NONPARAMETRIC IDENTIFICATION AND ESTIMATION OF TRANSFORMATION MODELS

PIERRE-ANDRE CHIAPPORI, IVANA KOMUNJER, AND DENNIS KRISTENSEN

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

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

[Preliminary and Incomplete. Please do not circulate.]

Date: September 17, 2010.

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

Paper presented at seminars at MIT, Princeton, and the University of Chicago. We thank all the participants, Victor Chernozhukov, Aureo de Palma, Jim Heckman, Jerry Hausman, Stefan Hoderlein, Bo Honore, Arthur Lewbel, Whitney Newey, and Bernard Salanié for useful comments. Errors are ours.
1. Introduction

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

The present paper focuses on the following two questions. First, under what conditions is the transformation model nonparametrically identified? And second, is it possible to estimate the model at the parametric rate? Regarding the first question, our main result is to show that transformation models are nonparametrically globally identified under conditions that are significantly weaker than full independence. Our identification strategy is constructive in a sense that it naturally leads to a nonparametric estimator for the model. This then leads to the second main result of the paper, which is to show that our nonparametric estimator attains the parametric convergence rate.

We now discuss how the identification result of our paper relates to the literature. It is well-known that in nonparametric linear models \( Y = g(X) + \epsilon \), the unknown function \( g \) can be identified from \( E(\epsilon|Z) = 0 \) w.p.1 if the conditional distribution of the endogenous regressor \( X \) given the instrument \( Z \) is complete (see, e.g., Darolles, Florens, and Renault, 2002; Blundell and Powell, 2003; Newey and Powell, 2003; Hall and Horowitz, 2005; Severini and Tripathi, 2006; d'Haultfoeuille, 2006). Given that the model is linear in \( g \), this identification result is global in nature. Nothing
is said, however, about the identification of the conditional distribution $F_{\epsilon|X}$ of the disturbance.

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

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

Our identification results are close in spirit to those obtained by Ridder (1990), Ekeland, Heckman, and Nesheim (2004), and Jacho-Chávez, Lewbel, and Linton (2008). Using the independence of $\epsilon$ and $X$, Ridder (1990) establishes the nonparametric identifiability of $(Z,g,F_{\epsilon})$ in a Generalized Accelerated Failure-Time (GAFT) model $\ln Z(Y) = g(X) + \epsilon$, where $Y$ is the duration, $Z' > 0$ and $F_{\epsilon}$ is the distribution of the unobserved heterogeneity term $\epsilon$.\footnote{Duration models have been studied in Elbers and Ridder (1982), Heckman and Singer (1984b,a), Heckman and Honore (1989), Honore (1990), Ridder (1990), Heckman (1991), Honore (1993), Hahn (1994), Ridder and Woutersen (2003), Honore and de Paula (2008), and Abbring and Ridder (2010) among others; van den Berg (2001) provides a complete survey of this literature.} Letting $T \equiv \ln \circ Z$, this result is related to that of Ekeland, Heckman, and Nesheim (2004) who show that assuming $\epsilon \perp X$ is sufficient to establish nonparametric identifiability (up to unknown constants) of
$T$, $g$ and $F_\epsilon$ in a nonparametric transformation model of the kind studied here.\textsuperscript{2} A similar result has been obtained by Jacho-Chávez, Lewbel, and Linton (2008).

We extend the identification results of Ridder (1990), Ekeland, Heckman, and Nesheim (2004), and Jacho-Chávez, Lewbel, and Linton (2008) in two important directions: first, we prove nonparametric identification of the function $T$ even when the regressor $X$ contains an endogenous component; and second, we show that if there exists nonparametric instrumental variables $Z$ such that the conditional distribution of $X^{-I}$ given $Z$ is complete, then the conditional moment conditions $E(\epsilon|Z) = 0$ w.p.1 are sufficient as well as necessary to identify $g$ nonparametrically.\textsuperscript{3} It is worth pointing out that our identification strategy allows to nonparametrically identify the transformation $T$ even if the completeness assumption fails; the latter is only used to identify $g$ and $F_\epsilon|X$.

The results of this paper are also related to the literature on nonparametric identification under monotonicity assumptions (see Matzkin, 2007, for a recent survey). For example, Matzkin (2003) provides conditions under which in models of the form $Y = m(X, \epsilon)$ with $m$ strictly monotone, the independence assumption $\epsilon \perp X$ is sufficient to globally identify $m$ and $F_\epsilon$ (see also Chesher, 2003, for additional local results). In a sense, our result shows that the independence condition can be substantially relaxed, if a certain form of separability between $Y$, $X$ and $\epsilon$ holds, namely, if we have $T^{-1}(Y) = g(X) + \epsilon$.\textsuperscript{4}

\textsuperscript{2}In the same paper, the authors derive an additional result that relaxes the independence assumption and replaces it with $E(\epsilon|X) = 0$ w.p.1. They show that the latter is sufficient to identify general parametric specifications for $T(y, \phi)$ and $g(x, \theta)$ where $\phi$ and $\theta$ are finite dimensional parameters. Once $T(y, \phi)$ and $g(x, \theta)$ are specified, the results derived by Komunjer (2008) can be used to further check whether global GMM identification of $\phi$ and $\theta$ holds.

