n \in \mathcal{H}_n$ such that$\|h - h_n\|_{X_1,Z} \leq \text{const.} \times J_n^{-r_m/(1+d_z)}$.

Claim 2:

$$\begin{aligned} \text{(i)} \quad & ||\tilde{h} - h||_{X_1,Z} = O_p\left(J_n^{-r_m/(1+d_z)} + \sqrt{J_n/n}\right), \\ \text{(ii)} \quad & \sup_{g_n \in \mathcal{G}_n} ||\hat{M}(\cdot|g_n) - M(\cdot|g_n)||_{X_1,Z} = O_p\left(J_n^{-r_m/(1+d_z)} + \sqrt{J_n/n}\right). \end{aligned}$$

By inspection of the remaining arguments used in the proof of Theorem 2 in Blundell, Chen, and Kristensen (2007), we see that these remain correct without further modifications with multiple regressors and IVs. Thus, combining the above Claims 1-2 with the remaining arguments of Theorem 2 in Blundell, Chen, and Kristensen (2007), we conclude that the infeasible estimator $\tilde{g}$ (assuming T known) satisfies

$$||\tilde{g} - g||_X \leq ||g - g_n||_X + \tau_n \times O_p\left(J_n^{-r_m/(1+d_z)} + \sqrt{J_n/n} + ||M(\cdot|g - g_n)||_{X_1,Z}\right).$$

Using Assumptions A16 and A17 together with the fact that $||g - g_n||_X \leq$ const. $\times$ k_n^{-r/d_x}, we obtain

$$\begin{aligned} ||\tilde{g} - g||_X &= O_P\left(k_n^{-r/d_x}\right) + \tau_n \times O_p\left(J_n^{-r_m/(1+d_z)} + \sqrt{J_n/n}\right) \\ &= O_P\left(k_n^{-r/d_x}\right) + \tau_n \times O_p\left(\sqrt{k_n/n}\right). \end{aligned}$$

Next, by inspection of the above proof for the convergence rate of the infeasible estimator, observe that $\Theta(Y)$ only enters the arguments in Claim 2(i) through $\tilde{h}(z)$. In particular, the above arguments remain correct with $\tilde{h}(z)$ replaced by any other estimator which satisfies Claim 2(i). By definition of $\tilde{h}$ and $\hat{h}$ and Theorem 2, $||\hat{h} - \tilde{h}||_{X_1,Z} \leq \sup_{y\in\mathcal{Y}} |\hat{\Theta}(y) - \Theta(y)| = O_P(1/\sqrt{n})$, and so Claim 2(i) remains intact when replacing $\tilde{h}$ by $\hat{h}$. And this yields exactly the feasible estimator, $\hat{g}$. □

Appendix C. Lemmas

In the following, we let $\Phi(y,x)$, $p(y,x)$ and $f(x)$ denote the true, data-generating cdf, joint density and marginal density respectively. We define the following functionals for any functions $dp(y,x)$ and $df(x)$:

$$\begin{aligned} (25)\ \bigtriangledown_p S(y,x)[dp] &:= \int_0^y D_{p,0}(u,x)\, dp(u,x)\, du + \int_0^y D_{p,y}(u,x)\, dp_y(u,x)\, du \\ &\quad + \int_0^y D_{p,1}(u,x)\, dp_1(u,x)\, du, \end{aligned}$$

$$(26)\qquad \bigtriangledown_f S(y,x)[df] \equiv \int_0^y D_{f,0}(u,x)\, du \times df(x) + \int_0^y D_{f,1}(u,x)\, du \times df_1(x),$$

where $dp_y(y,x) = \partial dp(y,x)/(\partial y)$ and so forth, and

$$
\begin{aligned}
&D_{p,0}(y,x) \equiv \frac{\Phi_y(y,x) f_1(x)}{\Phi_1^2(y,x) f^2(x)}, \quad D_{p,y}(y,x) \equiv \frac{1}{f(x)\Phi_1(y,x)}, \\
&D_{f,0}(y,x) \equiv \frac{p_y^2(y,x)}{f(x)\Phi_1(y,x)} - \frac{\Phi_y(y,x)}{\Phi_1^2(y,x)}\left\{\frac{2p(y,x)}{f^3(x)} f_1(x) + \frac{p_1^2(y,x)}{f(x)}\right\}, \\
&D_{f,1}(y,x) \equiv \frac{\Phi_y(y,x) p(y,x)}{\Phi_1^2(y,x) f^2(x)}, \quad D_{p,1}(y,x) \equiv -\frac{\Phi_y(y,x)}{f(x)\Phi_1^2(y,x)}.
\end{aligned} \tag{27}
$$

