NONPARAMETRIC IDENTIFICATION AND ESTIMATION OF TRANSFORMATION MODELS

PIERRE-ANDRÉ 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 allows for endogenous regressors. 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 nonparametrically 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 different normalizations of the model lead to different asymptotic properties of the estimators with one normalization in particular resulting in an estimator for the unknown transformation function that converges at a parametric rate. A test for whether a candidate regressor satisfies the conditional independence assumption required for identification is developed. A Monte Carlo experiment illustrates the performance of our method in the context of a duration model with endogenous regressors.

JEL codes: C14, C26.

Keywords: Nonparametric identification; transformation models; endogeneity; special regressor; kernel estimation.

Date: January 5, 2015. *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, University College London, CeMMAP and CREATES. Email: d.kristensen@ucl.ac.uk

Acknowledgements: We would like to thank the Editor, Jianqing Fan, an Associate Editor, and three anonymous Referees, for their comments and suggestions which greatly improved the manuscript. Earlier versions of this paper were presented at the CAM workshop 2010 at University of Copenhagen, the $(\mathrm{EC})^2$ conference 2010 in Toulouse, 2010 CIRANO-CIREQ conference on "Revealed Preferences and Partial Identification" in Montreal, 2014 conference on non- and semiparametric Econometrics at University of York, and seminars at Boston College, MIT, Princeton, University of Chicago and University of Amsterdam. We thank all the participants for useful comments. All errors are ours. Kristensen gratefully acknowledges research support by the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, the Danish National Research Foundation (through a grant to CREATES) and the European Research Council (grant no. ERC-2012-StG 312474).

1. Introduction

A variety of structural econometric models comes in form of a transformation model, in which a scalar dependent variable Y is related to a vector of regressors X and a scalar unobservable ϵ through

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

The model is characterized by a strictly monotonic transformation T, a regression function g, and a cumulative distribution function (cdf) $F_{\epsilon|X}$ of ϵ given X, all of which are unknown. An important economic application of the model (1) is to the study of duration data (see, e.g., Van den Berg, 2001, for a survey). In this context, dependence between ϵ and some components of X is often a concern, which can arise for a variety of reasons. For instance, if the duration outcome depends on another duration variable with both durations affected by the same unobserved heterogeneity term (Abbring and van den Berg, 2003); or because duration data is only observed for those individuals that comply with some treatment and compliance is not random but selective (Bijwaard and Ridder, 2005); or else in a strategic environment in which durations of two or more players interact with each other (Honore and de Paula, 2010); or because of reverse causality as when duration represents time-to-default and defaults affect regressors such as prices (Palmer, 2014). More generally, omission of relevant regressors or presence of measurement errors might give rise to endogeneity.

We develop novel nonparametric identification results for $\left(T, g, F_{\epsilon|X}\right)$ when some of the regressors X are correlated with ϵ. Our identification strategy is *constructive* in the sense that we obtain explicit expressions of the components in terms of the cdf of Y given X, $F_{Y|X}$. This in turn allows us to develop simple nonparametric estimators of $\left(T, g, F_{\epsilon|X}\right)$ which we analyze. An important feature is that the convergence rate of the estimator of T critically depends on the normalization conditions we impose: The "smoother" the normalization, the faster the estimator converges. To the best of our knowledge, our paper is the first to show that normalization conditions are not innocuous, with different normalization choices leading to nonparametric estimators with radically different properties.[1] When the normalization

[1] At least in the context of nonparametric "plug-in" kernel estimators, which are the ones we propose here. Whether or not the same results obtain for other classes of estimators is an interesting question that we leave for future research.

used for identification of T does not involve derivatives of T, our estimator attains parametric rate. This in turn implies that for inference regarding g and $F_{\epsilon|X}$ we can treat T as known.

The identification argument proceeds in two steps: We first show that $\Theta \equiv T^{-1}$ is identified under the assumption that X can be decomposed into $X = (X_I, X_{-I})$ where the subset of regressors X_I is conditionally exogenous, $\epsilon \perp X_I \mid X_{-I}$. As such X_I play a role similar to the "special regressor" of Lewbel (1998); however, in contrast to his study, we do not require X_I to satisfy any "large-support" conditions. Once Θ has been identified, we can identify g and $F_{\epsilon|X}$ using existing results on nonparametric instrumental variables (IV); see, e.g. Darolles, Fan, Florens, and Renault (2011), and references therein.

The estimation strategy builds upon our identification result where we demonstrate that Θ can be expressed as a functional of $F_{Y|X}$. A pointwise estimator of Θ is then obtained by replacing $F_{Y|X}$ with a nonparametric estimator. Once Θ has been estimated, g can be estimated using, for example, nonparametric IVs with $\hat{\Theta}(Y)$ replacing the unknown dependent variable $\Theta(Y)$. Given the parametric convergence rate of $\hat{\Theta}$, our nonparametric IV estimator of g is asymptotically equivalent to the oracle estimator with Θ known. Having recovered Θ and g, we can compute residuals and use these to estimate $F_{\epsilon|X}$.

The identification and estimation schemes critically rely on the availability of at least one regressor being conditionally exogeneous. If, for a given choice of X_I, this assumption is violated the proposed estimators are inconsistent. It is therefore important to be able to check the validity of a candidate regressor. As part of the identification argument, we derive a set of over-identifying restrictions implied by the conditional independence assumption, which in turn is used to develop a statistical test for it.

We investigate the finite-sample performance of our estimators in a Monte Carlo simulation study designed around a popular duration model. We find that the estimators perform well with moderate biases and variances. Moreover, they appear to be quite robust to the choice of the various smoothing parameters used in their implementation.

Our identification results are close in spirit to those obtained by Ridder (1990) and Ekeland, Heckman, and Nesheim (2004) who focus on exogenous regressors. Fève and Florens (2010) allow for endogenous regressors when g is linear or partially linear using a so-called measurable separability assumption in placeof our conditional exogeneity condition. More in line with our identification strategy, Vanhems and Van Keilegom (2013) allow for endogeneity in a semiparametric version of the model with a finitely parameterized transformation. Finally, Chernozhukov, Imbens, and Newey (2007) and Chen, Chernozhukov, Lee, and Newey (2011) provide identification conditions that allow for endogeneity in a general class of models, including ours. These are, however, only local identification results and rely on high-level assumptions. We complement these papers by providing primitive conditions for global nonparametric identification.

Nonparametric estimators of Θ under exogeneity have been developed in, e.g., Horowitz (1996), Chen (2002) and Jochmans (2011). These require as input an initial parametric estimator of g and are thus difficult to extend to the fully nonparametric case. Matzkin (1991) and Jacho-Chávez, Lewbel, and Linton (2010) develop fully nonparametric estimators. However, the asymptotic properties of the former are still not fully understood, and the latter only achieves nonparametric convergence rate. None of the above papers allow for endogenous regressors. Finally, the sieve estimators developed in Chernozhukov, Imbens, and Newey (2007) and Chen and Pouzo (2012) should in principle be applicable to our model.

The remainder of the paper is organized as follows. Section 2 contains the identification result, while estimators are proposed and analyzed in Section 3. The test for conditional independence is developed and analyzed in Section 4. Section 5 illustrates the performance of the proposed estimators and test through a Monte Carlo experiment. The last section concludes. Additional technical assumptions and proofs are relegated to an Appendix.

2. Identification

2.1. **Model and Assumptions.**

We consider the model in (1) where Y has support $\mathcal{Y} \subseteq \mathbb{R}$, $X = (X_1, \ldots, X_{d_x})$ has support $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, and ϵ belongs to $\mathcal{E} \subseteq \mathbb{R}$. The variables Y and X are observed, while ϵ remains latent. We decompose the regressors into $X = (X_I, X_{-I})$ where the subvector $X_I \in \mathbb{R}^{|I|}$ is assumed to be exogeneous while $X_{-I} \in \mathbb{R}^{d_x - |I|}$ contains the potentially endogenous components. The supports of X_I and X_{-I} are denoted by $\mathcal{X}_I$ and $\mathcal{X}_{-I}$, respectively.

Assumption A1. *For a.e. $x \in \mathcal{X}$, the conditional distribution $F_{\epsilon|X}(\cdot|x)$ of ϵ given $X = x$ is absolutely continuous (with respect to the Lebesgue measure on $\mathbb{R}$) with a density $f_{\epsilon|X}(\cdot|x)$ that is continuous on its support $\mathcal{E}_x \subseteq \mathbb{R}$.*

Assumption A2. *(i) $\epsilon \perp X_I \mid X_{-I}$; (ii) X_I is continuously distributed on $\mathcal{X}_I \subseteq \mathbb{R}$.*

Assumption A1 together with the assumption of T being monotonic implies that $F_{Y|X}$ is absolutely continuous, which is important for our identification argument. Assumption A2(i) formally states that at least one of the regressors is conditionally exogenous. As such, X_I plays a role similar to the "special regressor" of Lewbel (1998). However, no large support conditions are imposed on X_I; this is unlike in Lewbel (1998) who requires the support to be either the entire real line, or else large enough if the supports of X_{-I} and ϵ are bounded. If an exogenous variable is discrete, it can be moved from X_I to X_{-I} since Assumption A2 imposes no restrictions on X_{-I}. Moreover, the assumption of X_I being continuously distributed can be weakened: We know from Ridder (1990), for example, that, in absence of endogeneity, nonparametric identification is possible even if the regressors are discrete. In Appendix D we show that Assumption A2(ii) can be dropped provided, however, additional restrictions are put on the regression function g. This alternative identification strategy is not constructive though, in a sense that it does not lead to a natural nonparametric estimator for Θ. We thus choose not to pursue this approach further.

Next, we put restrictions on the support of Y and the behavior of the transformation T:

Assumption A3. *The support $\mathcal{Y}$ of Y is a connected subset of $\mathbb{R}$ (i.e. an interval) that contains zero.*

Assumption A4. *T is invertible with inverse $\Theta \equiv T^{-1}$ that is increasing and continuously differentiable on $\mathcal{Y}$.*

Since our identification argument will be based on integrating certain partial differential equations w.r.t. y, we need the domain of integration to be an interval in $\mathbb{R}$. This is ensured by Assumption A3. That zero belongs to this interval will be used in our normalization conditions to follow. If needed, zero can be replaced with any other value $y_0 \in \mathcal{Y}$. Assumption A4 requires T to be invertible with a continuously differentiable inverse $\Theta = T^{-1}$ that is increasing on $\mathcal{Y}$ so $\Theta(t) \leq \Theta(v)$ whenever $t \leq v$.

Assumptions A1-A4 have strong implications which we now derive. First, observe that equation (1) can be rewritten as

$$\Theta\left(Y\right)=g(X)+\epsilon. \tag{2}$$

Since $\Theta'\left(y\right)\geq0$, the conditional cdf of Y given X, which we denote by $\Phi\left(y|x\right)\equiv F_{Y|X}\left(y|x\right)$, satisfies:

$$\Phi\left(y|x\right)=F_{\epsilon|X}\left(\Theta\left(y\right)-g(x)|x\right)=F_{\epsilon|X_{-I}}\left(\Theta\left(y\right)-g(x)|x_{-I}\right), \tag{3}$$

for all $y\in\mathcal{Y}$ and $x=(x_I,x_{-I})\in\mathcal{X}$, where the second equality follows from the conditional independence of ϵ and X_I given X_{-I}. Moreover, $\Phi(y|x)$ is absolutely continuous with a continuous density.

The identification argument will rely on the ability to generate variation in Y through X_I while keeping ϵ fixed. Importantly, under our Assumption A2(i), any variation in X_I will only affect Φ through the regression function g. Identification is then achieved through the derivatives of $\Phi\left(y|x\right)$ w.r.t. y and x_I. For these to be well-defined, we impose the following additional smoothness restriction on g:

Assumption A5. *$g(x)$ is continuously differentiable w.r.t. x_I on $\mathcal{X}$.*

Similar to A2(ii), Assumption A5 only restricts the smoothness of $g\left(x\right)$ with respect to x_I. Nothing is being said about the behavior of g with respect to the remaining components x_{-I}. When we analyze the nonparametric estimators of Θ and g, we will however impose additional smoothness conditions on g as a function of x_{-I}.

The identification of Θ will then rely on the following two sets of equations,

$$\Phi_y(y|x) \;=\; \Theta'(y)f_{\epsilon|X_{-I}}(\Theta(y)-g(x)|x_{-I}), \tag{4}$$

$$\Phi_i(y|x) \;=\; -\frac{\partial g(x)}{\partial x_i}f_{\epsilon|X_{-I}}(\Theta(y)-g(x)|x_{-I}),\quad i\in I, \tag{5}$$

where $\Phi_y(y|x)\equiv\partial\Phi(y|x)/\partial y$, $\Phi_i(y|x)\equiv\partial\Phi(y|x)/\partial x_i$, and $\Theta'\left(y\right)$ is the derivative of Θ. In particular, dividing equation (4) by (5) and rearranging,

$$\Theta'(y)=-\frac{1}{\partial g(x)/\partial x_i}s_i(y,x),\quad\text{where }s_i(y,x)\equiv\frac{\Phi_y(y|x)}{\Phi_i(y|x)}, \tag{6}$$

for any $i\in I$ whenever $\Phi_i(y|x)\neq0$. This expression is key to the identification of Θ with the conditional independence assumption A2(i) guaranteeing that $\frac{s_i(y,x)}{\partial g(x)/\partial x_i}$ is constant with respect to x. Equation (6) only holds for pairs (y, x) for which $\Phi_i(y|x) \neq 0$. We therefore impose the following assumption:

Assumption A6. *The set* $\mathcal{A}_i \equiv \{x \in \mathcal{X} : \Phi_i(y, x) \neq 0$ *for every* $y \in \mathcal{Y}\}$ *is nonempty for some* $i \in I$.

The requirement that $\mathcal{A}_i$ is nonempty can be thought of as a generalized rank condition saying that a given exogenous regressor X_i $(i \in I)$ has a causal impact on Y. Equation (5) shows that A6 has two parts: First, we need that for some $i \in I$ there exist an $x \in \mathcal{X}$ such that $\partial g(x)/\partial x_i \neq 0$. This requirement excludes the situation in which g is a constant function of all the exogenous regressors. The requirement is rather weak compared with the specific structure on g imposed in Lewbel (1998). 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 $f_{\epsilon|X}(\Theta(y) - g(x), x_{-I}) > 0$ 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.

It is worth pointing out that the larger the set of exogenous regressors, the easier it is to satisfy A6. The intuition is that we only need one exogenous regressor to generate variability in the regression function g: if X_I has several components it is sufficient that g be a non-constant function of one of them. This highlights the role of having multiple exogenous regressors available.

2.2. **Normalizations and Identification.** It is clear from Equation (2) that some normalization of the model is needed; indeed, for any $\lambda > 0$ and $(\mu, \nu) \in \mathbb{R}^2$, the structure $(\Theta, g, F_{\epsilon|X})$ in (2) is observationally equivalent to the structure $(\tilde{\Theta}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ given by

$$\tilde{\Theta} \equiv \mu + \lambda\Theta, \quad \tilde{g} \equiv \nu + \lambda g, \quad \tilde{\epsilon} \equiv \mu - \nu + \lambda\epsilon. \tag{7}$$

In particular, a location and a scale normalization of Θ is needed to pin down the constants μ and λ. Conditions that ensure $\nu = 0$ shall be imposed later on, when we discuss the identification of $(g, F_{\epsilon|X})$.