\textsuperscript{3}The results also extend those of Hoderlein (2009) who considers identification and estimation of semiparametric endogenous binary choice models in which $T(X) = \beta' X$. As shown in Hoderlein (2009), the slope parameter $\beta$ can then be identified as the mean ratio of derivatives of two functions of the instrument $Z$.

\textsuperscript{4}See also the discussion on page 24 in Blundell and Powell (2003).
We now discuss several results related to our nonparametric estimation of the transformation model. In a special case where \( g(X) = \beta' X \), Horowitz (1996) develops \( n^{1/2} \)-consistent, asymptotically normal, nonparametric estimators of \( T \) and \( F_\epsilon \). Estimators of \( \beta \) are available since Han (1987). In a special case where the transformation \( T \) is finitely parameterized by a parameter \( \phi \), Linton, Sperlich, and van Keilegom (2008) construct a mean square distance from independence estimator for the transformation parameter \( \phi \). Finally, it is worth pointing out that even though they do not provide primitive conditions for global nonparametric identification of \( \theta \) in the model \( E[\rho(Y, X, \theta)|Z] = 0 \) w.p.1, the estimation methods developed in Ai and Chen (2003) and Chernozhukov, Imbens, and Newey (2007) yield consistent estimators for \( \theta \), and are readily applicable in our setup.

IVANA: PERHAPS ADD MORE TO THE ESTIMATION PART HERE? IN PARTICULAR, I THINK WE SHOULD SAY THAT (Jacho-Chávez, Lewbel, and Linton, 2008) ONLY OBTAIN A NONPARAMETRIC RATE FOR THEIR ESTIMATOR.

The remainder of the paper is organized as follows. Section 2 introduces the transformation model and recalls basic definitions. In Section 3, we derive necessary and sufficient conditions for the model to be nonparametrically identified. Our identification strategy is constructive in a sense that it leads to a natural estimator. In Section 4 we show that the estimator converges at the parametric rate. Section 5 concludes. All of our proofs are relegated to an Appendix.

2. Model

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

\[
Y = T(g(X) + \epsilon),
\]

where \( Y \) belongs to \( \mathbb{R} \), \( X = (X_1, \ldots, X_{d_x}) \) belongs to \( \mathcal{X} \subseteq \mathbb{R}^{d_x} \), \( \epsilon \) is in \( \mathbb{R} \), and \( T : \mathbb{R} \to \mathbb{R} \) and \( g : \mathcal{X} \to \mathbb{R} \) are unknown functions. The variables \( Y \) and \( X \) are
observed, while $\epsilon$ remains latent. The support restrictions on $Y$, $X$, and $\epsilon$ shall be made explicit below.\footnote{Following the usual convention, the support of a random variable is defined as the smallest closed set whose complement has probability zero.}

Hereafter we maintain the following assumptions.

**Assumption A1.** $T$ is twice continuously differentiable on $\mathbb{R}$, $T'(t) > 0$ for every $t \in \mathbb{R}$, and $T(\mathbb{R}) = \mathbb{R}$.

We restrict our attention to the transformations $T$ in (1) that are smooth and strictly increasing from $\mathbb{R}$ onto $\mathbb{R}$. In particular, the limit conditions $\lim_{t \to \pm \infty} T(t) = \pm \infty$ hold true under assumption A1.\footnote{Our setup can easily accommodate situations in which the dependent variable $\tilde{Y} = \tilde{T}(g(X) + \epsilon)$ belongs to some known subset $\mathcal{Y}$ of $\mathbb{R}$, and $\tilde{T} : \mathbb{R} \to \mathcal{Y}$, provided $0 \in \mathcal{Y}$. It suffices to consider any strictly increasing transformation $\Lambda : \mathcal{Y} \to \mathbb{R}$ and to let $Y = \Lambda(\tilde{Y})$ and $T = \Lambda \circ \tilde{T}$. For instance, in duration models $\tilde{Y} \in [0, +\infty)$ so we can use the transformation $\Lambda = \ln$. Our limit conditions then reduce to $\lim_{t \to -\infty} \tilde{T}(t) = 0$ and $\lim_{t \to +\infty} \tilde{T}(t) = +\infty$ (see, e.g., Assumption (A-2) in Abbring and Ridder, 2010).}