The first lemma then shows that these two functionals are the pathwise differentials of $S(y,x)$ w.r.t. g and f respectively:

Lemma 1. *Under Assumptions A1-A11: With $\bigtriangledown_p S(y,x)[dp]$ and $\bigtriangledown_f S(y,x)[df]$ defined in Equations (25)-(26), the following expansion holds uniformly over $(y,x) \in \mathcal{Y} \times \mathcal{X}_0$:*

$$
\hat{S}(y,x) - S(y,x) = \bigtriangledown_p S(y,x)[\hat{p} - p] + \bigtriangledown_f S(y,x)[\hat{f} - f] + o_P\left(1/\sqrt{n}\right),
$$

Proof of Lemma 1. Let $\hat{\Phi} = \hat{p}/\hat{f}$ denote the kernel estimator. First, by a standard Taylor expansion (where we suppress dependence on y and x)

$$
\frac{\hat{\Phi}_y}{\hat{\Phi}_1} - \frac{\Phi_y}{\Phi_1} = \frac{1}{\Phi_1}\{\hat{\Phi}_y - \Phi_y\} - \frac{\Phi_y}{\Phi_1^2}\{\hat{\Phi}_1 - \Phi_1\} + O\left(|\hat{\Phi}_y - \Phi_y|^2\right) + O\left(|\hat{\Phi}_1 - \Phi_1|^2\right),
$$

where the derivatives w.r.t. y and x_1 respectively are on the form

$$
\Phi_y = \frac{p_y}{f}, \quad \Phi_1 = \frac{p_1}{f} - \frac{p f_1}{f^2}.
$$

We then Taylor expand those w.r.t. p and f:

$$
\hat{\Phi}_y - \Phi_y = \frac{1}{f}\{\hat{p}_y - p_y\} + \frac{p_y^2}{f}\{\hat{f} - f\} + O\left(|\hat{p}_y - p_y|^2\right) + O\left(|\hat{f} - f|^2\right),
$$

and

$$
\begin{aligned}
\hat{\Phi}_1 - \Phi_1 \;=\; & -\frac{f_1}{f^2}\{\hat{p} - p\} + \frac{1}{f}\{\hat{p}_1 - p_1\} + \left(\frac{2pf_1}{f^3} + \frac{p_1^2}{f}\right)\{\hat{f} - f\} - \frac{p}{f^2}\{\hat{f}_1 - f_1\} \\
& + O\left(|\hat{p} - p|^2\right) + O\left(|\hat{p}_1 - p_1|^2\right) + O\left(|\hat{f} - f|^2\right) + O\left(|\hat{f}_1 - f_1|^2\right).
\end{aligned}
$$

Combining these expressions we now obtain

$$\begin{aligned}\frac{\hat{\Phi}_y}{\hat{\Phi}_1}-\frac{\Phi_y}{\Phi_1} &= \frac{\Phi_y f_1}{\Phi_1^2 f^2}\{p-p_0\}+\frac{1}{f\Phi_1}\{\hat{p}_y-p_y\}+\frac{p_y^2}{f\Phi_1}\{\hat{f}-f\}-\frac{\Phi_y}{f\Phi_1^2}\{\hat{p}_1-p_1\}\\ &\quad -\frac{\Phi_y}{\Phi_1^2}\left(\frac{2p}{f^3}f_1+\frac{p_1^2}{f}\right)\{\hat{f}-f\}+\frac{\Phi_y p}{\Phi_1^2 f^2}\{\hat{f}_1-f_1\}+R\\ &= D_{p,0}\{\hat{p}-p_0\}+D_{p,y}\{\hat{p}_y-p_{0,y}\}+D_{p,1}\{\hat{p}_1-p_{0,1}\}\\ &\quad +D_{f,0}\{\hat{f}-f_0\}+D_{f,1}\{\hat{f}_1-f_{0,1}\}+R,\end{aligned}$$