To identify Θ, we will impose either of the following two different normalizations:

$$\Theta(0) = 0 \quad \text{and} \quad \Theta'(0) = 1, \tag{N1}$$

$$\Theta(0) = 0 \quad \text{and} \quad \Theta(1) = 1. \tag{N2}$$

The values 0 and 1 at which the normalizations are imposed are without loss of generality; if needed, they can be replaced by any $(y_0, y_1) \in \mathcal{Y}$ with $y_0 \neq y_1$. While both normalizations pin down the location through $\Theta(0) = 0$, they differ in the way they fix the scale. Normalization (N1) fixes the derivative of Θ at a particular point, while (N2) instead constrains the level of Θ at some additional point different from zero. Thus, the two normalizations have increasing degrees of smoothness. The following theorem shows that these different normalizations imply different expressions for the function Θ:

Theorem 1. *Let Assumptions A1-A5 hold. Then, with* $S_i\left(y,x\right) \equiv \int_0^y s_i\left(u,x\right)du$*, the following identification results hold for any regressor* $i \in I$ *that satisfies Assumption A6:*

(i) *under* (N1)*,* Θ *is globally identified as*

$$\Theta(y) = \vartheta_i\left(y,x\right), \quad \vartheta_i\left(y,x\right) \equiv \frac{S_i\left(y,x\right)}{s_i\left(0,x\right)}, \tag{8}$$

and the right-hand side of (8) *does not depend on* i *nor* x*.*

(ii) *under* (N2)*,* Θ *is globally identified as*

$$\Theta(y) = \theta_i\left(y,x\right), \quad \theta_i\left(y,x\right) \equiv \frac{S_i\left(y,x\right)}{S_i\left(1,x\right)}, \tag{9}$$

and the right-hand side of (9) *does not depend on* i *nor* x*.*

The theorem is constructive in the sense that $\vartheta_i\left(y,x\right)$ and $\theta_i\left(y,x\right)$, and thereby $\Theta\left(y\right)$, are functionals of $\Phi\left(y|x\right)$ with the latter being estimable given data of (Y,X). As we shall see, the two estimators for Θ corresponding to the two different normalizations will have radically different asymptotic properties: while the one based on (N1) will converge at a nonparametric rate, parametric rate is achieved by the estimator based on (N2).[2] This to the best of our knowledge is the first time in the literature that a formal study of the effect of normalizations is undertaken which shows that different normalizations can lead to estimators with radically different asymptotic properties. This result warns against the popular belief that a "normalization is innocuous."

The above theorem also highlights the role played by multiple exogenous regressors. As already pointed out, the larger $|I|$, the more likely is Assumption A6 to be satisfied. Put

[2] In Appendix E we consider a yet different "integral" normalization: $\Theta(0) = 0$ and $\int_{\mathcal{Y}} \Theta(y) f_0\left(y\right) = 1$ for some known function f_0. As we would expect from the smoothness of the latter, the corresponding nonparametric estimator retains the parametric convergence rate.

in words, this assumption requires some variation in the conditional distribution of Y given X when the exogenous regressor X_i varies. The more exogenous regressors, the easier it is to obtain the required variation. Given that the identified expression for Θ does not depend on which exogenous regressor is chosen, the presence of multiple X_i's gives rise to over-identifying restrictions which can in principle be used to test correct specification of the transformation model.

Once Θ is identified, we can treat $\Theta(Y)$ as observed and so the remaining task is to identify g and $F_{\epsilon|X}$ from the model (2) given observations of $\Theta(Y)$ and X. This is a standard additive nonparametric regression model, and we can therefore import existing results from the literature on nonparametric identification of regression models with endogenous regressors. Popular identification restrictions put forth in the literature include the existence of a set of instruments combined with either a mean restriction, as in Newey and Powell (2003), or a median restriction $P(\epsilon \leq 0|Z) = 1/2$, as in Horowitz and Lee (2007). One can alternatively take a control function approach, as pursued by Newey, Powell, and Vella (1999). Any of these three approaches will lead to the identification of g and $F_{\epsilon|X}$. We here follow the literature on nonparametric IV and assume the existence of a set of instruments $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ that satisfy standard conditions of this literature.

Assumption A7. *There exists a set of instruments Z such that: (i) $E[\epsilon|Z] = 0$ almost surely (a.s.); (ii) the conditional distribution of X_{-I} given Z is complete: for every function $m : \mathcal{X}_{-I} \to \mathbb{R}$ such that $E[m(X_{-I})]$ exists, $E[m(X_{-I})|Z] = 0$ a.s. implies $m(X_{-I}) = 0$ a.s.*

In the special case where all the regressors are exogenous, i.e. $|I| = d_x$, we can choose $Z = X$ and Assumption A7(i) collapses to $E[\epsilon] = 0$, which is a normalization condition that pins down the location ν of g and ϵ in equation (7). If one is willing to assume $E[\epsilon|X_{-I}] = 0$, which still allows for some dependence between ϵ and X_{-I}, then Assumption A7 holds with $Z = X_{-I}$. As for the exogenous regressors, note that Assumption A7(i) implies that $E[\epsilon] = 0$ which together with the conditional independence restriction A2 yields $E[\epsilon|X_I] = 0$. Thus, the required mean independence restrictions between ϵ and the regressors X are as we would expect from the nonparametric IV literature. If one is willing to impose the additional restriction that g is bounded, the completeness condition A7(ii) can be replaced by the weaker assumption of *bounded* completeness; see Blundell, Chen, and Kristensen (2007).

Corollary 1. *Let all the assumptions of Theorem 1 and Assumption A7(i) hold. Then g and $F_{\epsilon|X}$ are identified if and only if Assumption A7(ii) holds.*

It is worth pointing out that the completeness condition in A7(ii) is both necessary and sufficient for identification of the regression function g and the conditional distribution of the latent term $F_{\epsilon|X}$. In this sense, the restrictions in A7 can be seen as minimal.

3. Estimation

We use the identification results of the previous section to derive explicit estimators of $(T, g, F_{\epsilon|X})$. We will for notational simplicity assume that X_{-I} has a continuous distribution which we then estimate using kernel smoothing techniques. If some of the regressors in X_{-I} have a discrete distribution, the corresponding kernel function used in the nonparametric smoothing should be replaced by an indicator function.

3.1. Estimation of Θ.

Suppose we have a random sample (Y_i, X_i, Z_i) $(i = 1, \ldots, n)$ drawn from the model in Equation (1). Depending on whether we impose the normalization (N1) or (N2), we then build an estimator of $\Theta(y)$ based on Equation (8) or (9), respectively. Consider first the case where (N2) is imposed. For a given exogeneous regressor $i \in \{1, ..., |I|\}$ satisfying Assumption A6, for some weighting function $w(x)$ satisfying $\int_{\mathcal{X}} w(x)\, dx = 1$ with support $\mathcal{X}_w \subseteq \mathcal{A}_i$, and a given bowlshaped loss function L, Theorem 1 implies that $\Theta(y) \equiv \arg\min_{q\in\mathbb{R}} \int_{\mathcal{X}} w(x)\, L(\theta_i(y,x) - q)\, dx$. An estimator of $\Theta(y)$ is now easily obtained: Given some nonparametric "plug-in" estimator of $\theta(y,x)$, $\hat{\theta}_i(y,x)$, we compute

$$\hat{\Theta}(y) \equiv \arg\min_{q\in\mathbb{R}} \int_{\mathcal{X}} w(x)\, L\big(\hat{\theta}_i(y,x) - q\big) dx. \tag{10}$$

The estimator of $\Theta(y)$ for the case where (N1) is imposed is implemented in the same way, except that we replace $\hat{\theta}_i(y,x)$ by $\hat{\vartheta}_i(y,x)$, which again is obtained from a first-step estimator of Φ.

The weighting function w serves three purposes: First, it is used to control for the usual denominator problem present in many semiparametric estimators that involves division by a first-step nonparametric density estimator. Specifically, we will require that the support of w, $\mathcal{X}_w$, has been chosen so that $\inf_{x\in\mathcal{X}_w} f(x) > 0$, where $f(x)$ is the density of X. Second, the support $\mathcal{X}_w$ should only include those values of x that can be used to identify Θ from the variation in the regressor X_i, in the sense that $\inf_{(y,x)\in\mathcal{Y}\times\mathcal{X}_w} |\Phi_i(y,x)| > 0$. Third, w could be used to improve the efficiency of the estimator by reweighing $\hat{\theta}(y,x)$ as a function of x.

Two obvious choices of the loss function L are: the least-squares (LS) loss, $L(q) = q^2$, and the least-absolute deviation (LAD) loss, $L(q) = |q|$. When the normalization (N2) is imposed, these losses lead to the following estimators:

$$\hat{\Theta}^{\mathrm{LS}}(y) \equiv \int_{\mathcal{X}} w(x)\,\hat{\theta}_i(y,x)\,dx, \tag{11}$$

$$\hat{\Theta}^{\mathrm{LAD}}(y) \equiv \arg\min_{q\in\mathbb{R}} \int_{\mathcal{X}} w(x)\left|\hat{\theta}_i(y,x) - q\right| dx.$$

The LS estimator in (11) is similar to the one of Horowitz (1996) in that it involves integrals over derivatives of the conditional cdf Φ. However, Horowitz's estimator takes as input an estimator of $g(x) = \beta' x$, and is therefore based on a very different identification argument. Moreover, since the regression function is assumed to be linear and known, Horowitz's estimator is of a simpler form than ours.

Through simulations, we found that $\hat{\Theta}^{\mathrm{LS}}(y)$ did not always perform well; similar results are found for Horowitz's estimator (see Chen, 2002, for simulation results). More specifically, we find that for x taking values in the tails of the empirical distribution of X, $\hat{\theta}_i(y,x)$ proved to be a poor estimate of $\theta_i(y,x)$. One could in principle handle this issue by choosing the weights $w(x)$ so as to trim away the "extreme" values of x. It proves, however, simpler to instead use the LAD version of the estimator, which is well-known to be less sensitive to "outliers" in $\hat{\theta}_i(y,x)$ as we vary x. This is confirmed in the simulation study where $\hat{\Theta}^{\mathrm{LAD}}(y)$ performs significantly better than $\hat{\Theta}^{\mathrm{LS}}(y)$. To simplify the theoretical analysis, we follow Horowitz (1998) and introduce a smoothed version of the above LAD estimator,

$$\hat{\Theta}_b^{\mathrm{LAD}}(y) \equiv \arg\min_{q\in\mathbb{R}} Q_b(q|\hat{\theta}_i(y,\cdot)), \tag{12}$$

where

$$Q_b(q|\theta(y,\cdot)) \equiv \int_{\mathcal{X}} w(x)\left\{\theta(y,x) - q\right\}\left\{2F_b(\theta_i(y,x) - q) - 1\right\} dx,$$

with $F_b(\cdot) \equiv F(\cdot/b)$ for some cdf F with median at zero and some bandwidth $b > 0$. It is easily seen that $\hat{\Theta}_b^{\mathrm{LAD}}(y) \to \hat{\Theta}^{\mathrm{LAD}}(y)$ as $b \to 0$ for a given sample size. However, as we shall see, $\hat{\Theta}_b^{\mathrm{LAD}}(y)$ is in fact consistent for any fixed value of $b > 0$. This is due to the fact that in large-samples $\hat{\theta}_i(y,x)$ is constant w.r.t. x and so the smoothing over x does not affect it asymptotically. Moreover, $\hat{\Theta}_b^{\mathrm{LAD}}(y)$ proves to be first-order equivalent to $\hat{\Theta}^{\mathrm{LS}}(y)$ so there is no asymptotic efficiency loss from the improved finite-sample performance of $\hat{\Theta}_b^{\mathrm{LAD}}(y)$.

Finally, if Assumption A6 holds for multiple exogenous regressors, say, for a subset $I_0 \subseteq I$, we can compute an estimator for each regressor $i \in I_0$ yielding $\{\hat{\Theta}_i(y) : i \in I_0\}$. These can be combined to obtain a final estimator $\hat{\Theta}(y) = \sum_{i \in I_0} \tilde{w}_i\left(y\right) \hat{\Theta}_i\left(y\right)$ using another set of weighting functions $\left\{\tilde{w}_i\left(y\right)\right\}_{i \in I_0}$ satisfying $\sum_{i \in I_0} \tilde{w}_i\left(y\right) = 1$. As with GMM-type estimators, given the (asymptotic) covariance structure of the estimators, $\{\hat{\Theta}_i\left(y\right)\}_{i \in I_0}$, the weights $\left\{\tilde{w}_i\left(y\right)\right\}_{i \in I_0}$ can be chosen to obtain (pointwise) efficiency. This highlights another advantage of having multiple exogenous regressors: These can be used to improve efficiency of the estimator of $\Theta\left(y\right)$.

We now derive the large-sample properties of the LS and smoothed LAD estimators defined in Equations (11) and (12), respectively. We first analyze the version based on the normalization (N2) and then discuss the one based on (N1). For notational convenience, we hereafter assume that Assumption A6 holds with $i = 1$ so that we can drop the subindex $i \in I$ that keeps track of which exogenous regressor is being employed in the estimation. In particular, we set $S\left(y, x\right) = S_1\left(y, x\right)$, $\theta\left(y, x\right) = \theta_1\left(y, x\right)$, and so forth.

The specific estimator of $\theta\left(y, x\right)$ will be based on a kernel estimator of $\Phi\left(y|x\right)$. In principle, any nonparametric estimator could be employed, but kernel estimators are computationally very easy to implement and so we focus on this class of estimators in the following. To define the estimator, first observe that the conditional cdf of Y given X can be written as $\Phi\left(y|x\right) = p\left(y, x\right)/f\left(x\right)$ where

$$p\left(y, x\right) \equiv \int_{-\infty}^{y} f_{Y,X}\left(u, x\right) du, \quad f\left(x\right) \equiv \int_{\mathcal{Y}} f_{Y,X}\left(u, x\right) du,$$

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

$$\hat{\Phi}\left(y, x\right) = \frac{\hat{p}\left(y, x\right)}{\hat{f}\left(x\right)}, \tag{13}$$

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

with $\mathcal{K}_{h_y}\left(y\right) = \mathcal{K}\left(y/h_y\right)$, $\mathbf{K}_{h_x}\left(x\right) = \mathbf{K}\left(x/h_x\right)/h_x^{d_x}$, and where $h_x, h_y > 0$ are two univariate bandwidths. The functions $\mathcal{K}\left(y\right)$ and $\mathbf{K}\left(x\right)$ are given as $\mathcal{K}\left(y\right) = \int_{-\infty}^{y} K\left(u\right) du$ and $\mathbf{K}\left(x\right) = \prod_{i=1}^{d_x} K\left(x_i\right)$ with $K : \mathbb{R} \to \mathbb{R}$ 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. Note that we use $\mathcal{K}_{h_y}\left(Y_i - y\right)$ instead of the indicator function $\mathbb{I}\left\{Y_i \leqslant y\right\}$ in the estimation since we will need $\hat{\Phi}$ to be differentiable w.r.t.

y. The estimator $\hat{\Phi}$ can then be used to estimate $\theta(y,x)$ by

$$\hat{\theta}(y,x)=\frac{\hat{S}(y,x)}{\hat{S}(1,x)},\quad \text{where } \hat{S}(y,x)=\int_0^y \frac{\hat{\Phi}_y(u,x)}{\hat{\Phi}_1(u,x)}du.$$

For the analysis of the estimators of $\Theta(y)$, we introduce additional assumptions on the model, the kernel function K and the weighting function w:

Assumption A8. *The univariate kernel K is differentiable, and there exists constants $C,\eta>0$ such that $\left|K^{(i)}(z)\right|\leq C|z|^{-\eta}$, $\left|K^{(i)}(z)-K^{(i)}(z')\right|\leq C|z-z'|$, for $i=0,1$, where $K^{(i)}(z)$ denotes the ith derivative. Furthermore, $\int_{\mathbb{R}}K(z)dz=1$, $\int_{\mathbb{R}}z^jK(z)dz=0$, $1\leq j\leq m-1$, and $\int_{\mathbb{R}}|z|^mK(z)dz<\infty$.*

Assumption A9. *The weighting function $w(x)$ is continuously differentiable with compact support $\mathcal{X}_w\subseteq\mathcal{A}_1$ which has non-empty interior.*

Assumption A10. *The joint density $f_{Y,X}(y,x)$ is bounded and m times differentiable with bounded derivatives; its m^{th} order partial derivatives are uniformly continuous. Furthermore, $\inf_{x\in\mathcal{X}_w}f(x)>0$.*

Assumption A11. *$\sqrt{n}h_x^m\to0$, $\sqrt{n}h_y^m\to0$, $\sqrt{n}h_x^{d_x+2}/\log n\to\infty$, and $\sqrt{n}h_yh_x^{d_x+1}/\log n\to\infty$.*

The class of kernels in Assumption A8 is fairly general and accommodates 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 needed to ensure that $\hat{\Phi}_1$ is well-defined. We allow for both standard second-order kernels ($m=2$) such as the Gaussian one, and higher-order kernels ($m>2$). Assumption A9 puts restrictions on the weighting function $w(x)$ in terms of its support $\mathcal{X}_w$. In particular, it ensures that $\inf_{(y,x)\in\mathcal{Y}\times\mathcal{X}_w}|\Phi_1(y,x)|>0$.

The use of higher-order kernels in conjunction with smoothness conditions on the densities stated in Assumption A10 allows us to control smoothing biases. In general, the kernel has to be of higher order in order for $\hat{\Theta}(y)$ to be $\sqrt{n}$-consistent. Note that the number of derivatives in A10, $m\geq2$, is assumed to match up with the order of the kernel K. The lower bound imposed on $f(x)$ allows us to avoid any denominator problems in the proofs, and allows us to establish uniform convergence of $\hat{S}(y,x)$ over $\mathcal{Y}\times\mathcal{X}_w$.