**Assumption A2.** Let $\mathcal{X}$ be the support of $X$. Then, for a.e. $x \in \mathcal{X}$, the conditional distribution $F_{\epsilon|X}$ of $\epsilon$ given $X = x$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$, and has density $f_{\epsilon|X}$ that is continuously differentiable on $\mathbb{R}$ and satisfies:

$$\int_{\mathbb{R}} f_{\epsilon|X}(t,x)dt = 1 \text{ and } f_{\epsilon|X}(\cdot,x) > 0 \text{ on } \mathbb{R}$$

Assumption A2 states that for almost every realization $x \in \mathcal{X}$ of $X$, the conditional density of $\epsilon$ given $X = x$ exists, is positive and continuously differentiable over its entire support $\mathbb{R}$. This assumption, combined with the fact that $T$ is a twice differentiable homeomorphism from $\mathbb{R}$ onto $\mathbb{R}$, guarantees that the conditional distribution $F_{Y|X}$ of $Y$ given $X = x$ has support $\mathbb{R}$, is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$, and has density $f_{Y|X}(\cdot,x)$ that is positive and continuously differentiable everywhere on $\mathbb{R}$.\footnote{In particular, note that the support of the conditional distribution of $Y$ given $X = x$ does not vary with $x.$}
We now further restrict the dependence between $\epsilon$ and $X$. For this, let $X_1$ denote the first component of $X$; whenever $d_x > 1$, we denote by $X_{-1}$ the remaining sub-
vector of $X$, i.e. $X_{-1} \equiv (X_2, \ldots, X_{d_x})$. The supports of $X_1$ and $X_{-1}$ are denoted $\mathcal{X}_1$ and $\mathcal{X}_{-1}$, respectively. We make the following assumption:

**Assumption A3.** $\epsilon \perp X_1 \mid X_{-1}$.

Assumption A3 states that $\epsilon$ is independent of at least one component of $X$, given the remaining components of $X$; we may, with no loss of generality, assume that this conditionally exogenous component is $X_1$. Put in words, the property in A3 says that the variable $X_1$ is excluded from the conditional distribution of $\epsilon$ given $X$. This is why we call exclusion restriction the conditional independence assumption in A3.

**Assumption A4.** The random variable $X_1$ is continuously distributed on $\mathcal{X}_1 \subseteq \mathbb{R}$.

According to Assumption A4, the first component $X_1$ of each observable vector $X$ is continuous. Note that except for the continuity of the random variable $X_1$, A4 does not restrict its support $\mathcal{X}_1$. In particular, $\mathcal{X}_1$ need not be equal to $\mathbb{R}$, and $X_1$ may well have bounded support. Perhaps more importantly, assumption A4 allows all the other components $X_2, \ldots, X_{d_x}$ to be either continuous or discrete with bounded or unbounded supports.

**Assumption A5.** For a.e. $x \in \mathcal{X}$, the second-order partial derivative $\partial^2 g(x)/(\partial x_1)^2$ exists and is continuous; moreover, the set $\{x \in \mathcal{X} : \partial g(x)/\partial x_1 \neq 0\}$ has a nonempty interior.

Similar to A4, Assumption A5 only restricts the behavior of the partial derivatives of $g$ with respect to $x_1$. Nothing is being said about the behavior of $g$ with respect to the remaining components $x_{-1}$. The restriction on the set of points $x \in \mathcal{X}$ at which the partial derivative of $g$ with respect to $x_1$ vanishes excludes situations in which $g$ is a constant function.

In addition to the restrictions on the joint distribution of $\epsilon$ and $X_1$ conditional on $X_{-1}$ stated in Assumption A3, we now restrict the joint distribution of $\epsilon$ and $X_{-1}$. 
Assumption A6. For a.e. $z \in Z$, $E(\epsilon|Z = z) = 0$ and the conditional distribution of $X_{-1}$ given $Z = z$ is complete: for every function $h : \mathcal{X}_{-1} \to \mathbb{R}$ such that $E[h(X_{-1})]$ exists and is finite, $E[h(X_{-1}) | Z = z] = 0$ implies $h(x_{-1}) = 0$ for a.e. $x_{-1} \in \mathcal{X}_{-1}$.

Recall from A3 that $\epsilon$ is assumed to be conditionally independent of $X_1$ given $X_{-1}$, i.e. the first component of $X$ is conditionally exogenous. The other components are on the other hand allowed to be endogenous provided there exists a vector of instruments $Z$ with respect to which the distribution of $X_{-1}$ is complete, and such that $\epsilon$ is mean independent of $Z$.