where R is the remainder term satisfying

$$R=O\left(|\hat{p}-p|^2\right)+O\left(|\hat{p}_1-p_1|^2\right)+O\left(|\hat{p}_y-p_y|^2\right)+O\left(|\hat{f}-f|^2\right)+O\left(|\hat{f}_1-f_1|^2\right),$$

and $D_{p,0}$, $D_{p,y}$, $D_{p,1}$, $D_{f,0}$ and $D_{f,1}$ are defined in Equation (27). Given the definitions of $\bigtriangledown_p S(y,x)[dp]$ and $\bigtriangledown_f S(y,x)[df]$, we now obtain

$$\hat{S}(y,x)-S(y,x)=\bigtriangledown_p S(y,x)[\hat{p}-p]+\bigtriangledown_f S(y,x)[\hat{f}-f]+R,$$

and what remains to be shown is that the remainder term $R=o_P(1/\sqrt{n})$ uniformly in $(x,y)\in\mathcal{X}_0\times\mathcal{Y}$. By standard results for kernel density smoothers of i.i.d. data (see e.g. Hansen (2008), Proof of Theorem 6) the following rates hold under Assumptions A8 and A9:

$$\begin{aligned}||\hat{p}-p||_\infty &= O_P\left(\max(h_x,h_y)^m\right)+O_P\left(\sqrt{\frac{\log(n)}{nh_x^{d_x}}}\right),\\ ||\hat{p}_1-p_1||_\infty &= O_P\left(\max(h_x,h_y)^m\right)+O_P\left(\sqrt{\frac{\log(n)}{nh_x^{d_x+1}}}\right),\\ ||\hat{p}_y-p_y||_\infty &= O_P\left(\max(h_x,h_y)^m\right)+O_P\left(\sqrt{\frac{\log(n)}{nh_yh_x^{d_x}}}\right),\\ ||\hat{f}-f||_\infty &= O_P\left(h_x^m\right)+O_P\left(\sqrt{\frac{\log(n)}{nh_x^{d_x}}}\right),\\ ||\hat{f}_1-f_1||_\infty &= O_P\left(h_x^m\right)+O_P\left(\sqrt{\frac{\log(n)}{nh_x^{d_x+1}}}\right).\end{aligned} \tag{28}$$

Now, under Assumption A11, we see that the squared estimation error of the kernel estimators $\hat{p}$ and $\hat{f}$ and their relevant derivatives all are of order $o_P\left(1/\sqrt{n}\right)$. In particular, $\sup_{(x,y)\in\mathcal{X}_0\times\mathcal{Y}} R = o_P\left(1/\sqrt{n}\right)$ which completes the proof. □

Lemma 2. *Under Assumptions A1-A11: For any weighting function $\bar{w}$ with support $\mathcal{X}_0$, the functionals $\bigtriangledown_p S\left(y,x\right)[dp]$ and $\bigtriangledown_f S\left(y,x\right)[df]$ defined in Equations (25)-(26) satisfy uniformly over $y\in\mathcal{Y}$:*

$$\int_{\mathcal{X}} \bar{w}\left(x\right)\left\{\bigtriangledown_p S\left(y,x\right)[\hat{p}-p]+\bigtriangledown_f S\left(y,x\right)[\hat{f}-f]\right\}dx = \frac{1}{n}\sum_{i=1}^{n}\psi^{\bar{w}}\left(U_i|y\right)+o_P\left(1/\sqrt{n}\right),$$

where $\psi^{\bar{w}}\left(U_i|y\right)$ is defined in Equation (13).

Proof of Lemma 2. By definition,

$$\begin{aligned}\bigtriangledown_p S\left(y,x\right)[\hat{p}] &= \int_0^y D_{p,0}\left(u,x\right)\hat{p}\left(u,x\right)du+\int_0^y D_{p,y}\left(u,x\right)\hat{p}_y\left(u,x\right)du\\ &\quad+\int_0^y D_{p,1}\left(u,x\right)\hat{p}_1\left(u,x\right)du\\ &= :\bigtriangledown_p^{(1)} S\left(y,x\right)[\hat{p}]+\bigtriangledown_p^{(2)} S\left(y,x\right)[\hat{p}]+\bigtriangledown_p^{(3)} S\left(y,x\right)[\hat{p}].\end{aligned}$$