Finally, assumption A11 restricts the set of feasible bandwidths to ensure 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,x). As is standard for kernel estimators, there is a curse-of-dimensionality which appears in the last two restrictions on h_x: When $d_x = \dim(X)$ 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 $m > (d_x+2)/2$ and $1/(4m) < r_x, r_y < 1/[2(d_x+2)]$.

The analysis of the estimator proceeds along the same lines as for two-step semiparametric estimators. We first linearize the LS estimator w.r.t. the first-step estimator $\hat{S}(y,x)$ to obtain

$$\begin{aligned}\hat{\Theta}^{\mathrm{LS}}(y) - \Theta(y) = & \int_{\mathcal{X}} \frac{w(x)}{S(0,x)}\left[\hat{S}(y,x) - S(y,x)\right]dx \\ & - \int_{\mathcal{X}} \frac{w(x)S(y,x)}{S^2(0,x)}\left[\hat{S}(0,x) - S(0,x)\right]dx + o_P\left(n^{-1/2}\right).\end{aligned} \tag{14}$$

While $\hat{S}(y,x)$ does not converge with $\sqrt{n}$-rate, the integration over x speeds up the convergence rate and we show that each of the two integrals converges with $\sqrt{n}$-rate towards Normal distributions. This yields the following result:

Theorem 2. *Let Assumptions A1 through A11 and the normalization condition* (N2) *hold. Then, for any $b > 0$, the following functional weak convergence results hold over any compact set $\mathcal{Y}_0 \subseteq \mathcal{Y}$:*

$$\sqrt{n}(\hat{\Theta}^{LS}(y) - \Theta(y)) \Rightarrow \mathbb{W}(y), \quad \sqrt{n}(\hat{\Theta}_b^{LAD}(y) - \Theta(y)) \Rightarrow \mathbb{W}(y),$$

where $y \mapsto \mathbb{W}(y)$ is a zero-mean Gaussian process with covariance kernel $\Omega(y_1,y_2) = E[\delta_i^w(y_1)\delta_i^w(y_2)]$, and $\delta_i^w(y)$ is as defined in Equation (34) *in Appendix B.*

The large-sample variance of the estimators is determined by $\delta_i^w(y)$. Though somewhat complicated, $\delta_i^w(y)$ is a known functional of the weighing function $w(x)$ and the conditional cdf $\Phi(y|x)$ of Y given X. Thus, replacing Φ with $\hat{\Phi}$ in the definition of $\delta_i(w)$ leads to a consistent estimator $\hat{\delta}_i(y)$ of $\delta_i(y)$,[3] which in turn allows us to consistently estimate the

[3] 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 complicated problem outside the scope of this paper.

asymptotic covariance kernel using

$$\hat{\Omega}\left(y_1, y_2\right) \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}_i(y_2)\hat{\delta}_i(y_2).$$

An interesting feature of the smoothed LAD estimator is that its first-order asymptotic properties are invariant to the choice of bandwidth b which can be kept fixed as the sample size grows. This is different from the analysis in Horowitz (1998) who has to restrict b to shrink at a suitable rate to eliminate smoothing biases. The reason for this discrepancy is that in the limit $\hat{\theta}\left(y, x\right)$ is constant with respect to x and so the effect of smoothing is asymptotically negligible. In practice, the LAD estimator will be affected by the bandwidth choice but the impact should be small.

Next, consider the estimator of $\Theta\left(y\right)$ based on the alternative normalization (N1). Following the same proof strategy as for the previous estimator we obtain that

$$\begin{aligned} \hat{\Theta}^{\mathrm{LS}}(y) - \Theta\left(y\right) = & \int_{\mathcal{X}} \frac{w\left(x\right)}{s\left(0, x\right)} \left[\hat{S}\left(y, x\right) - S\left(y, x\right)\right] dx \\ & - \int_{\mathcal{X}} \frac{w\left(x\right) S\left(y, x\right)}{s^2\left(0, x\right)} \left[\hat{s}\left(0, x\right) - s\left(0, x\right)\right] dx + o_P\left(n^{-1/2}\right). \end{aligned}$$

Compare this to Equation (14), and note that the first term still involves $\hat{S}\left(y, x\right)$ and so by the same arguments as before the first integral converges with $\sqrt{n}$-rate. However, the second term now involves $\hat{s}\left(0, x\right) = \partial \hat{S}\left(y, x\right) / \left(\partial y\right)\Big|_{y=0}$ instead of $\hat{S}\left(0, x\right)$, which is due to the fact that the second normalization in (N1), $\Theta'\left(0\right) = 1$, involves the derivative of $\Theta\left(y\right)$ instead of its level. As is well-known, derivatives are harder to estimate nonparametrically and so the second integral in the last expression only converges with rate $\sqrt{nh_y}$. Thus, the estimator based on the normalization (N1) does not attain the parametric rate.

3.2. **Estimation of g.** Once $\hat{\Theta}\left(y\right)$ has been computed, the regression function and the conditional cdf of the error term can be estimated using nonparametric IV techniques: First, suppose that Θ is known. Then, $\Theta(Y) = g(X) + \epsilon$ where $\Theta(Y)$ and X are observed and $E\left(\epsilon \mid X_1, Z\right) = 0$; thus, estimation of g is a standard nonparametric IV regression problem. In this case, we can now in principle employ any existing nonparametric IV estimator proposed in literature such as the kernel estimator of Hall and Horowitz (2005) or the sieve estimator of Blundell, Chen, and Kristensen (2007). We here focus on sieve estimators since these are computationally very simple to implement as explained in Blundell, Chen, and

Kristensen (2007); we expect the following theoretical results derived for the sieve estimators to carry over to alternative estimators. The oracle sieve estimator, assuming Θ is known, takes the form

$$\tilde{g} \equiv \arg\min_{g_n \in \mathcal{G}_n} \sum_{i=1}^{n} \{\tilde{h}\left(X_{1,i}, Z_i\right) - \hat{M}\left(X_{1,i}, Z_i | g_n\right)\}^2, \tag{15}$$

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

$$\begin{aligned} h\left(x_1, z\right) &\equiv E\left[\Theta\left(Y\right)|X_1 = x_1, Z = z\right] \\ M\left(x_1, z|g_n\right) &\equiv E\left[g_n\left(X\right)|X_1 = x_1, Z = z\right], \end{aligned} \tag{16}$$

and $\mathcal{G}_n$ is a sieve space. We have here left out the weighting function used in Blundell, Chen, and Kristensen (2007) for simplicity. Since Θ is unknown, we replace Θ by the first-step estimator:

$$\hat{g} \equiv \arg\min_{g_n \in \mathcal{G}_n} \sum_{i=1}^{n} \{\hat{h}\left(X_{1,i}, Z_i\right) - \hat{M}\left(X_{1,i}, Z_i | g_n\right)\}^2, \tag{17}$$

where $\hat{h}\left(x_1, z\right)$ is a first-step nonparametric estimator of $E[\hat{\Theta}\left(Y\right)|X_1 = x_1, Z = z]$.

Finally, given $\hat{\Theta}\left(y\right)$ and $\hat{g}\left(x\right)$, we can compute the corresponding residuals, $\hat{\epsilon}_i = \hat{\Theta}\left(Y_i\right) - \hat{g}\left(X_i\right)$, $i = 1, \ldots, n$. 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,

$$\hat{F}_{\epsilon|X_{-I}}\left(t, x_{-I}\right) \equiv \frac{\sum_{i=1}^{n} \mathcal{K}_{h_\epsilon}\left(\hat{\epsilon}_i - t\right) \mathbf{K}_{h_x}\left(X_{-I,i} - x_{-I}\right)}{\sum_{i=1}^{n} \mathbf{K}_{h_x}\left(X_{-I,i} - x_{-I}\right)}.$$

As a first step in the analysis of $\hat{g}$, 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 (15) is consistent; these are straightforward but rather technical extensions which we relegate to the Appendix. We note that the most substantive of these additional assumptions is the requirement of compact support of (X_I, Z). In addition, we also impose the restriction that $\mathcal{Y}$ is bounded so that we can choose the set $\mathcal{Y}_0$, over which we showed uniform convergence of $\hat{\Theta}\left(y\right)$, equal to $\mathcal{Y}$ in the following:

Assumption A12. *The support of Y is bounded so that $\mathcal{Y} = (y_l, y_u)$ where $-\infty < y_l < y_u < +\infty$. Moreover, $\lim_{y \to y_l} f_{Y|X}\left(y|x\right) = \lim_{y \to y_u} f_{Y|X}\left(y|x\right) = 0$ for all $x \in \mathcal{X}_w$.*

The second part of the last assumption is a technical one which ensures that the kernel estimators of the conditional density does not suffer from boundary biases. This could be removed, but we would then need to employ boundary kernels in the first-step estimation of Φ. Theorem 2 now yields that $\hat{\Theta}^{\mathrm{LS}}(y)$ and $\hat{\Theta}_b^{\mathrm{LAD}}(y)$ both converge uniformly over $\mathcal{Y}$ with rate $O_P(1/\sqrt{n})$. This in turn enables us to show that the feasible estimator $\hat{g}$ is asymptotically equivalent to $\tilde{g}$, thereby yielding the following result:

Theorem 3. *Let Assumptions A1 through A12 and the normalization condition (N2) hold. Assume in addition that Assumptions A14 through A18 in Appendix A 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:

$$\tau_n\equiv\sup_{g_n\in\mathcal{G}_n:g_n\neq 0}\frac{\sqrt{E\{g_n(X)\}^2}}{\sqrt{E\{E[g_n(X)|X_I,Z]\}^2}}. \tag{18}$$

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 $\{\mu_k\}$ of the conditional mean operator $g\mapsto M(x_I,z|g)$ defined in Equation (16); see Section 4 in Blundell, Chen, and Kristensen (2007) for further discussion. If for example, the singular values satisfy $\mu_k\asymp k^{-s/d_x}$, for some $s>0$ then $\tau_n\leq \text{const}\times k_n^{s/d_x}$ 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 oracle estimator $\tilde{g}$ that assumes knowledge of Θ; thus, there is no (asymptotic) loss from not knowing Θ in the estimation of g. This is due to the fact that $\hat{\Theta}$ 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. As shown there, the rate of convergence towards an asymptotic distribution is slower than $\sqrt{n}$, and so the asymptotic distribution is unaffected by the first-step estimation of Θ.

We conjecture that Theorem 3 remains true without restricting $\mathcal{Y}$ to be bounded. By inspection of the proof of Theorem 3, it is easily checked that the theorem holds as long as $||\hat{\Theta}-\Theta||_Y=o_P\left(n^{-r/[2(r+s)+1]}\right)$, where $\|\cdot\|_Y$ denotes 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, but in order to establish this result we would need to introduce trimming of $\hat{\Theta}$ to control for denominator issues that usually arise when deriving convergence results over unbounded sets. In addition, our current set of assumptions and proofs will become more complicated since we need to control the tail behavior of Θ.

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. A full analysis of the asymptotic properties of $\hat{F}_{\epsilon|X}(t, x_{-1})$ is outside of the scope of this paper. We expect that the techniques developed in Mammen, Rothe, and Schienle (2012) could be adapted to our setting and thereby allow for a more complete analysis of $\hat{F}_{\epsilon|X}(t, x_{-1})$. This is left for future research.

4. Testing Exogeneity

The identification and estimation results developed in the two previous sections rest on two fundamental assumptions regarding the chosen "special" regressor X_i: First, X_i needs to be relevant in a sense that $\partial g(x)/\partial x_i \neq 0$; and second it needs to be exogenous in a sense that:

$$H_0 : \epsilon \perp X_i \mid X_{-i}.$$

If either of these two restrictions is violated, the proposed estimator will in general be inconsistent. It is therefore of interest to develop tools to examine whether a candidate regressor indeed satisfies these assumptions. Regarding the first hypothesis, note that $\partial g(x)/\partial x_i = 0$ if and only if $\Phi_i(y|x) = 0$ for all $y \in \mathcal{Y}$. Given our nonparametric estimator of $\Phi_i(y|x)$, this restriction can be formally tested using standard tools. We therefore in the following focus on the exogeneity condition H_0.

Taking as maintained hypothesis that the transformation model in (1) is correct, the following testable implications of the exogeneity assumption H_0 obtain:

Theorem 4. *Let Assumptions A1, A2(ii), A3, A4, and A5 hold. Then, for any index $i \in I$ that satisfies Assumption A6 and any $x \in \mathcal{A}_i$, the following testable implications hold:*

(i) *under* (N1), $\vartheta_i(y, x) = \Theta(y)$ *for every* $y \in \mathcal{Y}$ *if and only if* H_0 *holds;*

(ii) *under* (N2), $\theta_i(y, x) = \Theta(y)$ *for every* $y \in \mathcal{Y}$ *if and only if* H_0 *holds.*

In general, testing exogeneity of a regressor requires the availability of an instrument to generate overidentifying restrictions; this is, for example, the case in the Hausmann test. In our case, since we require conditional strict independence instead of just conditional mean independence, the maintained model assumption generates overidentifying restrictions that allow us to test H_0 without the use of additional instruments.

The above theorem suggests a natural test for exogeneity by comparing estimators of $\vartheta_i(y,x)$ and $\Theta(y)$ as obtained under the null. As in the section on estimation, we focus for notational simplicity on testing for exogeneity of X_1 in the following and drop the regressor index $i=1$, and so will, for example, write $\hat{\theta}\left(y,x\right)$ for $\hat{\theta}_1\left(y,x\right)$. Moreover, we only consider the case where we the normalization (N2) has been imposed; the testing procedure is easily adapted to the case of (N1), and we expect that the theoretical results derived under (N2) carry over to (N1) with only minor adjustments.

To allow for added flexibility in the testing procedure, we will use two different sets of bandwidths, (h_x, h_y) and $(h_{0,x}, h_{0,y})$, for the estimation of $\Theta\left(y\right)$ and $\theta\left(y,x\right)$, respectively. We will then restrict (h_x, h_y) so that $\hat{\Theta}\left(y\right)$ converges with $\sqrt{n}$-speed. This in turn ensures that the asymptotic distribution of our test statistic will be determined by the nonparametric estimator of $\theta\left(y,x\right)$ alone. To emphasize that different bandwidths are used, we use $\hat{\theta}_0\left(y,x\right)$ to denote the estimator based on $(h_{0,x}, h_{0,y})$. We then propose to compare the two nonparametric estimators through the following L_2-statistic,

$$Q \equiv \int_{\mathcal{Y}}\int_{\mathcal{X}} W\left(y,x\right)[\hat{\theta}_0\left(y,x\right) - \hat{\Theta}\left(x\right)]^2 dydx, \tag{19}$$

where $W\left(y,x\right)$ is a weighting function with compact support satisfying $\int_{\mathcal{Y}}\int_{\mathcal{X}} W\left(y,x\right)dydx = 1$. We will reject H_0 if Q is “large.”

The test based on Q is related to standard nonparametric misspecification tests where a “parametric” estimator, $\hat{\Theta}\left(x\right)$, is compared with a nonparametric one, $\hat{\theta}_0\left(y,x\right)$; see e.g. Härdle and Mammen (1993) and Kristensen (2011). However, in comparison to these papers, the asymptotic analysis of Q is complicated by the fact that the nonparametric estimator $\hat{\theta}_0\left(y,x\right)$ is more complicated compared to the kernel regression and density estimators considered in these two papers. The test is also similar to the specification test proposed in Lewbel, Lu, and Liangjun (2013). However, Lewbel, Lu, and Liangjun (2013) maintain the assumption of exogenous regressors and then wish to test for the functional form restrictions implied by a transformation model. In our case, we maintain the functional form and wish to test for exogeneity of X_1.

For the asymptotic analysis, we restrict the set of feasible bandwidths used in the computation of $\hat{\theta}_0\left(y,x\right)$ to satisfy:

Assumption A13. $nh_{0,x}^{d_x+2}\to\infty$, $nh_{0,x}^{d_x/2+2+m}\to 0$, $nh_{0,x}^{d_x/2+2+4m}\to 0$, $nh_{0,x}^{d_x/2+2}h_{0,y}^{4m}\to 0$, $nh_{0,x}^{3/2d_x}/\left(\log\left(n\right)\right)^2\to\infty$, $nh_{0,y}^{2}h_{0,x}^{3/2d_x-2}/\left(\log\left(n\right)\right)^2\to\infty$.