Following the related literature (e.g., Koopmans and Reiersøl, 1950; Brown, 1983; Roehrig, 1988; Matzkin, 2003) we hereafter call structure a particular value of the triplet $(T, g, F_{\epsilon|X})$ in (1), where $T : \mathbb{R} \to \mathbb{R}$, $g : \mathcal{X} \to \mathbb{R}$, and $F_{\epsilon|X} : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$. The model then simply corresponds to the set of all structures $(T, g, F_{\epsilon|X})$ that satisfy the restrictions given by Assumptions A1 through A6. Each structure in the model induces a conditional distribution $F_{Y|X}$ of the observables, and two structures $(\tilde{T}, \tilde{g}, \tilde{F}_{\epsilon|X})$ and $(T, g, F_{\epsilon|X})$ are observationally equivalent if they generate the same $F_{Y|X}$. When the set of all conditional distributions $F_{Y|X}$ generated by the model contains the true conditional distribution we say that the model is correctly specified. In that case, the model contains at least one true structure $(T, g, F_{\epsilon|X})$ that induces $F_{Y|X}$. The model is then said to be identified, if the set of structures that are observationally equivalent to $(T, g, F_{\epsilon|X})$ reduces to a singleton. In what follows, we shall provide conditions under which the model (1)—assumed to be correctly specified—is also identified. We then use our identification strategy to derive an estimator for $(T, g, F_{\epsilon|X})$.

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

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

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

The conditional independence property in A3 has strong implications which we now derive. In what follows, $\Theta : \mathbb{R} \to \mathbb{R}$ denotes the inverse mapping $T^{-1}$. Under A1, $\Theta$ is twice continuously differentiable and strictly increasing on $\mathbb{R}$. Note that in addition $\Theta(\mathbb{R}) = \mathbb{R}$. Equation (1) is equivalent to $\epsilon = \Theta(Y) - g(X)$, so by $\Theta' > 0$ and the conditional independence of $\epsilon$ and $X_1$ given $X_{-1}$,

$$F_{Y|X}(y, x) = \Pr(Y \leq y \mid X = x) = \Pr(\epsilon \leq \Theta(y) - g(x) \mid X = x) = F_{\epsilon|X}(\Theta(y) - g(x), x_{-1})$$

for all $(y, x) \in \mathbb{R} \times \mathcal{X}$. The problem we now examine can then be restated as follows: to what extent is it possible to recover the functions $T : \mathbb{R} \to \mathbb{R}$, $g : \mathcal{X} \to \mathbb{R}$, and $F_{\epsilon|X} : \mathbb{R} \times \mathcal{X}_{-1} \to \mathbb{R}$, which for every $y \in \mathbb{R}$ and a.e. $x \in \mathcal{X}$ satisfy:

$$(2) \quad \Phi(y, x) = F_{\epsilon|X}(\Theta(y) - g(x), x_{-1})$$

where $\Phi(y, x) \equiv F_{Y|X}(y, x) = \Pr(Y \leq y \mid X = x)$ denotes the conditional cdf of $Y$ given $X$?

For one thing, it is clear from (1) that some normalization of the model is needed; indeed, for any $(\lambda, \mu) \in \mathbb{R}^2$, the transformation model (1) is equivalent to

$$Y = \tilde{T}(\lambda g(X) + \mu + \lambda \epsilon)$$
where $\tilde{T}$ is defined by $
abla t \equiv T ((t - \mu)/\lambda)$. We therefore impose that any structure $(T, g, F_{\epsilon|X})$ in (1) satisfies the normalization condition:

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

The main identification result is provided by the following statement:

**Theorem 1.** Let $(T, g, F_{\epsilon|X})$ and $(\tilde{T}, \tilde{g}, \tilde{F}_{\epsilon|X})$ be two observationally equivalent structures that generate $F_{Y|X}$, and satisfy assumptions A1-A5 and the normalization condition (3). Then, assumption A6 is necessary and sufficient to globally identify $(T, g, F_{\epsilon|X})$.

Theorem 1 shows two results. First, that under Assumptions A1 to A5, the completeness condition is sufficient to nonparametrically identify the transformation model (1). This identification result is global even though the model (1) is nonlinear in $g$ and $F_{\epsilon|X}$. The second result of Theorem 1 is that the completeness condition is also necessary, in the following sense: assume that there exists some function $h: \mathcal{X}_{-1} \to \mathbb{R}$ that (i) does not vanish a.e., but (ii) is such that $E[h(X_{-1}) | Z = z] = 0$ for a.e. $z \in Z$. Then there exists two different but observationally equivalent structures that generate $F_{Y|X}$ while satisfying assumptions A1-A5 and the normalization condition (3).