Here, with $x=(x_1,x_{-1})$,

$$\begin{aligned}\bigtriangledown_p^{(1)} S\left(y,x\right)[\hat{p}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{h_x}\left(X_i-x\right)\int_0^y D_{p,0}\left(u,x\right)\mathcal{K}_{h_y}\left\{Y_i-u\right\}du\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{h_x}\left(X_i-x\right)\left[\int_0^y D_{p,0}\left(u,x\right)\mathbb{I}\left\{Y_i\le u\right\}du+O_P\left(h_y^m\right)\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{h_x}\left(X_i-x\right)\left[\int_{Y_i}^y D_{p,0}\left(u,x\right)du+O_P\left(h_y^m\right)\right].\end{aligned}$$

Similarly,

$$\begin{aligned}\bigtriangledown_p^{(2)} S\left(y,x\right)[\hat{p}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{h_x}\left(X_i-x\right)\int_0^y D_{p,y}\left(u,x\right)K_{h_y}\left\{Y_i-u\right\}du\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{h_x}\left(X_i-x\right)\left[\mathbb{I}\left\{Y_i\le y\right\}D_{p,y}\left(Y_i,x\right)+O_P\left(h_y^m\right)\right],\end{aligned}$$

and, writing $\mathbf{K}_{h_x}(X_i - x) = K_{h_x}(X_{1,i} - x_1)\,\mathbf{K}_{-1,h_x}(X_{-1,i} - x_{-1})$ with $x = (x_1, x_{-1})$,

$$\begin{aligned}
& \bigtriangledown_p^{(3)} S(y,x)[\hat{p}] \\
= & \frac{1}{n}\sum_{i=1}^{n} K'_{h_x}(X_{1,i} - x_1)\,\mathbf{K}_{-1,h_x}(X_{-1,i} - x_{-1}) \int_0^y D_{p,1}(u,x)\,K_{h_y}\{Y_i - u\}\,du \\
= & \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(X_{1,i} - x_1)\,\mathbf{K}_{-1,h_x}(X_{-1,i} - x_{-1}) \times \left[\int_{Y_i}^y D_{p,1}(u,x)\,du + O_P\left(h_y^m\right)\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
& \int_{\mathcal{X}} \bar{w}(x) \bigtriangledown_p^{(1)} S(y,x)[\hat{p}]\,dx \\
= & \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{X}} \bar{w}(x)\,\mathbf{K}_{h_x}(X_i - x) \int_0^y D_{p,0}(u,x)\,\mathcal{K}_{h_y}\{Y_i - u\}\,du \\
= & \frac{1}{n}\sum_{i=1}^{n} \int_{Y_i}^y \int_{\mathcal{X}} \bar{w}(x)\,D_{p,0}(u,x)\,\mathbf{K}_{h_x}(X_i - x)\,dx du \times \left[1 + O_P\left(h_y^m\right)\right] \\
= & \frac{1}{n}\sum_{i=1}^{n} \bar{w}(X_i) \int_{Y_i}^y D_{p,0}(u,X_i)\,du \times \left[1 + O_P\left(h_y^m\right) + O_P\left(h_x^m\right)\right].
\end{aligned}$$

By similar arguments,

$$\begin{aligned}
& \int_{\mathcal{X}} \bar{w}(x) \bigtriangledown_p^{(2)} S(y,x)[\hat{p}]\,dx \\
= & \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \le y\} \int_{\mathcal{X}} w(x)\,\mathbf{K}_{h_x}(X_i - x)\,D_{p,y}(Y_i,x)\,dx + O_P\left(h_y^m\right) \\
= & \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \le y\}\,w(X_i)\,D_{p,y}(Y_i,X_i)\left[1 + O_P\left(h_y^m\right) + O_P\left(h_x^m\right)\right],
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\mathcal{X}} \bar{w}(x) \bigtriangledown_p^{(3)} S(y,x)[\hat{p}]\, dx \\
= \;& \frac{1}{n}\sum_{i=1}^{n} \int_{Y_i}^{y} \int_{\mathcal{X}} \bar{w}(x) K'_{h_x}(X_{1,i}-x_1)\, \mathbf{K}_{h_x}(X_{-1,i}-x_{-1})\, D_{p,1}(u,x)\, dxdu \\
& \times \left[1+O_P\left(h_y^m\right)\right] \\
= \;& -\frac{1}{n}\sum_{i=1}^{n} \int_{Y_i}^{y} \int_{\mathcal{X}} K_{h_x}(X_{1,i}-x_1)\, \mathbf{K}_{h_x}(X_{-1,i}-x_{-1}) \frac{\partial}{\partial x_1}\left[\bar{w}(x) D_{p,1}(u,x)\right] dxdu \\
& \times \left[1+O_P\left(h_y^m\right)\right] \\
= \;& -\frac{1}{n}\sum_{i=1}^{n} \int_{Y_i}^{y} \frac{\partial\left[\bar{w}(X_i) D_{p,1}(u,X_i)\right]}{\partial x_1} du \left[1+O_P\left(h_y^m\right)+O_P\left(h_x^m\right)\right].
\end{aligned}$$