The asymptotic distribution of the test statistic in (19) is then as follows:

Theorem 5. *Assumptions A1, A2(ii), A3, A4, and A5 hold, and the bandwidths for $\hat{\Theta}\left(y\right)$ and $\hat{\theta}_0\left(y,x\right)$ satisfy A11 and A13 respectively. Then, under H_0,*

$$nh_{0,x}^{d_x/2+2}\frac{Q-m_Q}{v_Q}\to^d N\left(0,1\right),$$

where, with $\mathbf{K}_1\left(x\right)=\partial\mathbf{K}\left(x\right)/\left(\partial x_1\right)$, $\sigma_k^2\left(y,x\right)\equiv Var\left(\bar{D}_k\left(y,Y_i,X_i\right)|X_i=x\right)$, and $\bar{D}_k\left(y,Y_i,X_i\right)$ $(k=1,2)$ defined in Equation (39) in Appendix B,

$$\begin{aligned} m_Q &= \frac{1}{nh_{0,x}^{d_x}}\int\mathbf{K}^2\left(x\right)dx\times\int\int\sigma_1^2\left(y,z\right)W\left(y,x\right)f\left(x\right)dydx \\ &\quad+\frac{1}{nh_{0,x}^{d_x+2}}\int\mathbf{K}_1^2\left(x\right)dx\times\int\int\sigma_2^2\left(y,z\right)W\left(y,x\right)f\left(x\right)dydx, \end{aligned}$$

$$v_Q=2\int\left[\mathbf{K}_1*\mathbf{K}_1\right]^2\left(x\right)dx\times\int\int\sigma_2^4\left(y,x\right)W^2\left(y,x\right)f^2\left(x\right)dydx.$$

If H_0 does not hold, then $nh_{0,x}^{d_x/2+2}\left|\frac{Q-m_Q}{v_Q}\right|\to+\infty$.

The above result is similar to the ones in Härdle and Mammen (1993) and Kristensen (2011) except that the expressions of the location and scale parameters, m and v^2, are somewhat more involved. We propose to use subsampling in order to implement the test as also advocated in Lewbel, Lu, and Liangjun (2013) who provide Monte Carlo evidence of that this procedure leads to good size and power properties for their test; we expect the same to hold true for our related test.

5. Monte Carlo Application to Duration Models

We here illustrate how the proposed identification and estimation strategy can be used in the study of duration models, and provide Monte Carlo results for estimators and tests in this context.

5.1. **Identification of Duration Models under Endogeneity.** First, we recall some basic facts about duration models. Let $\tau \in (0, +\infty)$ denote the duration, $X \in \mathcal{X}$ be a vector of observed covariates, $U \in (0, +\infty)$ an unobserved individual heterogeneity term, and $H(t, x, u)$ denote the conditional hazard function:

$$H(t,x,u) \equiv \lim_{dt \to 0} \frac{P\left(t \leqslant \tau < t+dt \middle| X = x, U = u\right)}{dt}$$

We assume that both X and U are time-invariant, in which case the integrated conditional hazard is distributed as a unit exponential random variable, i.e. for a.e. $(x, u) \in \mathcal{X} \times (0, +\infty)$ we have $\xi \equiv \int_0^\tau H(t,x,u)dt \sim \text{Exp}(1)$. In the mixed proportional hazard model, $H(t, x, u) = H_0(t) \exp[-\phi(x)]u$ where $H_0(t) > 0$ is the baseline hazard. The corresponding log-integrated conditional hazard transform, $\lambda(t) \equiv \ln \int_0^t \theta_0(s)ds$, can be expressed as

$$\lambda(\tau) = \phi(X) + \ln \xi - \ln U. \tag{20}$$

Note that λ satisfies $\lambda'(t) > 0$ and $\lim_{t \to 0} \lambda(t) = -\infty$ and $\lim_{t \to +\infty} \lambda(t) = +\infty$. The model can be written on the form (1) by defining $Y \equiv \tau - 1$, $\Theta\left(y\right) \equiv \lambda\left(y+1\right)/\sigma$, $g\left(x\right) \equiv \phi\left(x\right)/\sigma$ and $\epsilon \equiv \left(\ln \xi - \ln U\right)/\sigma$ where $\sigma \neq 0$ is a scale parameter. The normalization $\Theta\left(0\right) = 0$ then amounts to setting $\lambda(1) = 0$ (which normalizes the baseline hazard to $\int_0^1 H_0(s)ds = 1$), and $E[\ln U] = \mathrm{e}$ where $\mathrm{e} \approx 0.577$ denotes Euler's constant. The scale parameter is chosen as $\sigma = \lambda'\left(1\right)$ if we impose (N1), while if (N2) is imposed then $\sigma = \lambda\left(2\right)$. Thus, up to the scale parameter σ, we can nonparametrically estimate the hazard rate model using the techniques developed in the previous section. The estimators of the normalized hazard rate function and regression function, $\Theta\left(\tau\right)$ and $g\left(x\right)$, are entirely new and not yet seen in the literature.

Once Θ, g and F_ϵ have been estimated, we can estimate σ along the same lines as in Horowitz (1999): If X is exogenous, we can follow Horowitz (1999) and obtain that $\sigma = \lim_{t \to 0} \sigma\left(t\right)$, $\sigma\left(t\right) = -\int G_v\left(t|v\right) p^2\left(v\right) dv \, / \int G\left(t|v\right) p^2\left(v\right) dv$, where $G\left(t|v\right) = P\left(\tau \leq t|V = v\right)$ and $p\left(v\right)$ is the density of $V \equiv g\left(X\right)$. If X_{-1} is endogenous, the above identification result is no longer valid. Instead, we can use that $\Phi\left(t,x\right) = 1 - \int \exp\left[-\Lambda\left(t\right) e^{-\sigma g(x) - u}\right] dF_{\ln U|X_{-1}}\left(u|x_{-1}\right)$ and $\Phi_1\left(t,x\right) = -\sigma\Lambda\left(t\right) g_1\left(x\right) \int \exp\left[-\Lambda\left(t\right) e^{-\sigma g(x) - u}\right] dF_{\ln U|X_{-1}}\left(u|x_{-1}\right)$. In particular, we can then express the scale as $\sigma = \lim_{t \to 0} \sigma\left(t\right)$, where $\sigma\left(t\right) = -\int \Phi_1\left(t,x\right)/g_1\left(x\right) f^2\left(x\right) dx \, / \int \Phi\left(t,x\right) f^2\left(x\right) dx$. This appears to be a new identification result which should be of independent interest.

5.2. **Monte Carlo Results.** For the Monte Carlo study, we generate data from (20) with $X = (X_1, X_2)$ being bivariate and generated as $X_1 = \nu_1$, $X_2 = \alpha_1 Z + \alpha_2 Z^2 + \nu_2 + \rho\epsilon$, where $(\nu_1, \nu_2, \epsilon, Z)$ are mutually independent standard normal random variables. Thus, $X_1 \perp \epsilon$ is exogenous while X_2 remains endogenous whenever $\rho \neq 0$. We consider both the case of exogenous regressors ($\rho = 0$) and endogenous ones ($\rho = 0.5$). Finally, the regression function is specified as $\phi(X) = \beta_1 \Phi(X_1) + \beta_2 X_1 + \beta_3 X_2^2$, with $(\beta_1, \beta_2, \beta_3) = (2.0, 0.1, -0.1)$, while $\lambda(t)$ is chosen as $\lambda(t) = \log(t)$ corresponding to a proportional hazard duration model with a Weibull baseline hazard. In the estimation, we impose the following normalization: $\Theta(0) = 0$ and $\int_{\mathcal{Y}} \Theta(y) f_0(y)\, dy = 1$ for some known density $f_0(y)$. By following the same arguments as used in the proof of Theorem 1, we obtain $\Theta(y) = \theta(y, x)$ for all x, where $\theta(y, x) := S(y, x) / \int_{\mathcal{Y}} S(y, x) f_0(y)\, dy$.

For the implementation of the estimators, we have to choose the bandwidths used to estimate $\Phi(y, x)$ and its derivatives together with a weighting function w. In addition, for the computation of $\hat{\Theta}(y)$, we have to numerically evaluate the integrals that enter the expression of our estimator. The bandwidths are chosen by first implementing Silverman's Rule-of-thumb and then scaling these down since our theoretical results state that we should undersmooth in order to obtain $\sqrt{n}$-consistency. To be more specific, our bandwidths for the two estimators are chosen as follows:

$$\hat{\Phi}_y(y, x) : h_y = (4/3)^{1/5} \hat{\sigma}_Y n^{-(1+\delta)/5}, \quad h_{x_k} = \hat{\sigma}_k n^{-(1+\delta)/6}, \ k = 1, 2. \tag{21}$$

where $\hat{\sigma}_Y^2$ and $\hat{\sigma}_{x_k}^2$ are the sample variances of Y and X_k ($k = 1, 2$), and δ controls the degree of undersmoothing; we set $\delta = 1$. Next, the support of the weighting function was chosen as the uniform density with support $\mathcal{X}_w$ chosen in data-driven way to avoid the aforementioned denominator issues,

$$\mathcal{X}_w = \left\{ (x_1, x_2) : \hat{\Phi}_x\left(\bar{Y}, x\right) > c, \quad \hat{q}_{X_k}(2.5) \leq x_k \leq \hat{q}_{X_k}(97.5) \quad k = 1, 2 \right\},$$

where $\bar{Y}$ is sample mean of Y and $\hat{q}_{X_k}(\cdot)$ the sample quantile function of X_k, $k = 1, 2$.

The estimators were then implemented as follows: First, simulate $N \geq 1$ uniform bivariate draws on $\mathcal{X}_0$, say $x_i^* = \left(x_{i,1}^*, x_{i,2}^*\right)$ for $i = 1, \ldots, N$ and compute $\hat{\theta}_i^*(y) = \hat{S}(y, x_i^*) / \int_{\mathcal{Y}} \hat{S}(y, x_i^*) f_0(y)\, dy$ for each draw. Given these S alternative estimators evaluated at randomly chosen values of x across $\mathcal{X}_0$, we then computed the "empirical" mean, $\hat{\Theta}^{\mathrm{LS}}(y) = \sum_{i=1}^{N} \hat{\theta}_i^*(y)/N$, and the smoothed empirical median, $\hat{\Theta}_b^{\mathrm{LAD}}(y)$; for the latter, we chose a bandwidth of $b = 0.01$.

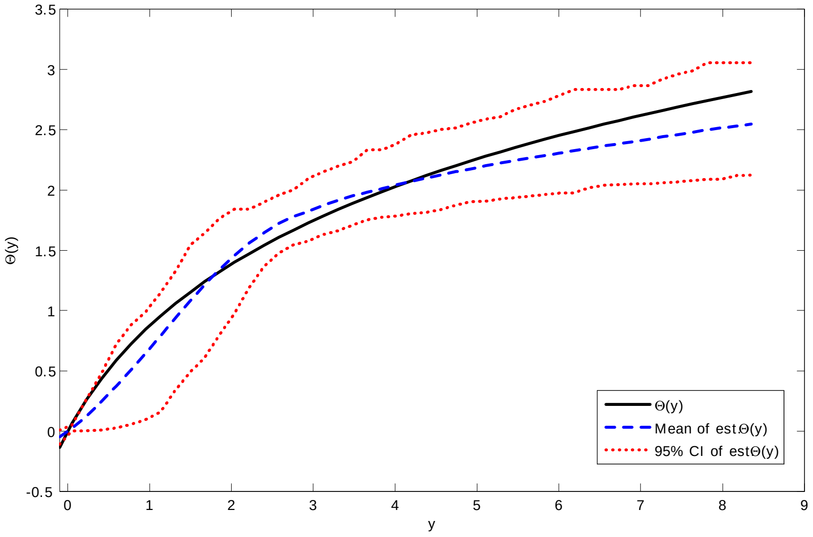

Figure 1: Median-estimator of $\Theta\left(y\right)$ with exogenous regressors.

The performance of $\hat{\Theta}_{b}^{\text{LAD}}\left(y\right)$ for the exogenous and endogenous case are reported in Figures 1 and 2 respectively. We see that the estimator performs very well, with little bias and variance for the most part of the domain of Y, despite the fact that the estimator is based on only $n=250$ observations. As such, the attractive properties asymptotic properties of the estimator appear to also hold in finite samples. Moreover, there is only small differences in the performance of the estimator when comparing the exogenous and endogenous case.

Finally, Figure 3 shows the performance of the least-squares version of our estimator, $\hat{\Theta}^{\text{LS}}\left(y\right)$, for the case of endogenous regressors. The performance of $\hat{\Theta}^{\text{LS}}\left(y\right)$ is clearly inferior to $\hat{\Theta}_{b}^{\text{LAD}}\left(y\right)$ as shown in Figure 3. The poor performance is due to the fact that $\hat{\theta}_{s}^{*}\left(y\right)$, $s=1,\ldots,S$, contain a relatively large number of "outliers" which here are given equal weight. In contrast, the LAD estimator discards these outliers and so is not affected.

Next, we analyze the sensitivity of the estimators to bandwidth choice. To this end, we kept the same design as before, and then re-computed the LAD estimator of $\Theta\left(y\right)$ with bandwidth chosen as (i) $h_{y}^{u}=1.2\times h_{y}$ and $h_{x_{k}}^{u}=1.2\times h_{x_{k}}$, and (ii) $h_{y}^{l}=0.8\times h_{y}$ and $h_{x_{k}}^{l}=0.8\times h_{x_{k}}$, where h_{y} and $h_{x_{k}}$ are given in Equation (21). Thus, we first increrase the bandwidths by 20% ("oversmoothing") and then decrease them by 20% ("undersmoothing") relative to the benchmark reported above. In Table 1, we report the integrated bias, variance and mean-square-error (MSE) of $\hat{\Theta}_{b}^{\text{LAD}}\left(y\right)$ for the bandwidth choice in Equation (21) and

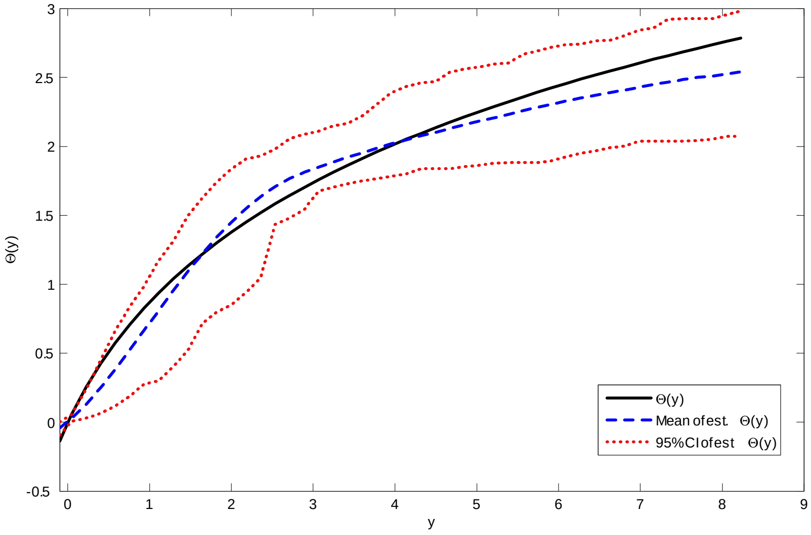

Figure 2: Median estimator of $\Theta(y)$ with endogenous regressors.

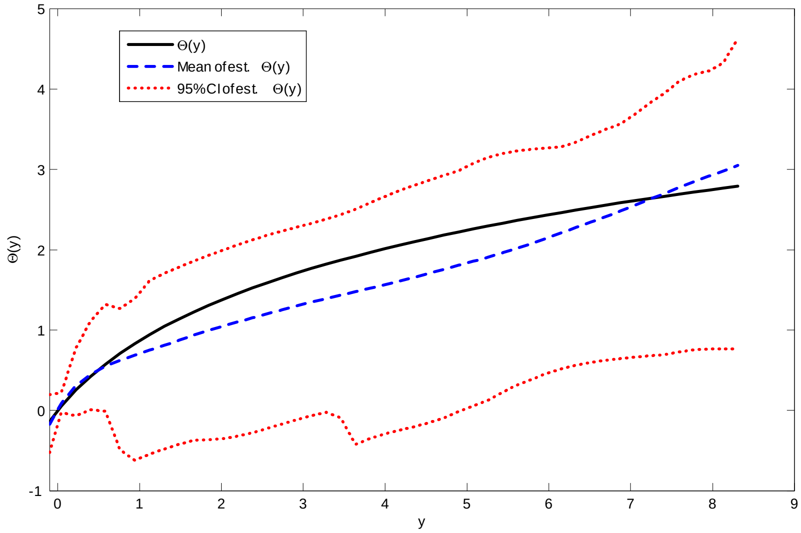

Figure 3: Mean-estimator of $\Theta(y)$ with endogenous regressors.

the "oversmoothed" and "undersmoothed" versions. Here, the integration is done over the interval ranging from the 2.5% quantile of Y to its 97.5% quantile. From the results in the table, we observe that the estimator is somewhat sensitive to the bandwidth choice. However, one should here keep in mind that the reported variation in bias, variance and MSE is taking place over a quite wide range of bandwidths. Moreover, while there is some variation in the performance of the estimator across different bandwidths, the overall integrated MSE remains quite small.