While necessary to identify $(g, F_{\epsilon|X})$, the completeness assumption A6 is not used to identify the transformation $T$. In fact, the proof of Theorem 1 shows that under A1-A5 and the normalization condition (3), the inverse transformation $T^{-1}$ can be written as:

$$T^{-1}(y) = \frac{\int_{y}^{\infty} \Phi_{y}(u, x)[\Phi_{1}(u, x)]^{-1}du}{\int_{\mathbb{R}} f_{Y}(y) \int_{0}^{\infty} \Phi_{y}(u, x)[\Phi_{1}(u, x)]^{-1}dudy}$$

where $x$ is any point in $\mathcal{X}$ for which $g_{1}(x) \neq 0$, as before $\Phi(y, x) = F_{Y|X}(y, x)$ is the conditional cdf of $Y$ given $X$, $f_{Y}(y)$ is the unconditional density of $Y$, and where we use subscripts to denote partial derivatives.\(^9\) Key to the identification of $T$ is the

---

\(^9\)Specifically, $g_{1}(x) = \frac{\partial g(x)}{\partial x_{1}}$, $\Phi_{y}(y, x) = \frac{\partial \Phi(y, x)}{\partial y}$ and $\Phi_{1}(y, x) = \frac{\partial \Phi(y, x)}{\partial x_{1}}$.\(^9\)
conditional independence assumption (A3) which in particular guarantees that the right hand side in (4) is not a function of $x$.

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

As Equation (4) shows, our identification strategy is constructive in a sense that it leads for a closed form expression of the unknown inverse transformation $T^{-1}$ as a function of the observables. In the next section, we examine the properties of a nonparametric estimator for $T^{-1}$ that builds on the expression in (4).

4. Estimation

Suppose we have a random sample $(Y_i, X_i, Z_i)$ ($i = 1, \ldots, n$) drawn from the transformation model in Equation (1) and that Assumptions A1 to A6 hold. We study the estimation of the unknown transformation function $T$ in Equation (1) under the normalization (3).

To develop our estimator, write the inverse $\Theta = T^{-1}$ in Equation (4) as:

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

where, we have defined

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

The estimation method that we propose is straightforward in principle: We first obtain a nonparametric estimator of the conditional cdf. We then plug this estimator into Equation (6) to obtain an estimator of $S$ and its moment which in turn are substituted into Equation (5). This yields a nonparametric estimator of $\Theta (y)$.
To be more specific, observe that the conditional cdf can be written as

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

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

$$\hat{\Phi}(y, x) = \frac{\hat{g}(y, x)}{\hat{f}(x)},$$

where

$$\hat{g}(y, x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_y}(Y_i - y) K_{h_x}(X_i - x), \quad \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(X_i - x),$$

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

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

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

where $w(x)$ is a weighting function with compact support $X_0 \subseteq X$ satisfying

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

and

$$\hat{S}(y, x) = \int_{0}^{y} \frac{\hat{g}(u, x)}{\hat{f}(u, x)} \, du \quad \text{and} \quad \hat{E}[\hat{S}(Y, x)] = \frac{1}{n} \sum_{i=1}^{n} \hat{S}(Y_i, x).$$

We note that this estimator is similar to the estimator of Horowitz (1996) who considers the semiparametric model where the regression function is restricted to
$g(x) = \beta' x$. Assuming for simplicity that $\beta$ is known such that $Z = \beta' X$ is observed, Horowitz (1996) proposes the following estimator:

$$\tilde{\Theta}(y) = -\int_0^y \int_z \omega(z) \frac{\hat{\Phi}_y(u, z)}{\hat{\Phi}_1(u, z) / \partial z} dz du,$$

where $Z$ is the support of $Z$ and $\omega(z)$ is a weighting function with compact support on $Z$. As shown in Horowitz (1996), due to the double-integration, $\tilde{\Theta}(y)$ is $\sqrt{n}$-consistent despite the fact that it relies on nonparametric estimators. Similarly, in our case we also integrate over both $y$ and $x$, and as such we expect that our proposed estimator $\hat{\Theta}(y)$ will be $\sqrt{n}$-consistent.

In order to derive the asymptotic properties of $\hat{\Theta}$ we introduce additional assumptions on the model and the kernel function used in the estimation. The kernel $K$ used to define our estimator is assumed to belong to the following class of kernel function:

**Assumption A7.** The univariate kernel $K$ is differentiable, and there exists constants $C, \eta > 0$ such that

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

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

The above class is fairly general and allow for both unbounded support. We do however require the kernel $K$ to be differentiable which rules out uniform and Epanechnikov kernels. We here allow for both standard second-order kernels ($m = 2$) such as the Gaussian one, and higher-order kernel ($m > 2$). The use of higher-order kernels in conjunction with smoothness conditions on the densities in the model allow us to control for the smoothing bias induced by the use of kernels.

The smoothness conditions that we will impose on the density of data are as follows:
**Assumption A8.** The joint density, $f_{Y,X}(y,x)$ is bounded, $m$ times differentiable w.r.t. $(y,x)$ with bounded derivatives; its $m$th order partial derivatives are uniformly continuous. Furthermore, $\sup_{(x,y)} \|(x,y)\|^b f_{Y,X}(y,x) < \infty$ for some constant $b > 0$.