Since $\sqrt{n}\left[h_x^m+h_y^m\right]=o(1)$, the claimed result now holds. □

Lemma 3. *Under Assumptions A1-A11: For any weighting function $\bar{w}$ with support $\mathcal{X}_0$,*

$$\int_{\mathcal{X}} \bar{w}(x)\{\hat{E}[\hat{S}(Y,x)]-E[S(Y,x)]\}dx = \frac{1}{n}\sum_{i=1}^{n} \varphi^{\bar{w}}\psi^{\bar{w}}(U_i)+o_P\left(1/\sqrt{n}\right),$$

where $\varphi^{\bar{w}}(U_i)$ is defined in Equation (14).

Proof of Lemma 3. Applying Lemmas 1 and 2,

$$\begin{aligned}
&\int_{\mathcal{X}} \bar{w}(x)\hat{E}[\hat{S}(Y,x)]-E[S(Y,x)]dx \\
= \;& \int_{\mathcal{X}} \bar{w}(x)\{\hat{E}[\hat{S}(Y,x)]-\hat{E}[S(Y,x)]\}dx + \int_{\mathcal{X}} \bar{w}(x)\{\hat{E}[S(Y,x)]-E[S(Y,x)]\}dx \\
= \;& \frac{1}{n}\sum_{j=1}^{n} \int_{\mathcal{X}} \bar{w}(x)\{\hat{S}(Y_j,x)-S(Y_i,x)\}dx \\
& +\frac{1}{n}\sum_{j=1}^{n} \int_{\mathcal{X}} \bar{w}(x)\{S(Y_j,x)-E[S(Y,x)]\}dx \\
= \;& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \psi^{\bar{w}}(U_i|Y_j)+\frac{1}{n}\sum_{j=1}^{n} \int_{\mathcal{X}} \bar{w}(x)\{S(Y_j,x)-E[S(Y,x)]\}dx+o_P\left(1/\sqrt{n}\right),
\end{aligned}$$

The first term is a U-statistic, and by appealing to standard results (see e.g. Newey and McFadden (1994), Lemma 8.4),

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi^{\bar{w}}\left(U_i|Y_j\right)=\frac{1}{n}\sum_{i=1}^{n}\bar{\psi}^{\bar{w}}\left(U_i\right)+o_P\left(1/\sqrt{n}\right),$$

where $\bar{\psi}^{\bar{w}}\left(U_i\right)=E\left[\psi^{\bar{w}}\left(U_i|Y_j\right)|U_i\right]$. Thus,

$$\int_{\mathcal{X}}\bar{w}\left(x\right)\{\hat{E}[\hat{S}\left(Y,x\right)]-E[S\left(Y,x\right)]\}dx=\frac{1}{n}\sum_{i=1}^{n}\varphi^{\bar{w}}\left(U_i\right)+o_P\left(1/\sqrt{n}\right).$$

□

Lemma 4. *Under Assumptions A1-A11:*

$$\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|\bigtriangledown_p S\left(y,x\right)[\hat{p}-p]|^2=o_P\left(1/\sqrt{n}\right),\qquad\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|\bigtriangledown_f S\left(y,x\right)[\hat{f}-f]|^2=o_P\left(1/\sqrt{n}\right).$$