	Int. Squared Bias	Int. Variance	Int. MSE
Benchmark in Eq. (21)	0.0072	0.0010	0.0082
"Undersmoothing" by 20%	0.0170	0.0032	0.0201
"Oversmoothing" by 20%	0.0084	0.0023	0.0107

Table 1: Sensitivity of $\hat{\Theta}^{\mathrm{LAD}}(y)$ towards bandwidth choice

Finally, we investigate how the two-stage NPIV sieve estimator of $g(x)$ performs in the above design in the case of $n = 1000$ observations. As a benchmark we also computed the one-step oracle estimator of $g(x)$ that assumes knowledge of $\Theta(y)$. The results are reported in Table 2 for the same three bandwidth choices as examined in Table 1. We see that the feasible two-step estimators suffer from quite substantial biases compared to the oracle estimator with the bias increasing by a factor 4. We suspect that this is due to imprecise estimation of $\Theta(y)$ in the tails of the empirical support of Y, and conjecture that parts of these biases can be removed through trimming, something we have not explored here. On the other hand, while the variances of the two-step estimators also go up relative to the oracle estimator this increase is more moderate. Finally, we note that the bandwidth selection has some effect on the estimation of g as well, but less so compared to when $\Theta(y)$ is the target. Over all, the performance of the two-step estimators is satisfactory.

	Int. Squared Bias	Int. Variance	Int. MSE
Oracle one-step estimator	4.1823	8.2334	12.4156
2-step estimator using Eq. (21)	64.3620	14.1174	78.4794
2-step estimator w/ "undersmoothing" by 20%	74.9138	14.1620	89.0758
2-step estimator w/ "oversmoothing" by 20%	50.7798	13.7758	64.5556

Table 2: Performance of NPIV sieve estimator $\hat{g}(x)$

6. Discussion and Conclusion

We conclude by discussing possible extensions and applications of our results. First, note that additional instrumental variables are easily incorporated in our setup. Specifically, instead of assuming conditional independence between ϵ and X_I given X_{-I}, we could assume that some instrument W was available such that ϵ and X_I were conditionally independent given (X_{-I}, W), i.e. $\epsilon \perp X_I \mid (X_{-I}, W)$. This would amount to considering the conditional distribution $F_{Y|X,W}$ of Y given (X, W) which now satisfies:

$$F_{Y|X,W}(y|x, w) \equiv \Phi(y|x, w) = F_{\epsilon|X_{-I},W}\left(\Theta(y) - g(x), x_{-I}, w\right).$$

Redefining X to be (X, W), the above expression falls exactly in the framework obtained in (3), with an additional restriction on the function g which now no longer depends on the components of X corresponding to W. When the conditional distribution of the redefined vector X_{-I} given Z is complete, we know that g is identifiable. This identification result holds even without restricting the way that g depends on W; a fortiori, the identification result remains valid in this case.

Finally, we illustrate a way in which our results may be useful in economic applications. Say one is interested in counterfactual analysis of the situation in which the value of one of the regressors X is changed. For example, if in Equation (1), Y is the demand for some product, one may be interested in evaluating the effect of a change in the price of this product (one of the endogenous X's), while keeping all the other variables fixed. Then, the quantity of interest is the marginal effect:

$$\begin{aligned} E\left[\frac{\partial Y}{\partial X_j}\middle| X = x\right] &= E\left[T'\left(g(x) + \epsilon\right)\frac{\partial g(x)}{\partial x_j}\middle| X = x\right] \\ &= \int_{\mathcal{E}_x} T'\left(g(x) + \epsilon\right)\frac{\partial g(x)}{\partial x_j} f_{\epsilon|X}(\epsilon|x)d\epsilon, \end{aligned}$$

where we have let X_j denote the (endogenous) price, and X_{-j} denotes all the remaining regressors. Since all the terms on the right-hand side of the above equality are identified, so is the counterfactual on the left-hand side. Moreover, the marginal effect is consistently estimable using

$$\int_{\mathbb{R}} \hat{T}'\left(\hat{g}(x) + \epsilon\right)\frac{\partial \hat{g}(x)}{\partial x_j}\hat{f}_{\epsilon|X}(\epsilon|x)d\epsilon,$$

with $(\hat{T}, \hat{g}, \hat{F}_{\epsilon|X})$ as defined in the previous sections. As pointed out by Horowitz (1996), however, though the effects such as $E[\partial Y/\left(\partial X_j\right)| \, X = x]$ are consistently estimable, their rate of convergence is less than $\sqrt{n}$. This is because though T is estimable at the parametric rate, only nonparametric rates obtain for $(\hat{g}, \hat{F}_{\epsilon|X})$.

In certain situations, one may be able to work around this by looking at the conditional quantiles rather than expectations. For example, say that one is interested in predicting Y conditional on $X = x$. The most familiar predictor is a consistent estimator of $E\left[Y|X=x\right]$,

$$\int_{\mathbb{R}} \hat{T}\left(\hat{g}(x) + \epsilon\right) \hat{f}_{\epsilon|X}(\epsilon|x) d\epsilon.$$

As pointed out before, the above estimator is not $\sqrt{n}$ consistent. An alternative is to then use a conditional α-quantile $(0 < \alpha < 1)$ of the distribution of Y given $X = x$. In the context of the transformation model, the latter is given by

$$T(g(x) + q_\alpha(x)), \quad \text{where} \quad q_\alpha(x) = F^{-1}_{\epsilon|X}(\alpha|x),$$

is the conditional α-quantile of the conditional distribution of ϵ given $X = x$. Note that though the above quantity bypasses the need to consistently estimate (at $\sqrt{n}$ rate) the entire distribution $F_{\epsilon|X}$, one still needs to do so for $g(x)$. Thus, if $g(x)$ is only estimable at nonparametric rates, so will be the conditional quantiles of Y given $X = x$. This is unlike in Horowitz (1996), where it is assumed that g is parametric, $g(x) = \beta' x$, and that a $\sqrt{n}$-consistent estimator for β is already available.

References

Abbring, J. H., and G. J. van den Berg (2003): "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 71(5), 1491–1517.

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

Bijwaard, G., and G. Ridder (2005): "Correcting for selective compliance in a re-employment bonus experiment," *Journal of Econometrics*, 125, 77–111.

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

Chen, S. (2002): "Rank Estimation of Transformation Models," *Econometrica*, 70, 1683–1697.

Chen, X., V. Chernozhukov, S. Lee, and W. Newey (2011): "Local Identification of Nonparametric and Semiparametric Models," Discussion paper, Cowles Foundation Discussion Paper No. 1795.

Chen, X., and D. Pouzo (2012): “Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals,” *Econometrica*, 80, 277Ű321.

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

Darolles, S., Y. Fan, J. Florens, and E. Renault (2011): “Nonparametric Instrumental Regression,” *Econometrica*, 79, 1541–1565, Centre de Recherche et Développement Économique, 05-2002.

Ekeland, I., J. J. Heckman, and L. Nesheim (2004): “Identification and Estimation of Hedonic Models,” *The Journal of Political Economy*, 112, S60–S109.

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.

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

Härdle, W., and E. Mammen (1993): “Comparing Nonparametric versus Parametric Regression Fits,” *Annals of Statistics*, 21, 1926–1947.

Honore, B., and A. de Paula (2010): “Interdependent Durations,” *The Review of Economic Studies*, 77, 1138–1163.

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

——— (1998): “Bootstrap Methods for Median Regression Models,” *Econometrica*, 66, 1327–1351.

——— (1999): “Semiparametric Estimation of a Proportional Hazard Model with Unobserved Heterogeneity,” *Econometrica*, 67, 1001–1028.

Horowitz, J. L., and S. Lee (2007): “Nonparametric Instrumental Variables Estimation of a Quantile Regression Model,” *Econometrica*, 75, 1191–1208.

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.

Jochmans, K. (2011): "Pairwise-comparison Estimation with Nonparametric Controls," manuscript, Sciences Po Département d'économie.

Kristensen, D. (2011): "Semi-Nonparametric Estimation and Misspecification Testing of Diffusion Models," *Journal of Econometrics*, 164, 382–403.

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

Lewbel, A., X. Lu, and S. Liangjun (2013): "Specification Testing for Transformation Models with Applications to Generalized Accelerated Failure-Time Models," manuscript, Boston College.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics*, 40, 1132–1170.

Matzkin, R. L. (1991): "A Nonparametric Maximum Rank Correlation Estimator," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. Barnett, J. Powell, and G. Tauchen, chap. 11. Cambridge University Press.

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

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Palmer, C. (2014): "Why Did So Many Subprime Borrowers Default During the Crisis: Loose Credit or Plummeting Prices?," manuscript, MIT.

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

Van den Berg, G. J. (2001): "Duration Models: Specification, Identification and Multiple Durations," in *Handbook of Econometrics, Vol. 5*, ed. by J. J. Heckman, and E. Leamer, pp. 3381–3460. Elsevier.

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

Vanhems, A., and I. Van Keilegom (2013): "Semiparametric transformation model with endogeneity: a control function approach," manuscript, Université de Toulouse and Université catholique de Louvain.

Appendix A. Sieve IV Assumptions

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

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

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

Assumption A14. *(i) $g \in \mathcal{G} \equiv \Lambda_c^r(\mathcal{X})$ for some $r > 1/2$; (ii) $E[||X||^{2a}] < \infty$ for some $a > r$.*

Assumption A15. *The functions $h\left(x_I, z\right) \equiv E[\Theta\left(Y\right)|X_I = x_I, Z = z]$ and $M\left(x_I, z|g_n\right) \equiv E[g_n\left(X\right)|X_I = x_I, Z = z]$ belong to $\mathcal{H} \equiv \Lambda_c^{r_m}(\mathcal{X}_I \times \mathcal{Z})$, $r_m > 1/2$, for any $g_n \in \mathcal{G}_n$.*

Assumption A16. *(i) The smallest eigenvalue and the largest eigenvalue of $E[p^{J_n}(X_I, Z)p^{J_n}(X_I, Z)']$ 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 $\mathcal{X}_I \times \mathcal{Z}$, which is a compact set with non-empty interior.*

Assumption A17. *There is a $g_n \in \mathcal{G}_n$ such that $\tau_n^2 \times E[E[g(X) - g_n\left(X\right)|X_I, Z]^2] \leq const \times ||g - g_n||_X^2$.*

Assumption A18. *(i) $k_n \to \infty$, $J_n/n \to 0$; (ii) $nJ_n^{-2r_m/(1+d_z)-1} \to 0$ and $\lim_{n\to\infty}\left(J_n/k_n\right) = c_0 > 1$;*

Appendix B. Proofs

Proof of Theorem 1. Consider a structure $(\Theta, g, F_{\epsilon|X})$ that satisfies assumptions A1-A5, and generates $\Phi\left(y, x\right)$ in the sense of equation (3) in the main text. To establish the results of Theorem 1 we proceed in two steps. The first step establishes the identification of Θ under the normalization (N1). The second step shows that Θ is also identified under the normalization N2.

Step 1: Identification of Θ under (N1). Under assumptions A1, A4, and A5, the partial derivatives $\Phi_y(y,x)$ and $\Phi_i(y,x)$ $(i \in I)$ exist so that eqs. (4) by (5) hold. Under Assumption A6, one of the sets $\mathcal{A}_i$ $(1 \leq i \leq |I|)$ is nonempty. Pick an i for which this is true and take any point $\bar{x} \in \mathcal{A}_i$. Then for every $y \in \mathcal{Y}$, $\Theta'(y) = -s_i(y,\bar{x})\partial g(\bar{x})/\partial x_i$, where s_i is defined in eq. (6). Under Assumption A3 $\mathcal{Y}$ is a connected subset of $\mathbb{R}$ (i.e. an interval) that contains 0 so we can integrate on both sides from 0 to any $y \in \mathcal{Y}$ to get:

$$\Theta(y) = -\frac{\partial g(\bar{x})}{\partial x_i} S_i(y,\bar{x}) \quad \text{where} \quad S_i(y,\bar{x}) \equiv \int_0^y s_i(t,\bar{x})dt, \tag{22}$$

where we have used the normalization $\Theta(0) = 0$. Now to get rid of the partial of g, observe that $1 = \Theta'(0) = -s_i(0,\bar{x})\partial g(\bar{x})/\partial x_i$. Since $\bar{x} \in \mathcal{A}_i$, $\partial g(\bar{x})/\partial x_i \neq 0$ and is finite; hence, $s_i(0,\bar{x}) \neq 0$ and is finite as well, and we can write:

$$\frac{\partial g(\bar{x})}{\partial x_i} = -\frac{1}{s_i(0,\bar{x})}. \tag{23}$$

Combining (23) and (22) then yields

$$\Theta(y) = \frac{S_i(y,\bar{x})}{s_i(0,\bar{x})}, \tag{24}$$

so Θ is identified under (N1). It remains to be shown that the right-hand side of (24) does not depend on $\bar{x}$ nor i. For this, assume that there is an index j $(1 \leq j \leq |I|)$ also satisfying assumption A6 such that $\tilde{x} \in \mathcal{A}_j$, where $(j,\tilde{x}) \neq (i,\bar{x})$. Then notice that for all $y \in \mathcal{Y}$,

$$\frac{s_i(y,\bar{x})}{s_i(0,\bar{x})} = \frac{s_j(y,\tilde{x})}{s_j(0,\tilde{x})}. \tag{25}$$

Since from (24), we can write $\Theta(y) = \int_0^y \frac{s_i(t,\bar{x})}{s_i(0,\bar{x})}dt$, the result follows by combining the above expression with the equality established in (25). This completes the proof of part (i) of Theorem 1.

Step 2: Identification of Θ under (N2). Use the same reasoning up to equation (22). To get rid of the g term we now use a different approach. Evaluating (22) at $y = 1$ we get:

$$1 = \Theta(1) = -\frac{\partial g(\bar{x})}{\partial x_i} S_i(1, \bar{x}),$$

where we have used the fact that under normalization N2 $\Theta(1) = 1$. Since $\bar{x} \in \mathcal{A}_i$, $\partial g(\bar{x})/\partial x_i \neq 0$ and is finite; hence, $S_i(1, \bar{x}) \neq 0$ and is finite as well, so we can write:

$$\frac{\partial g(\bar{x})}{\partial x_i} = -\frac{1}{S_i(1, \bar{x})}. \tag{26}$$

Combining (22) and (26) then gives for every $y \in \mathcal{Y}$:

$$\Theta(y) = \frac{S_i(y, \bar{x})}{S_i(1, \bar{x})}, \tag{27}$$

so Θ is identified under (N2). To show that the right-hand side of (27) does not depend on i nor $\bar{x}$ use the same reasoning as in Step 1 to establish that for all $y \in \mathcal{Y}$,

$$\frac{s_i(y, \bar{x})}{S_i(1, \bar{x})} = \frac{s_j(y, \tilde{x})}{S_j(1, \tilde{x})},$$

where j and $\tilde{x}$ are as in Step 1. Combining the above with the expression for Θ in (27) then yields the result. This completes the proof of Theorem 1. □

Proof of Corollary 1. Under the assumptions of Theorem 1, Θ is identified. We now proceed to establish the identification of g and $F_{\epsilon|X}$.