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

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

**Assumption A9.** $\sqrt{n} h_x^{2m} \to 0$, $\sqrt{n} h_y^{2m} \to 0$, $\sqrt{n} h_x^{d_x+2} \log(n) \to \infty$, $\sqrt{n} h_y h_x^{d_x+1} / \log(n) \to \infty$.

Assumption A9 puts restrictions on the two bandwidths sequences which ensures that the squared estimation error of the kernel estimators $\hat{g}$ and $\hat{f}$ and their relevant derivatives all are of order $o_P(1/\sqrt{n})$. As is standard for kernel estimators, there is a curse-of-dimensionality which appears in the last two restrictions on $h_x$: When the dimension of $X$, $d_x \geq 1$, is large, we in general need to use higher-order kernels in order for all four conditions to hold simultaneously.

Under the above conditions, we obtain the following asymptotic distribution of the proposed estimator:

**Theorem 2.** Let Assumptions A1 through A9 hold. Then,

$$\sqrt{n}(\hat{\Theta}(y) - \Theta(y)) \to^d N\left(0, \mathbb{E}\left[\delta_i^w(y)^2\right]\right),$$

where

$$\delta_i^w(y) = \psi_i^{w_1}(y) - \varphi_i^{w_2}(y),$$
with \( \bar{w}_1 (x) = w(x) / E [S(Y, x)] \), \( \bar{w}_2 (x) = w(x) / E [S(Y, x)]^2 \) and

\[
\psi_i^{\bar{w}} (y) := \bar{w}(X_i) \int_{Y_i} y D_{g,0} (u, X_i) \, du + \mathbb{1} \{ Y_i \leq y \} \bar{w}(X_i) D_{g,y} (Y_i, X_i) + \int_{Y_i} \frac{\partial [\bar{w}(X_i) D_{g,1} (u, X_i)]}{\partial x_1} \, du,
\]

\[
\varphi_i^{\bar{w}} := \bar{\psi}_i^{\bar{w}} + \int_X \bar{w}(x) (S(Y_i, x) - E[S(Y, x)]) \, dx, \quad \bar{\psi}_i^{\bar{w}} := E[\psi_i^{\bar{w}} (Y_j) | X_i, Y_i].
\]

Here, the functions \( D_{g,0} (y, x), D_{g,y} (y, x) \) and \( D_{g,1} (y, x) \) are defined in eqs. (17).

Discussion [TBC]: Efficiency (minimization of \( E[\delta_i^{w} (y)^2] \) as a functional of \( w \)). This looks like a rather complicated affair given the complex expression of the influence function \( \delta_i (y) \). I’ve tried to obtain a simpler expression of the function, but so far without much success :(

The above result only shows pointwise weak convergence of the estimator. We now extend this result to weak convergence of the stochastic process \( \sqrt{n} (\hat{\Theta}(\cdot) - \Theta(\cdot)) \):

**Theorem 3. [To be completed...]** Let Assumptions A1 through A9 hold. Then for any compact set \( \mathcal{Y}_0 \subset \mathbb{R} \):

\[
\sqrt{n} (\hat{\Theta}(\cdot) - \Theta(\cdot)) \overset{d}{\to} Z(\cdot),
\]

where \( \{ Z(y) : y \in \mathcal{Y}_0 \} \) is a Gaussian process with covariance kernel

\[
\Sigma(y_1, y_2) = E[\delta_i^{w} (y_1) \delta_i^{w} (y_2)],
\]

and \( \delta_i (y) \) is defined in Theorem 2.

5. Conclusion

We conclude by discussing a couple of extensions of our identification result. First, note that if the function \( g \) in the model (1) is further assumed to be bounded, then the completeness condition in assumption A6 can be replaced by a bounded completeness condition: for every bounded function \( h : \mathcal{X}_1 \to \mathbb{R}, E[h(X_{-1}) | Z] = 0 \) w.p.1 implies \( h(x_{-1}) = 0 \) for a.e. \( x_{-1} \in \mathcal{X}_{-1} \). The bounded completeness condition is weaker then
the completeness condition (see, e.g., Blundell, Chen, and Kristensen, 2007, for a discussion).

Second, assume that instead of relying on the conditional independence between \( \epsilon \) and \( X_1 \) given \( X_{-1} \), we use the fact that there exists an instrument \( V \), such that \( \epsilon \) and \( X_1 \) are conditionally independent given \( (X_{-1}, V) \), i.e. \( \epsilon \perp X_1 \mid (X_{-1}, V) \). This would amount to considering the conditional distribution \( F_{Y \mid X,V} \) of \( Y \) given \( (X, V) \) which now satisfies:

\[
F_{Y \mid X,V}(y, x, v) \equiv \Phi(y, x, v) = F_{\epsilon \mid X,V}(\Theta(y) - g(x), x_{-1}, v)
\]