Proof of Lemma 4. >From the definition of $\bigtriangledown_p S\left(y,x\right)[dp]$,

$$\begin{aligned}&\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|\bigtriangledown_p S\left(y,x\right)[dp]|\\ \leq\ &\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|D_{p,0}\left(y,x\right)|\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|dp\left(y,x\right)|+\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|D_{p,y}\left(y,x\right)|\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|dp_y\left(y,x\right)|\\ &+\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|D_{p,1}\left(y,x\right)|\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|dp_1\left(y,x\right)|,\end{aligned}$$

where $\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|D_{p,a}\left(y,x\right)|<\infty$, $a=0,y,1$, given the smoothness and bound conditions imposed in Assumptions A9 and A10. Next, with $dp=\hat{p}-p$, it follows from the convergence rate results in Equation (28) together with the bandwidth requirement in Assumption A11 that $\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}_0}|\hat{p}\left(y,x\right)-p\left(y,x\right)|=o_P\left(1/n^{1/4}\right)$ and similarly for its partial derivatives w.r.t. y and x. This proves the first claim. The proof of the second claim follows along the same lines and so is left out. □

References

Abbring, J. H., P. Chiappori, and T. Zavadil (2007): "Better Safe than Sorry? Ex Ante and Ex Post Moral Hazard in Dynamic Insurance Data," VU University Amsterdam and Columbia University.

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

Belloni, A., X. Chen, V. Chernozhukov, and Z. Liao (2010): "On Limiting Distributions of Possibly Unbounded Functionals of Linear Sieve M-Estimators," Yale University.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.

Blundell, R., and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics, Theory and Applications, Eighth World Congress, Vol. II*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnovsky. Cambridge University Press.

Brown, B. W. (1983): "The Identification Problem in Systems Nonlinear in the Variables," *Econometrica*, 51, 175–196.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4–14.

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

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

der Vaart, A. V., and J. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer-Verlag.

d'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, forthcoming.

EKELAND, I., J. J. HECKMAN, AND L. NESHEIM (2004): "Identification and Estimation of Hedonic Models," *The Journal of Political Economy*, 112, S60–S109.

ENGLE, R. F. (2000): "The Econometrics of Ultra-High-Frequency Data," *Econometrica*, 68, 1–22.

FÈVE, F., AND J.-P. FLORENS (2010): "The practice of non-parametric estimation by solving inverse problems: the example of transformation models," *The Econometrics Journal*, 13, S1–S27.

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

HAN, A. K. (1987): "A Non-Parametric Analysis of Transformations," *Joumal of Econometrics*, 35, 191–209.

HANSEN, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24, 726–748.

HAUSMAN, J. A., AND D. A. WISE (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46, 403–426.

HECKMAN, J. J., R. L. MATZKIN, AND L. NESHEIM (2005): "Estimation and Simulation of Hedonic Models," in *Frontiers in Applied General Equilibrium*, ed. by T. Kehoe, T. Srinivasan, and J. Whalley. Cambridge University Press, Cambridge.

HODERLEIN, S. (2009): "Endogenous Semiparametric Binary Choice Models with Heteroscedasticity," Brown University.

HOROWITZ, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103–137.

JACHO-CHÁVEZ, D., A. LEWBEL, AND O. LINTON (2010): "Identification and Nonparametric Estimation of a Transformed Additively Separable Model," *Journal of Econometrics*, 156, 392–407.

KEIFER, N. M. (1988): "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 26, 646–679.

KOMUNJER, I. (2008): "Global Identification in Nonlinear Semiparametric Models," University of California San Diego.

KOOPMANS, T. C., AND O. REIERSØL (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics*, 21, 165–181.

LINTON, O., S. SPERLICH, AND I. VAN KEILEGOM (2008): "Estimation of a Semiparametric Transformation Model," *The Annals of Statistics*, 36, 686–718.

LO, A. W., A. C. MACKINLAY, AND J. ZHANG (2002): "Econometric models of limit-order executions," *Journal of Financial Economics*, 65, 31–71.

MATA, J., AND P. PORTUGAL (1994): "Life Duration of New Firms," *The Journal of Industrial Economics*, 42, 227–245.

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

——— (2007): "Nonparametric Identification," in *Handbook of Econometrics, Vol. 6B*, ed. by J. J. Heckman, and E. Leamer, pp. 5307–5368. Elsevier.

NEWEY, W. K., AND D. L. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2111–2245. Elsevier.

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

RIDDER, G. (1990): "The Non-Parametric Identification of Generalized Accelerated Failure-Time Models," *The Review of Economic Studies*, 57, 167–181.

ROEHRIG, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433–447.

SEVERINI, T. A., AND G. TRIPATHI (2006): "Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors," *Econometric Theory*, 22, 258–278.