First, we consider the identification of g with respect to the exogenous regressors X_I. We start with $i = 1$. Take any $x \in \mathcal{X}$: then either $\Phi_1(y, x) = 0$ for all $y \in \mathcal{Y}$, or $\Phi_1(y, x) \neq 0$ for some $y \in \mathcal{Y}$. The first is true if and only if $\partial g(x)/\partial x_1 = 0$. If the latter is true, take y_x such that $\Phi_1(y_x, x) \neq 0$. Note that this y_x can be chosen so that $\Theta'(y_x) \neq 0$, i.e. $\Phi_y(y_x, x) \neq 0$ (this follows by the absolute continuity of $F_{\epsilon|X}$ in A1 and the fact that Θ' can be zero only at isolated points). Taking ratios in (4)-(5) with $i = 1$, it then follows that

$$\partial g(x)/\partial x_1 = -\frac{\Theta'(y_x)}{s_1(y_x, x)} \quad \text{where} \quad s_1(y_x, x) = \frac{\Phi_y(y_x, x)}{\Phi_1(y_x, x)},$$

and with Θ as identified in Theorem 1. Now let $\Gamma_1 : \mathcal{X} \to \mathbb{R}$ be defined as:

$$\Gamma_1(x) \equiv \begin{cases} 0, & \text{if } \Phi_1(y, x) = 0 \text{ for all } y \in \mathcal{Y}, \\ -\frac{\Theta'(y_x)}{s_1(y_x, x)}, & \text{otherwise.} \end{cases}$$

Note that the function Γ_1 is known, i.e. observable, and we have that $\partial g(x)/\partial x_1 = \Gamma_1(x)$ for every $x \in \mathcal{X}$. A particular solution $\bar{g}_1 : \mathcal{X} \to \mathbb{R}$ to this partial differential equation is

$$\bar{g}_1 \left(x_1, x_2, \ldots, x_{d_x}\right) = \int_c^{x_1} \Gamma_1(u, x_2, \ldots, x_{d_x}) du \tag{28}$$

for some $c_1 \in \mathcal{X}_1$. Obviously, any solution to $\partial g(x)/\partial x_1 = \Gamma_1(x)$ must have the same partial derivative with respect to x_1 as $\bar{g}_1$ in (28) and so

$$g(x) = \bar{g}_1(x) + \beta_1(x_2, \ldots, x_{d_x})$$

for some unknown function $\beta_1 : \mathcal{X}_{-1} \to \mathbb{R}$. If $|I| = 1$ we can stop here. If on the other hand $|I| \geq 2$, we can repeat the same reasoning as above with any value $x \in \mathcal{X}$ such that $\partial g(x)/\partial x_2 \neq 0$. This will give us a known function Γ_2 such that $\partial g(x)/\partial x_2 = \Gamma_2(x)$ for every $x \in \mathcal{X}$. Differentiating (28) with respect to x_2 then gives us

$$\frac{\partial \beta_1(x_2, \ldots, x_{d_x})}{\partial x_2} = \Gamma_2(x) - \frac{\partial \bar{g}_1(x)}{\partial x_2},$$

i.e.

$$\beta_1(x_2, \ldots, x_{d_x}) = \bar{g}_2(x_2, \ldots, x_{d_x}) + \beta_2(x_3, \ldots, x_{d_x}), \tag{29}$$

where $\bar{g}_2$ is a known function

$$\bar{g}_2(x_2, \ldots, x_{d_x}) \equiv \int_{c_2}^{x_2} \left[\Gamma_2(x_1, u, x_3, \ldots, x_{d_x}) - \frac{\partial \bar{g}_1(x_1, u, x_3, \ldots, x_{d_x})}{\partial x_2}\right] du,$$

with some $c_2 \in \mathcal{X}_2$. Combining (28) and (29) then gives, for all $x \in \mathcal{X}$:

$$g(x) = \bar{g}_1(x) + \bar{g}_2(x) + \beta_2(x_3, \ldots, x_{d_x}),$$

where both functions $\bar{g}_1$ and $\bar{g}_2$ are known. If $|I| = 2$ we stop here; otherwise, repeating the same reasoning until we have exhausted the exogenous regressors will lead to

$$g(x) = \bar{g}(x) + \beta(x_{-I}), \quad \text{for all } x \in \mathcal{X} \text{ where } \bar{g} \text{ is known.} \tag{30}$$

Thus, g is identified up to an additive unknown function of x_{-I}. Now let g be an arbitrary solution, and consider $E(\epsilon|Z)$ where $\epsilon = \Theta(Y) - g(X)$ with Θ as identified in Theorem 1 and g as in (30). Letting $F_{Y|Z}$ and $F_{X|Z}$ denote the conditional distributions of Y and X given Z, respectively, we have:

$$\begin{aligned} E\left[\epsilon|Z=z\right] &= \int_{\mathcal{Y}} \Theta(y)dF_{Y|Z}(y,z) - \int_{\mathcal{X}} g(x)dF_{X|Z}(x,z) \\ &= \int_{\mathcal{Y}} \Theta(y)dF_{Y|Z}(y,z) - \int_{\mathcal{X}} [\bar{g}(x)+\beta(x_{-1})]dF_{X|Z}(x,z) \end{aligned} \tag{31}$$

Now, consider a structure $(\Theta, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ that is observationally equivalent to $(\Theta, g, F_{\epsilon|X})$ and has the same properties as $(\Theta, g, F_{\epsilon|X})$. It follows from (31) that for a.e. $z \in \mathcal{Z}$:

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

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 (30), this implies that $g(x) = \tilde{g}(x)$ for a.e. $x \in \mathcal{X}$. Thus g is identified.

Since Θ and g are identified, $\epsilon = \Theta(Y) - g(X)$ is identified and so is its conditional distribution $F_{\epsilon|X}$.

To complete the proof we need to establish that Assumption A7(ii) is also necessary to identify g and $F_{\epsilon|X}$. To see this, assume that A7(ii) does not hold, i.e. there exists some nonzero function $h(x_{-I})$ such that $E[h(X_{-I})|Z] = 0$ a.s. It then suffices to consider $\tilde{g}(x) \equiv g(x) + h(x_{-I})$ and $\tilde{\epsilon} \equiv \epsilon - h(X_{-I})$ to show that the two structures $(\Theta, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$ and $(\Theta, g, F_{\epsilon|X})$ are different yet observationally equivalent. Thus, $(g, F_{\epsilon|X})$ is not identified. □

Proof of Theorem 2. We first linearize $\hat{\theta}(y,x)$ with respect to $\hat{S}(y,x)$ and $\hat{S}(1,x)$,

$$\begin{aligned} \hat{\theta}(y,x) - \Theta(y) &= \frac{1}{S(1,x)}\{\hat{S}(y,x) - S(y,x)\} - \frac{S(y,x)}{S^2(1,x)}\{\hat{S}(1,x) - S(1,x)\} \\ &\quad + O(\|\hat{S} - S\|_\infty^2), \end{aligned} \tag{32}$$

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

$$\begin{aligned}
&\int_{\mathcal{X}_w} \frac{w\left(x\right)}{S\left(1,x\right)}\{\hat{S}\left(y,x\right)-S\left(y,x\right)\}dx \\
&= \int_{\mathcal{X}_w} \frac{w\left(x\right)}{S\left(1,x\right)}\{\bigtriangledown_p S\left(y,x\right)\left[\hat{p}-p\right]+\bigtriangledown_f S\left(y,x\right)[\hat{f}-f]\}dx + o_P\left(1/\sqrt{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\delta_i^{\bar{w}_1}\left(1,y\right)+o_P\left(1/\sqrt{n}\right),
\end{aligned}$$

where $\bar{w}_1\left(y_0,x\right)=w\left(x\right)/S\left(y_0,x\right)$, and we have let:

$$\begin{aligned}
\delta_i^{\bar{w}}\left(y_0,y\right) \equiv\ & \bar{w}\left(y_0,X_i\right)\left\{\int_{\max\{0,Y_i\}}^{y} D_{p,0}\left(u,X_i\right)du+\int_0^y D_{f,0}\left(u,X_i\right)du\right\} \\
&+\int_0^y \frac{\partial\left[\bar{w}\left(y_0,X_i\right)D_{f,1}\left(u,X_i\right)\right]}{\partial x_1}du \\
&+\mathbb{I}\left\{0\leq Y_i\leq y\right\}\left\{\bar{w}\left(y_0,X_i\right)D_{p,y}\left(Y_i,X_i\right)-\frac{\partial\left[\bar{w}\left(y_0,X_i\right)D_{p,1}\left(Y_i,X_i\right)\right]}{\partial x_1}\right\}, \qquad (33)
\end{aligned}$$

with $D_{p,k}\left(y,x\right)$ and $D_{f,k}\left(y,x\right)$, $k\in\{0,1,y\}$, being as defined in Equation (42). Moreover,

$$\int_{\mathcal{X}_w}\frac{w\left(x\right)S\left(y,x\right)}{S^2\left(1,x\right)}\{\hat{S}\left(1,x\right)-S\left(1,x\right)\}dx=\frac{1}{n}\sum_{i=1}^{n}\delta_i^{\bar{w}_2}\left(y,1\right)+o_P\left(1/\sqrt{n}\right),$$

where $\bar{w}_2\left(y_0,x\right) = w\left(x\right)S\left(y_0,x\right)/S^2\left(1,x\right)$. Finally, by Lemmas 1 and 3, $||\hat{S}-S||_{\infty}^2 = o_P\left(1/\sqrt{n}\right)$. Collecting the above results,

$$\sqrt{n}\{\hat{\Theta}^{\mathrm{LS}}\left(y\right)-\Theta\left(y\right)\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i^w\left(y\right)+o_P\left(1\right),$$

uniformly over $\mathcal{Y}_0$, where $\delta_i^w\left(y\right)$ is the random function defined as

$$\delta_i^w\left(y\right)\equiv\delta_i^{\bar{w}_1}\left(1,y\right)-\delta_i^{\bar{w}_2}\left(y,1\right). \qquad (34)$$

It is easily checked that $E[\delta_i^w\left(y\right)]=0$, while we show below that $E[\delta_i^w\left(y\right)^2]<\infty$. Thus, pointwise weak convergence follows by the CLT. This extends to weak functional convergence over the compact set $\mathcal{Y}_0$ if we can show stochastic equicontinuity. However, this follows from, for example, van der Vaart and Wellner (1996) since $\left(y_0,y\right)\mapsto\delta_i^w\left(y_0,y\right)$ is continuous almost surely and has an L_2-envelope, $|\delta_i^w\left(y_0,y\right)|\leq\bar{\delta}_i^w$, $y\in\mathcal{Y}_0$, with $E[(\bar{\delta}_i^w)^2]<\infty$: It is easily checked that for an appropriate constant c (depending on the size of the support of w, it holds for both $\bar{w}=\bar{w}_1$ and $\bar{w}=\bar{w}_2$ as defined above,

$$\begin{aligned}\delta_i^{\bar{w}}(y_0,y)^2 &\leq c\sup_{(y,y_0,x)\in\mathcal{Y}_0^2\times\mathcal{X}_w}\bar{w}^2(y_0,x)\left\{D_{p,0}^2(y,x)+D_{p,y}^2(y,x)+D_{f,0}^2(y,x)\right\}\\&\quad+c\sup_{(y,y_0,x)\in\mathcal{Y}_0^2\times\mathcal{X}_w}\left\{\left|\frac{\partial\left[\bar{w}(y_0,x)D_{p,1}(y,x)\right]}{\partial x_1}\right|^2+\left|\frac{\partial\left[\bar{w}(y_0,x)D_{f,1}(u,x)\right]}{\partial x_1}\right|\right\}.\end{aligned}$$

Since all the functions on the right-hand side are continuous and $\mathcal{Y}_0^2\times\mathcal{X}_w$ is compact, the bound is finite.

Next, consider the LAD version. First, it is easily checked that, for any *fixed* b, $\Theta(y)=\arg\min_\theta Q_b(\theta|\Theta(y))$ is the unique minimum. Since $||\hat{\theta}(y,x)-\Theta(y)||_\infty=o_P(1)$, it follows by standard arguments that $||\hat{\Theta}_b^{\mathrm{LAD}}-\Theta||_\infty=o_P(1)$. Next, by the mean-value theorem,

$$0=\frac{\partial Q_b(\Theta(y)|\hat{\theta}(y,\cdot))}{\partial\theta}+\frac{\partial^2 Q_b(\bar{\Theta}(y)|\hat{\theta}(y,\cdot))}{\partial\theta^2}\{\hat{\Theta}_b(y)-\Theta(y)\},$$

for some $\bar{\Theta}(y)\in[\Theta(y),\hat{\Theta}_b(y)]$ where, by a functional Taylor expansion w.r.t. $\hat{\theta}(y,\cdot)$,

$$\frac{\partial Q_b(\Theta(y)|\hat{\theta}(y,\cdot))}{\partial\theta}=\Gamma[\hat{\theta}(y,\cdot)-\Theta(y)]+O(||\hat{\theta}(y,\cdot)-\Theta(y)||_\infty^2)$$

where we have used that $\partial Q_b(\Theta(y)|\Theta(y))/(\partial\theta)=0$ and

$$\Gamma[d\Theta]:=-4f_b(0)\int w(x)\,d\Theta(y,x)\,dx,$$

where $f_b(\theta)=f(\theta/b)/b$ and $f(\theta)=F'(\theta)$. Moreover,

$$\frac{\partial^2 Q_b(\Theta(y)|\theta(y,\cdot))}{\partial\theta^2}=4\int w(x)f_b(\theta(y,x)-\theta)\,dx+2\int w(x)\{\theta(y,x)-\theta\}f_b'(\theta(y,x)-\theta)\,dx,$$

and, again using the uniform convergence result for $\hat{\theta}(y,x)$,

$$\begin{aligned}\frac{\partial^2 Q_b(\bar{\Theta}(y)|\hat{\theta}(y,\cdot))}{\partial\theta^2}&=\frac{\partial^2 Q_b(\Theta(y)|\hat{\theta}(y,\cdot))}{\partial\theta^2}+o_P(1)\\&=4f_b(0)+o_P(1),\end{aligned}$$

uniformly over y. Collecting the above results, $\hat{\Theta}_b^{\mathrm{LAD}}(y)=\hat{\Theta}^{\mathrm{LS}}(y)+o_P\left(n^{-1/2}\right)$, and it now follows from Theorem 2 that $\sqrt{n}(\hat{\Theta}_b^{\mathrm{LAD}}(y)-\Theta(y))\Rightarrow\mathbb{W}(y)$ for any fixed bandwidth $b>0$. □

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 $\|g - g_n\|_X \leq$ 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$\|h - h_n\|_{X_1,Z} \leq$ const. $\times$ $J_n^{-r_m/(1+d_z)}$.

Claim 2: $||\tilde{h} - h||_{X_1,Z} = O_p\left(J_n^{-r_m/(1+d_z)} + \sqrt{J_n/n}\right)$ and $\sup_{g_n \in \mathcal{G}_n} ||\hat{M}(\cdot|g_n) - M(\cdot|g_n)||_{X_1,Z} = O_p\left(J_n^{-r_m/(1+d_z)} + \sqrt{J_n/n}\right)$.

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 Θ 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 A17 and A18 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, $\hat{g}$. □

Proof of Theorem 4. With no loss of generality consider $i = 1$. Without X_1 being exogenous, we have:

$$\begin{aligned}\Phi_y(y,x) &= \Theta'(y) f_{\epsilon|X}(\Theta(y) - g(x), x)\\ \Phi_1(y,x) &= -g_1(x) f_{\epsilon|X}(\Theta(y) - g(x), x) + \left.\frac{\partial F_{\epsilon|X}(t,x)}{\partial x_1}\right|_{t=\Theta(y)-g(x)}\end{aligned}$$

Then taking ratios for every $(y,x) \in \mathcal{Y} \times \mathcal{A}_1$, we have

$$s_1(y,x) \equiv \frac{\Phi_y(y,x)}{\Phi_1(y,x)} = \frac{\delta_1(y,x) - \Theta'(y)}{g_1(x)}, \tag{35}$$

where we have let $g_1(x) \equiv \partial g(x)/\partial x_1$, and

$$\pi_1(y,x) \equiv \Theta'(y) \frac{\partial F_{\epsilon|X}(t,x)/\partial x_1\big|_{t=\Theta(y)-g(x)}}{g_1(x) f_{\epsilon|X}(\Theta(y) - g(x), x) + \partial F_{\epsilon|X}(t,x)/\partial x_1\big|_{t=\Theta(y)-g(x)}}. \tag{36}$$

Proceeding as in the proof of Theorem 1, integrating (35) between 0 and any $y \in \mathcal{Y}$, and using $\Theta(0) = 0$, then gives

$$\Theta(y) = -g_1(x) S_1(y,x) + \Pi_1(y,x), \tag{37}$$

with $S_1(y,x) = \int_0^y s_1(u,x)du$ as before, and

$$\Pi_1(y,x) \equiv \int_0^y \pi_1(u,x)du.$$

We now proceed in two steps, one for each normalization.