Redefining \( X \) to be \( (X, V) \), the above expression falls exactly in the framework obtained in (2), with an additional restriction on the function \( g \) which now no longer depends on the components of \( X \) corresponding to \( V \). When the conditional distribution of the redefined vector \( X_{-1} \) given \( Z \) is complete, we know that \( g \) is identifiable. This identification result holds even without restricting the way that \( g \) depends on \( V \); a fortiori, the identification result remains true when \( g \) is restricted.
Appendix A. Proofs

Proof of Theorem 1. Consider a structure \((T, g, F_{|X})\) that satisfies assumptions A1-A5, and generates \(\Phi(y, x)\) in the sense of Equation (2). To establish that the normalization condition in (3) is sufficient to identify \((T, g, F_{|X})\) we proceed in two steps: the first establishes identification of \(\Theta\) while the second shows identification of \(g\) and \(F_{|X}\). The third and final step shows that the normalization condition is also necessary.

Step 1: Identification of \(\Theta\). Under assumptions A2, A4 and A5, the partial derivatives \(\partial \Phi(y, x)/\partial y\), \(\partial \Phi(y, x)/\partial x_1\) exist and are continuous. Differentiating Equation (2) in \(y\) and \(x_1\) gives:

\[
\frac{\partial \Phi}{\partial y}(y, x) = \Theta'(y) \frac{\partial F_{|X}}{\partial t}(\Theta(y) - g(x), x_{-1})
\]

\[
\frac{\partial \Phi}{\partial x_1}(y, x) = -\frac{\partial g}{\partial x_1}(x) \frac{\partial F_{|X}}{\partial t}(\Theta(y) - g(x), x_{-1})
\]

where \(\Theta'\) is the derivative of \(\Theta\), and \(\partial F_{|X}/\partial t\) denotes the partial derivative of \(F_{|X}\) with respect to its first variable. Let \(A \equiv \{x \in X : \partial \Phi(y, x)/\partial x_1 \neq 0 \text{ for every } y \in \mathbb{R}\}\). Under Assumptions A2 and A5, the set \(A\) has a nonempty interior. Take any point \(x \in A\). Then for every \(y \in \mathbb{R}\), we have:

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

Note that \(s(y, x)\) is nonzero and keeps a constant sign for all \(y \in \mathbb{R}\). Integrating the above from 0 to \(y\) and using the fact that \(\Theta(0) = 0\) we have:

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

Note that \(S(y, x)\) is nonzero and keeps a constant sign for all \(y \in \mathbb{R}\); hence \(E[S(Y, x)] = \int_y S(y, x)f_Y(y)dy \neq 0\). Using the fact that \(E[\Theta(Y)] = 1\) we then get that:

\[
\frac{\partial g}{\partial x_1}(x) = \frac{1}{E[S(Y, x)]}.\]
Combining (9) and (10) then gives:

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

Hence, \( \Theta \) is identified.

**Step 2: Identification of \( g \) and \( F_{\epsilon|X} \).** Now let \( \Gamma : \mathcal{X} \to \mathbb{R} \) be defined as:

\[ \Gamma(x) = \begin{cases} 
1/E[S(Y, x)], & \text{if } x \in A, \\
0, & \text{otherwise.} 
\end{cases} \]

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

\[ \bar{g}(x_1, x_2, \ldots, x_d) = \int_c^{x_1} \Gamma(u, x_2, \ldots, x_d) du \]

where \( c \in \mathcal{X}^1 \). Obviously, any solution to \( \partial g(x)/\partial x_1 = \Gamma(x) \) must have the same partial in \( x_1 \) as \( \bar{g} \) in (12); it must therefore be of the form:

\[ g(x) = \bar{g}(x) + \beta(x_{-1}) \]

for some function \( \beta : \mathcal{X}_{-1} \to \mathbb{R} \). Now let \( g \) be an arbitrary solution, and consider \( E(\epsilon|Z) \) where \( \epsilon = \Theta(Y) - g(X) \) with \( \Theta \) as in (11) and \( g \) as in (13). Letting \( F_{Y|Z} \) and \( F_{X|Z} \) denote the conditional distributions of \( Y \) given \( Z \) and of \( X \) given \( Z \), respectively, we have:

\[ E(\epsilon|Z = z) = \int_Y \Theta(y) dF_{Y|Z}(y, z) - \int_X g(x) dF_{X|Z}(x, z) \]

\[ = \int_Y \Theta(y) dF_{Y|Z}(y, z) - \int_X [\bar{g}(x) + \beta(x_{-1})] dF_{X|Z}(x, z) \]