Step 1: Under normalization N1 Plugging $\Theta'(0) = 1$ back into (35) yields

$$g_1(x) = \frac{\pi_1(0,x) - 1}{s_1(0,x)} = \frac{1}{s_1(0,x)},$$

where the second equality follows from the expression of $\pi_1(y,x)$ in (36). Combining the above with (37) then gives

$$\Theta(y) = \frac{S_1(y,x)}{s_1(0,x)} + \Pi_1(y,x) = \vartheta_1(y,x) + \Pi_1(y,x).$$

It follows directly from Theorem 1 that H_0 implies $\Pi_1(y,x) = 0$ for all $y \in \mathcal{Y}$. It now remains to show the converse, i.e. that $\Pi_1(y,x) = 0$ for all $y \in \mathcal{Y}$ implies H_0 (this in turn is equivalent to: H_a implies $\Pi_1(y,x) \neq 0$ for some $y \in \mathcal{Y}$). It follows directly from the expression of $\Pi_1(y,x)$ that $\Pi_1(y,x) = 0$ for all $y \in \mathcal{Y}$ only if the integrand is everywhere zero, i.e. only if $\pi_1(y,x) = 0$ for all $y \in \mathcal{Y}$. Since the set of points $y \in \mathcal{Y}$ where $\Theta'(y) = 0$ is isolated, it follows from (36) that $\pi_1(y,x) = 0$ for all $y \in \mathcal{Y}$ only if

$$\left.\frac{\partial F_{\epsilon|X}(t,x)}{\partial x_1}\right|_{t=\Theta(y)-g(x)} = 0, \qquad \text{for all } y \in \mathcal{Y},$$

i.e. $\partial F_{\epsilon|X}(t,x)/\left(\partial x_1\right) = 0$ for all $t \in \mathcal{E}_x$. The latter in turn is equivalent to $\epsilon \perp X_1 \mid X_{-1}$.

Step 2: Under Normalization N2. Plugging $\Theta(1) = 1$ into (37) gives

$$g_1(x) = \frac{\Pi_1(1,x) - 1}{S_1(1,x)},$$

which together with (37) again gives

$$\Theta(y) = \frac{S_1(y,x)}{S_1(1,x)} + \Pi_1(y,x) - \Pi_1(1,x)\frac{S_1(y,x)}{S_1(1,x)} = \theta_1(y,x)\left(1 - \Pi_1(1,x)\right) + \Pi_1(y,x). \tag{38}$$

Similar to before, H_0 implies $\Pi_1(y,x) = 0$ for all $y \in \mathcal{Y}$, and so $\Theta(y) = \theta_1(y,x)$. It remains to show the converse: Suppose $\theta_1(y,x) = \Theta(y)$ for all y. Then (38) can only hold if $\Pi_1(1,x)S_1(y,x) = \Pi_1(y,x)S_1(1,x)$, which by definition of $\Pi_1(y,x)$ and (35) implies

$$\int_0^y \pi_1(u,x)du \int_0^1 \frac{\pi_1(u,x) - \Theta'(u)}{g_1(x)}du = \int_0^1 \pi_1(u,x)du \int_0^y \frac{\pi_1(u,x) - \Theta'(u)}{g_1(x)}du,$$

so since $\Theta(1) = 1$, necessarily

$$\Theta(y)\int_0^1 \pi_1(u,x)du = 0, \quad \text{for all } y \in \mathcal{Y}.$$

This is only possible if $\Pi_1(1,x) = \int_0^1 \pi_1(u,x)du = 0$. Plugging back into (38), we then get that $\Theta(y) = \theta_1(y,x)$ only if $\Pi_1(y,x) = 0$ for all $y \in \mathcal{Y}$, which following the same reasoning as at the end of Step 1 implies $\epsilon \perp X_1 \mid X_{-1}$. □

Proof of Theorem 5. First note that since $\sup_{y\in\mathcal{Y}} |\hat{\Theta}\left(y\right) - \Theta\left(y\right)| = O_P\left(1/\sqrt{n}\right)$ we can treat $\Theta\left(y\right)$ as known in the analysis of Q. Next, combining Equation (32) with Lemma 1

$$\begin{aligned}\hat{\theta}_0\left(y,x\right) - \Theta\left(y\right) \quad &= \quad \frac{1}{S\left(1,x\right)}\left\{\bigtriangledown_p S\left(y,x\right)\left[\hat{p} - p\right] + \bigtriangledown_f S\left(y,x\right)\left[\hat{f} - f\right]\right\} \\ &\quad -\frac{S\left(y,x\right)}{S^2\left(1,x\right)}\left\{\bigtriangledown_p S\left(1,x\right)\left[\hat{p} - p\right] + \bigtriangledown_f S\left(1,x\right)\left[\hat{f} - f\right]\right\} + R,\end{aligned}$$

where R satisfies Equation (43). In particular, $nh_{0,x}^{d_x/2+2}R^2 = o_P(1)$ under Assumption A13. From the proof of Lemma 1,

$$\hat{\theta}_0(y,x) - \Theta(y) \simeq \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_{0,x}}(X_i - x)\,\bar{D}_{1,i}(y,x) + \mathbf{K}_{h_{0,x},1}(X_i - x)\,\bar{D}_{2,i}(y,x) + O_P\left(h_y^m\right),$$

where

$$\bar{D}_k(y, Y_i, x) = \frac{1}{S(1,x)}\tilde{D}_k(y, Y_i, x) - \frac{S(y,x)}{S^2(1,x)}\tilde{D}_k(1, Y_i, x), \quad k = 1, 2, \tag{39}$$

$$\begin{aligned}
\tilde{D}_1(y, Y_i, x) &= \int_{\max\{0,Y_i\}}^{y} D_{p,0}(u,x)\,du + \mathbb{I}\{0 \le Y_i \le y\}\,D_{p,y}(Y_i, x) + \int_0^y D_{f,0}(u,x)\,du, \\
\tilde{D}_2(y, Y_i, x) &= \mathbb{I}\{0 \le Y_i \le y\}\,D_{p,y}(Y_i, x) + \int_0^y D_{f,1}(u,x)\,du,
\end{aligned}$$

and $\mathbf{K}_{h_{0,x},1}(X_i - x) = \partial \mathbf{K}_{h_{0,x}}(X_i - x)\,/\,(\partial x_1)$. Substituting the resulting linearized version into Q,

$$\begin{aligned}
Q &\simeq \int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{K}_{h_{0,x}}(X_i - x)\,\bar{D}_1(y, Y_i, x) + \mathbf{K}_{1,h_{0,x}}(X_i - x)\,\bar{D}_2(y, Y_i, x)\right\}\right]^2 dydx \\
&\simeq \int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{h_{0,x}}(X_i - x)\,\bar{D}_1(y, Y_i, x) - f(x)\,E\left[\bar{D}_1(y, Y_i, x)\right]\right]^2 dydx \\
&\quad + \int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{1,h_{0,x}}(X_i - x)\,\bar{D}_2(y, Y_i, x) - f_1(x)\,E\left[\bar{D}_2(y, Y_i, x)\right]\right]^2 dydx \\
&\equiv Q_1 + Q_2.
\end{aligned}$$

For Q_1, we proceed as in, for example, the proof of Proposition 1 in Härdle and Mammen (1993) to obtain that, with $e_{k,i}(y,x) = \bar{D}_k(y, Y_i, x) - E\left[\bar{D}_k(y, Y_i, x)\right]$, $k = 0, 1$,

$$\begin{aligned}
Q_1 &\simeq \int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{K}_{h_{0,x}}(X_i - x)\,e_{1,i}(y,x)\right]^2 dydx \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}\int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\,\mathbf{K}_{h_{0,x}}(X_i - x)\,\mathbf{K}_{h_{0,x}}(X_j - x)\,e_{1,i}(y,x)\,e_{1,j}(y,x)\,dydx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\,\mathbf{K}_{h_{0,x}}^2(X_i - x)\,e_{1,i}^2(y,x)\,dydx \\
&\quad + \frac{1}{n^2}\sum_{i \ne j}\int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\,\mathbf{K}_{h_{0,x}}(X_i - x)\,\mathbf{K}_{h_{0,x}}(X_j - x)\,e_{1,i}(y,x)\,e_{1,j}(y,x)\,dydx \\
&\equiv Q_{11} + Q_{12},
\end{aligned}$$

where, as $nh_{0,x}^{d_x} \to \infty$ and $nh_{0,x}^{d_x/2+m} \to 0$, $Q_{11} = m_1 + o_P(1)$ and $nh_{0,x}^{d_x/2} Q_{12} \to^d N(0, v_1)$; here,

$$\begin{aligned} m_1 &= \frac{1}{nh_{0,x}^{d_x}} \int \mathbf{K}^2(x)\,dx \times \int\int \sigma_1^2(y,x)\,\bar{W}(y,x)\,f(x)\,dydx, \\ v_1 &= 2\int [\mathbf{K} * \mathbf{K}]^2(x)\,dx \times \int\int \sigma_1^4(y,x)\,\bar{W}^2(y,x)\,f^2(x)\,dydx. \end{aligned}$$

In particular, $nh_{0,x}^{d_x/2+2} Q_{12} = o_P(1)$. Similar arguments can be applied to Q_2, see, e.g., proof of Theorem 7 in Kristensen (2011), to obtain that, as $nh_{0,x}^{d_x+2} \to \infty$ and $nh_{0,x}^{d_x/2+2+m} \to 0$,

$$\begin{aligned} Q_2 &\simeq \int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x) \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_{0,x},1}(X_i - x)\,e_{1,i}(y,x)\right]^2 dydx \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\,\mathbf{K}_{h_{0,x},1}^2(X_i - x)\,e_i^2(y,x)\,dydx \\ &\quad + \frac{1}{n^2}\sum_{i\neq j} \int_{\mathcal{Y}}\int_{\mathcal{X}} \bar{W}(y,x)\,\mathbf{K}_{h_{0,x},1}(X_i - x)\,\mathbf{K}_{h_{0,x},1}(X_j - x)\,e_i(y,x)\,e_j(y,x)\,dydx \\ &\equiv Q_{21} + Q_{22}, \end{aligned}$$

where $Q_{21} = m_2 + o_P(1)$ and $nh_{0,x}^{2+d_x/2} Q_{22} \to^d N(0, v_Q)$ with

$$\begin{aligned} m_2 &= \frac{1}{nh_{0,x}^{d_x+2}} \int \mathbf{K}_1^2(x)\,dx \times \int\int \sigma_2^2(y,z)\,W(y,x)\,f(x)\,dydx, \\ v_Q &= 2\int [\mathbf{K}_1 * \mathbf{K}_1]^2(x)\,dx \times \int\int \sigma_2^4(y,x)\,W^2(y,x)\,f^2(x)\,dydx. \end{aligned}$$

The claimed result now follows. □

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 $df(x)$:

$$\begin{aligned} \nabla_p S(y,x)[dp] \equiv& \int_0^y D_{p,0}(u,x)\,dp(u,x)\,du + \int_0^y D_{p,y}(u,x)\,dp_y(u,x)\,du \\ &+ \int_0^y D_{p,1}(u,x)\,dp_1(u,x)\,du, \end{aligned} \tag{40}$$

$$\nabla_f S(y,x)[df] \equiv \int_0^y D_{f,0}(u,x)\,du \times df(x) + \int_0^y D_{f,1}(u,x)\,du \times df_1(x), \tag{41}$$

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

$$\begin{aligned} & D_{p,0}(y,x) \equiv \frac{\Phi_y(y|x) f_1(x)}{\Phi_1^2(y|x) f^2(x)}, \quad D_{p,y}(y,x) \equiv \frac{1}{f(x)\Phi_1(y,x)}, \\ & D_{f,0}(y,x) \equiv \frac{\Phi_y(y|x)}{\Phi_1(y,x) f(x)} \left[1 - \frac{2\Phi(y,x) f_1(x)}{f(x)\Phi_1(y,x)} + f(x) + \frac{\Phi(y,x) f_1(x)}{\Phi_1(y,x)}\right], \\ & D_{f,1}(y,x) \equiv \frac{\Phi_y(y,x)\Phi(y,x)}{\Phi_1^2(y,x) f(x)}, \quad D_{p,1}(y,x) \equiv -\frac{\Phi_y(y,x)}{f(x)\Phi_1^2(y,x)}. \end{aligned} \tag{42}$$

The first lemma then shows that these two functionals are the pathwise differentials of $S(y,x)$ with respect to 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 (40)-(41), the following expansion holds uniformly over $(y,x) \in \mathcal{Y}_0 \times \mathcal{X}_w$:*

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

Proof of Lemma 1. Suppressing dependence on y and x, let $\hat{\Phi} = \hat{p}/\hat{f}$ denote the kernel estimator. We in the following use repeatedly the following identity:

$$\frac{\hat{a}}{\hat{b}} - \frac{a}{b} = \frac{1}{b}\{\hat{a} - a\} - \frac{a}{b^2}\left\{\hat{b} - b\right\} + \frac{\left\{\hat{b} - b\right\}}{b\hat{b}}\left\{\hat{a} - a - \frac{a(\hat{b} - b)}{b}\right\}.$$

First,

$$\frac{\hat{\Phi}_y}{\hat{\Phi}_1} - \frac{\Phi_y}{\Phi_1} = \frac{1}{\Phi_1}\{\hat{\Phi}_y - \Phi_y\} - \frac{\Phi_y}{\Phi_1^2}\{\hat{\Phi}_1 - \Phi_1\} + \frac{\{\hat{\Phi}_1 - \Phi_1\}}{\hat{\Phi}_1\Phi_1}\left[\{\hat{\Phi}_y - \Phi_y\} - \frac{\Phi_y\{\hat{\Phi}_1 - \Phi_1\}}{\Phi_1}\right],$$

where $\Phi_y = p_y/f$ and $\Phi_1 = p_1/f - pf_1/f^2$. Thus,

$$\hat{\Phi}_y - \Phi_y = \frac{1}{f}\{\hat{p}_y - p_y\} + \frac{p_y}{f^2}\{\hat{f} - f\} + \frac{\{\hat{f} - f\}}{\hat{f}f}\left[\{\hat{p}_y - p_y\} - \frac{p_y\{\hat{f} - f\}}{f}\right],$$

and

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

Combining the last three expressions and then rearranging,

$$\begin{aligned}
\frac{\hat{\Phi}_y}{\hat{\Phi}_1} - \frac{\Phi_y}{\Phi_1} &= \frac{1}{\Phi_1}\left\{\frac{1}{f}\left\{\hat{p}_y - p_y\right\} + \frac{p_y}{f^2}\{\hat{f} - f\} + \frac{\{\hat{f} - f\}}{\hat{f}f}\left[\{\hat{p}_y - p_y\} - \frac{p_y\{\hat{f} - f\}}{f}\right]\right\} \\
&\quad -\frac{\Phi_y}{\Phi_1^2}\left\{-\frac{f_1}{f^2}\left\{\hat{p} - p\right\} + \frac{1}{f}\left\{\hat{p}_1 - p_1\right\} + \left[\frac{2pf_1}{f^3} - \frac{p_1}{f^2}\right]\{\hat{f} - f\} - \frac{p}{f^2}\{\hat{f}_1 - f_1\}\right\} \\
&= \frac{\Phi_y}{\Phi_1^2}\frac{f_1}{f^2}\left\{\hat{p} - p\right\} + \frac{1}{\Phi_1 f}\left\{\hat{p}_y - p_y\right\} - \frac{\Phi_y}{\Phi_1^2}\frac{1}{f}\left\{\hat{p}_1 - p_1\right\} \\
&\quad + \left[\frac{p_y}{\Phi_1 f^2} - \frac{\Phi_y}{\Phi_1^2}\left(\frac{2pf_1}{f^3} - \frac{p_1}{f^2}\right)\right]\{\hat{f} - f\} + \frac{\Phi_y}{\Phi_1^2}\frac{p}{f^2}\{\hat{f}_1 - f_1\} + R \\
&= D_{p,0}\left\{\hat{p} - p\right\} + D_{p,y}\left\{\hat{p}_y - p_y\right\} + D_{p,1}\left\{\hat{p}_1 - p_1\right\} \\
&\quad + D_{f,0}\{\hat{f} - f\} + D_{f,1}\{\hat{f}_1 - f_1\} + R,
\end{aligned}$$

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

where R is the remainder term satisfying

$$R = O\left(|\hat{p} - p|^2\right) + O\left(|\hat{p}_1 - p_1|^2\right) + O\left(|\hat{p}_y - p_y|^2\right) + O\left(|\hat{f} - f|^2\right) + O\left(|\hat{f}_1 - f_1|^2\right), \tag{43}$$

and $D_{p,0}$, $D_{p,y}$, $D_{p,1}$, $D_{f,0}$ and $D_{f,1}$ are defined in Equation (42). Given the definitions of $\triangledown_p S(y,x)[dp]$ and $\triangledown_f S(y,x)[df]$, this shows that

$$\hat{S}(y,x) - S(y,x) = \triangledown_p S(y,x)[\hat{p} - p] + \triangledown_f S(y,x)[\hat{f} - f] + R(y,x).$$

What remains to be shown is that the remainder term satisfies $\sup_{(y,x)\in\mathcal{Y}_0\times\mathcal{X}_w} R(y,x) = o_P(1/\sqrt{n})$. 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 A10:

$$
\begin{aligned}
||\hat{p}-p||_{\infty} &= O_P\left(\max\left(h_x,h_y\right)^m\right)+O_P\left(\sqrt{\frac{\log n}{nh_x^{d_x}}}\right),\\
||\hat{p}_1-p_1||_{\infty} &= O_P\left(\max\left(h_x,h_y\right)^m\right)+O_P\left(\sqrt{\frac{\log n}{nh_x^{d_x+1}}}\right),\\
||\hat{p}_y-p_y||_{\infty} &= O_P\left(\max\left(h_x,h_y\right)^m\right)+O_P\left(\sqrt{\frac{\log n}{nh_yh_x^{d_x}}}\right),\\
||\hat{f}-f||_{\infty} &= O_P\left(h_x^m\right)+O_P\left(\sqrt{\frac{\log n}{nh_x^{d_x}}}\right),\\
||\hat{f}_1-f_1||_{\infty} &= O_P\left(h_x^m\right)+O_P\left(\sqrt{\frac{\log n}{nh_x^{d_x+1}}}\right).
\end{aligned}
\tag{44}
$$

Now, under Assumption A11, we see that the squared uniform estimation error of the kernel estimators $\hat{p}$ and $\hat{f}$ and their relevant derivatives all are of order $o_P\left(1/\sqrt{n}\right)$. Given the definition of $R\left(y,x\right)$, this completes the proof. □

Lemma 2. *Under Assumptions A1-A11, the following holds uniformly over $y, y_0 \in \mathcal{Y}_0$ for any continuous function $\bar{w}\left(y,x\right)$ with compact support contained in $\mathcal{X}_0$:*

$$\int_{\mathcal{X}}\bar{w}\left(y_0,x\right)\{\triangledown_p S\left(y,x\right)\left[\hat{p}-p\right]+\triangledown_f S\left(y,x\right)[\hat{f}-f]\}dx=\frac{1}{n}\sum_{i=1}^{n}\delta_i^{\bar{w}}\left(y_0,y\right)+o_P\left(1/\sqrt{n}\right),$$

where $\delta_i^{\bar{w}}\left(y_0,y\right)$ is defined in Equation (33).

Proof of Lemma 2. First note that $\triangledown_p S\left(y,x\right)\left[p\right]+\triangledown_f S\left(y,x\right)\left[f\right]=0$. Next,

$$
\begin{aligned}
\triangledown_p S\left(y,x\right)\left[\hat{p}\right] &= \int_0^y D_{p,0}\left(u,x\right)\hat{p}\left(u,x\right)du+\int_0^y D_{p,y}\left(u,x\right)\hat{p}_y\left(u,x\right)du\\
&\quad+\int_0^y D_{p,1}\left(u,x\right)\hat{p}_1\left(u,x\right)du\\
&\equiv A_1\left(y,x\right)+A_2\left(y,x\right)+A_3\left(y,x\right).
\end{aligned}
$$

Here, uniformly over $y \in \mathcal{Y}_0$, $u \leq y$, $u \geq 0$ and $u \geq Y_i$

$$\begin{aligned} A_1(y,x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x}(X_i - x)\int_0^y D_{p,0}(u,x)\mathcal{K}_{h_y}\{Y_i - u\}\,du \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x}(X_i - x)\left[\int_0^y D_{p,0}(u,x)\mathbb{I}\{Y_i \leq u\}\,du + O_P\left(h_y^m\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x}(X_i - x)\left[\int_{\max\{0,Y_i\}}^y D_{p,0}(u,x)\,du + O_P\left(h_y^m\right)\right], \end{aligned}$$

$$\begin{aligned} A_2(y,x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x}(X_i - x)\int_0^y D_{p,y}(u,x)K_{h_y}\{Y_i - u\}\,du \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x}(X_i - x)\left[\mathbb{I}\{0 \leq Y_i \leq y\}D_{p,y}(Y_i,x) + O_P\left(h_y^m\right)\right], \end{aligned}$$

and, with $\mathbf{K}_{h_x,1}(X_i - x) = \partial\mathbf{K}_{h_x}(X_i - x)/(\partial x_1)$,

$$A_3(y,x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x,1}(X_{-1,i} - x_{-1})\int_0^y D_{p,1}(u,x)K_{h_y}\{Y_i - u\}\,du$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x,1}(X_{-1,i} - x_{-1})\left[\mathbb{I}\{0 \leq Y_i \leq y\}D_{p,y}(Y_i,x) + O_P\left(h_y^m\right)\right].$$

Thus,

$$\begin{aligned} &\int_{\mathcal{X}} \bar{w}(y_0,x)A_1(y,x)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{\max\{0,Y_i\}}^y\int_{\mathcal{X}} \bar{w}(y_0,x)D_{p,0}(u,x)\mathbf{K}_{h_x}(X_i - x)\,dxdu \times \left[1 + O_P\left(h_y^m\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\bar{w}(y_0,X_i)\int_{\max\{0,Y_i\}}^y D_{p,0}(u,X_i)\,du \times \left[1 + O_P\left(h_y^m\right) + O_P\left(h_x^m\right)\right], \end{aligned}$$

$$\begin{aligned} &\int_{\mathcal{X}} \bar{w}(y_0,x)A_2(y,x)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{0 \leq Y_i \leq y\}\int_{\mathcal{X}} \bar{w}(y_0,x)\mathbf{K}_{h_x}(X_i - x)D_{p,y}(Y_i,x)\,dx + O_P\left(h_y^m\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{0 \leq Y_i \leq y\}\bar{w}(y_0,X_i)D_{p,y}(Y_i,X_i)\left[1 + O_P\left(h_y^m\right) + O_P\left(h_x^m\right)\right], \end{aligned}$$

and

$$\begin{aligned}
&\int_{\mathcal{X}} \bar{w}\left(y_0, x\right) A_3\left(y, x\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{0 \leq Y_i \leq y\right\} \int_{\mathcal{X}} \mathbf{K}_{h_x,1}\left(X_{-1,i} - x_{-1}\right) \bar{w}\left(y_0, x\right) D_{p,y}\left(Y_i, x\right) dx \\
&\quad \times \left[1 + O_P\left(h_y^m\right)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{0 \leq Y_i \leq y\right\} \int_{\mathcal{X}} \mathbf{K}_{h_x}\left(X_i - x\right) \frac{\partial}{\partial x_1}\left[\bar{w}\left(y_0, x\right) D_{p,1}\left(Y_i, x\right)\right] dx \\
&\quad \times \left[1 + O_P\left(h_y^m\right)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{0 \leq Y_i \leq y\right\} \frac{\partial\left[\bar{w}\left(y_0, X_i\right) D_{p,1}\left(Y_i, X_i\right)\right]}{\partial x_1}\left[1 + O_P\left(h_y^m\right) + O_P\left(h_x^m\right)\right].
\end{aligned}$$

By similar arguments,

$$\begin{aligned}
&\int_{\mathcal{X}} \bar{w}\left(y_0, x\right) \bigtriangledown_f S\left(y, x\right)[\hat{f}]dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_0^y \left\{\int_{\mathcal{X}} \left[\bar{w}\left(y_0, x\right) D_{f,0}\left(u, x\right) + \frac{\partial\left[\bar{w}\left(y_0, x\right) D_{f,1}\left(u, x\right)\right]}{\partial x_1}\right] \mathbf{K}_{h_x}\left(X_i - x\right) dx\right\} du \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{w}\left(y_0, X_i\right) \int_0^y D_{f,0}\left(u, X_i\right) du + \frac{1}{n}\sum_{i=1}^{n} \int_0^y \frac{\partial\left[\bar{w}\left(y_0, X_i\right) D_{f,1}\left(u, X_i\right)\right]}{\partial x_1} du + O_P\left(h_x^m\right).
\end{aligned}$$

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

Lemma 3. *Under Assumptions A1-A11:*

$$\left\|\bigtriangledown_p S\left[\hat{p} - p\right]\right\|_\infty^2 = o_P\left(1/\sqrt{n}\right), \quad \left\|\bigtriangledown_f S[\hat{f} - f]\right\|_\infty^2 = o_P\left(1/\sqrt{n}\right).$$

Proof of Lemma 3. From the definition of $\bigtriangledown_p S\left(y, x\right)\left[\hat{p} - p\right]$,

$$\left\|\bigtriangledown_p S\left[\hat{p} - p\right]\right\|_\infty \leq \left\|D_{p,0}\right\|_\infty \left\|\hat{p} - p\right\|_\infty + \left\|D_{p,y}\right\|_\infty \left\|\hat{p}_y - p_y\right\| + \left\|D_{p,1}\right\|_\infty \left\|\hat{p}_1 - p_1\right\|,$$

where $\left\|D_{p,a}\right\|_\infty < \infty$, $a = 0, y, 1$, given the smoothness and bound conditions imposed in Assumption A10. Next, it follows from the convergence rate results in Equation (44) together with the bandwidth requirement in Assumption A11 that $\left\|\hat{p} - p\right\|_\infty = o_P\left(1/n^{1/4}\right)$ and similarly for its partial derivatives with respect to y and x_1. This proves the first claim. The proof of the second claim follows along the same lines and so is left out. □

Appendix D. Identification without continuity

In this section of the Appendix, we exhibit a proof of nonparametric identification of Θ that does not rely on the continuity of the exogenous regressor. The proof strategy closely follows that of Ridder (1990). We first strengthen our Assumptions A1 and A5.

Assumption A1'. *For a.e. $x \in \mathcal{X}$, the conditional distribution $F_{\epsilon|X}(\cdot|x)$ of ϵ given $X = x$ is absolutely continuous (with respect to the Lebesgue measure on $\mathbb{R}$) with a density $f_{\epsilon|X}(\cdot|x)$ that is continuous and strictly positive on $\mathbb{R}$.*

Assumption A1' imposes that conditional on $X = x$, ϵ has full support on $\mathbb{R}$. This is a strengthening of our Assumption A1 which amounts to setting $\mathcal{E}_x = \mathbb{R}$. Next, we replace the continuity Assumption A2(ii) with the following support condition.

Assumption A2'. *(ii) The support of X_I given $X_{-I} = x_{0,-I}$ contains at least two distinct points $x_{0,I}$ and $x_{1,I}$.*

The following assumption requires that g takes distinct values at points in Assumption A2', with the value at $x_0 = (x_{0,I}, x_{0,-I})$ normalized to zero.

Assumption A5'. *For $x_0 = (x_{0,I}, x_{0,-I})$ and $x_1 = (x_{1,I}, x_{0,-I})$, $g(x_0) = 0 \neq g(x_1)$.*

The nonparametric identification result is then as follows:

Theorem 6. *Let Assumptions A1', A2(i), A2'(ii), A3, A4, and A5' hold. Then either of the normalization conditions N1, N2 and N3 is sufficient to nonparametrically identify Θ.*

Proof. Consider two observationally equivalent structures $(\Theta, g, F_{\epsilon|X})$ and $(\tilde{\Theta}, \tilde{g}, \tilde{F}_{\tilde{\epsilon}|X})$. Evaluating Equation (3) at $x_0 = (x_{0,I}, x_{0,-I})$ (x_0 as in Assumption A5') and $x_1 \equiv (x_{1,I}, x_{0,-I})$ we obtain:

$$\Phi(y, x_0) = F_{\epsilon|X}\big(\Theta(y), x_{0,-I}\big) = \tilde{F}_{\tilde{\epsilon}|X}\big(\tilde{\Theta}(y), x_{0,-I}\big)$$

$$\Phi(y, x_1) = F_{\epsilon|X}\big(\Theta(y) - g(x_1), x_{0,-I}\big) = \tilde{F}_{\tilde{\epsilon}|X}\big(\tilde{\Theta}(y) - \tilde{g}(x_1), x_{0,-I}\big)$$

Note that under conditional independence assumption A2 guarantees that the conditional distributions above only do not depend on the values $x_{0,I}$ and $x_{1,I}$ of the exogenous regressors, which in turn are the only coordinates that vary from x_0 to x_1. Now consider the change of variable $t = \Theta(y)$ in the first equation and $t = \Theta(y) - g(x_1)$ in the second equation. Then,

$$\begin{aligned} F_{\epsilon|X}\big(t, x_{0,-I}\big) &= \tilde{F}_{\tilde{\epsilon}|X}\big((\tilde{\Theta} \circ \Theta^{-1})(t), x_{0,-I}\big) \\ F_{\epsilon|X}\big(t, x_{0,-I}\big) &= \tilde{F}_{\tilde{\epsilon}|X}\big((\tilde{\Theta} \circ \Theta^{-1})(t + g(x_1)) - \tilde{g}(x_1), x_{0,-I}\big) \end{aligned}$$

and by virtue of Assumption A1', the above needs to hold for every $t \in \mathbb{R}$. Since $\tilde{F}_{\tilde{\epsilon}|X}$ is strictly increasing on $\mathbb{R}$, the above implies

$$\kappa(t + g(x_1)) = \kappa(t) + \tilde{g}(x_1) \quad \text{where} \quad \kappa \equiv \tilde{\theta} \circ \Theta^{-1}.$$

Using the same reasoning as in the proof of Theorem 1 in Ridder (1990) (equations (13)-(23) on p.180) then shows that

$$\tilde{\Theta} = \gamma + \delta\Theta, \quad \delta > 0, \gamma \in \mathbb{R},$$

that is Θ is nonparametrically identified up to a location and a scale. Any of the normalizations N1-N3 is then sufficient to pin down γ and δ. □

Appendix E. Integral Normalization $\Theta(0) = 0$ and $E[\Theta(Y)] = 1$

In this section, we consider yet a third normalization:

$$\text{(N3)} \qquad \Theta(0) = 0 \quad \text{and} \quad E[\Theta(Y)] = 1.$$

Our nonparametric identification result is as follows:

Corollary 2. *Let all the assumptions of Theorem 1 hold, and consider*

$$E_Y[S_i\,(Y, x)] = \int_{\mathcal{Y}} S_i\,(y, x)\, f_Y\,(y)\, dy.$$

Then, under normalization (N3), Θ *is globally identified as:*

$$\text{(45)} \qquad \Theta(y) = \psi_i\,(y, x), \quad \psi_i\,(y, x) \equiv \frac{S_i\,(y, x)}{E_Y\,[S_i\,(Y, x)]},$$

and the right-hand side of (45) *does not depend on i nor x.*

Proof of Corollary 2. We use the same reasoning as in the proof of Theorem 1 up to equation (22). To get rid of the g term we now use a different approach. Multiplying (22) by the pdf $f_Y(\cdot)$ of Y and then integrating with respect to y, we get:

$$1 = E[\Theta(Y)] = -\frac{\partial g(\bar{x})}{\partial x_i} \int_{\mathcal{Y}} S_i(y, \bar{x}) f_Y(y) dy = -\frac{\partial g(\bar{x})}{\partial x_i} E_Y[S_i(Y, \bar{x})],$$

where we have used the fact that under normalization N3 $E[\Theta(Y)] = 1$. Since $\bar{x} \in \mathcal{A}_i$, $\partial g(\bar{x})/\partial x_i \neq 0$ and is finite; hence, $E_Y[S_i(Y, \bar{x})] \neq 0$ and is finite as well, so we can write:

$$\frac{\partial g(\bar{x})}{\partial x_i} = -\frac{1}{E_Y[S_i(Y, \bar{x})]}. \tag{46}$$

Thus, $\Theta\left(y\right)$ is identified under (N3) by

$$\Theta(y) = \frac{S_i(y, \bar{x})}{E_Y[S_i(Y, \bar{x})]}. \tag{47}$$

To show that the right-hand side of (47) does not depend on i nor $\bar{x}$ use the same reasoning as in Step 1 of the proof of Theorem 1 to establish that for all $y \in \mathcal{Y}$,

$$\frac{s_i(y, \bar{x})}{E_Y[S_i(Y, \bar{x})]} = \frac{s_j(y, \tilde{x})}{E_Y[S_j(Y, \tilde{x})]},$$

where j and $\tilde{x}$ are as in Step 1 of the proof of Theorem 1. Combining the above with the expression for Θ in (47) then yields the result. □