Now, consider a structure \((\tilde{T}, \tilde{g}, F_{\tilde{\epsilon}|X})\) that is observationally equivalent to \((T, g, F_{\epsilon|X})\) and has the same properties as \((T, g, F_{\epsilon|X})\). It follows from (14) that for a.e. \( z \in Z \):

\[ E(\epsilon|Z = z) = 0 = E(\tilde{\epsilon}|Z = z) \Rightarrow \int_X [\beta(x_{-1}) - \tilde{\beta}(x_{-1})] dF_{X|Z}(x, z) = 0 \]

\[ \Rightarrow E((\beta(X_{-1}) - \tilde{\beta}(X_{-1})|Z = z) = 0 \]
where \( \tilde{\epsilon} = \tilde{\Theta}(Y) - \tilde{g}(X) \). Then, the completeness assumption A6 implies \( \beta(x_{-1}) = \tilde{\beta}(x_{-1}) \) for a.e. \( x_{-1} \in X_{-1} \). Combined with equation (13), this implies that

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

Thus \( g \) is identified. From \( F_{\epsilon|X}(\Theta(y) - g(x), x_{-1}) = \tilde{F}_{\epsilon|X}(\Theta(y) - g(x), x_{-1}) \) for every \( y \in Y \) and a.e. \( x \in X \), and the fact that \( \Theta(Y) = \mathbb{R} \), we conclude that for every \( t \in \mathbb{R} \) and a.e. \( x_{-1} \in X_{-1} \), \( F_{\epsilon|X}(t, x_{-1}) = \tilde{F}_{\epsilon|X}(t, x_{-1}) \), which establishes the identification of \( F_{\epsilon|X} \) and completes the proof of sufficiency.

**Step 3. Necessity.** Finally, assume that the completeness condition is violated, in the sense that there exists some function \( h : X_{-1} \to \mathbb{R} \) that (i) does not vanish a.e., but (ii) is such that \( E[h(X_{-1}) \mid Z = z] = 0 \) for a.e. \( z \in Z \). Let \((T, g, F_{\epsilon|X})\) be a structure generating \( \Phi \), that satisfies assumptions A1-A5 and the normalization condition (3). Define \( (\tilde{T}, \tilde{g}, \tilde{F}_{\epsilon|X}) \) by

\[
\begin{align*}
    \tilde{\Theta}(y) &\equiv \Theta(y) \\
    \tilde{g}(x) &\equiv g(x) + h(x_{-1}) \\
    \tilde{F}_{\epsilon|X}(t, x) &\equiv F_{\epsilon|X}(t + h(x_{-1}), x_{-1})
\end{align*}
\]

for every \( y \in Y \), every \( t \in \mathbb{R} \), and a.e. \( x \in X \). Then, the structure \((\tilde{T}, \tilde{g}, \tilde{F}_{\epsilon|X})\) satisfies the normalization condition (3), as well as assumptions A1-A5. Note that assumption A5 only requires \( \tilde{g} \) to be smooth with respect to the first component \( x_1 \); hence, it is satisfied even if the function \( h(x_{-1}) \) is discontinuous. Since the structure \((\tilde{T}, \tilde{g}, \tilde{F}_{\epsilon|X})\) is observationally equivalent to \((T, g, F_{\epsilon|X})\), \((T, g, F_{\epsilon|X})\) is not identified. \( \Box \)
Proof of Theorem 2. Write

\[ \hat{\Theta}(y) - \Theta(y) = \int_{x_0} \frac{w(x)}{E[S(Y,x)]} \left\{ \hat{S}(y,x) - S(y,x) \right\} dx \]

\[ = \int_{x_0} \frac{w(x)}{E[S(Y,x)]} \left\{ \hat{S}(y,x) - S(y,x) \right\} dx \]

\[ - \int_{x_0} \frac{w(x)}{E[S(Y,x)]^2} \left\{ \hat{E}[\hat{S}(Y,x)] - E[S(Y,x)] \right\} dx \]

\[ + O \left( \left\| \hat{S} - S \right\|_\infty^2 \right) + O \left( \left\| \hat{E}[\hat{S}] - E[S] \right\|_\infty^2 \right), \]

where \( \|g\|_\infty \) denote the supremum norm of a given function \( g \), \( \|g\|_\infty = \sup_{u \in U} |g(u)| \).

Applying in turn Lemmas 1 and 2,

\[ \int_{x_0} \frac{w(x)}{E[S(Y,x)]} \left\{ \hat{S}(y,x) - S(y,x) \right\} dx \]

\[ = \int_{x_0} \frac{w(x)}{E[S(Y,x)]} \left\{ \nabla_y S(y,x) [\hat{g} - g] + \nabla_f S(y,x) [\hat{f} - f] \right\} dx + o_P \left( 1/\sqrt{n} \right) \]

\[ = \frac{1}{n} \sum_{i=1}^n \psi^\theta_i (y) + o_P \left( 1/\sqrt{n} \right), \]

with \( \psi^\theta_i (y) \) defined in Lemma 2 and \( \bar{w}_1 (x) := w(x) / E[S(Y,x)] \). Next, from Lemma 3, we obtain

\[ \int_{x_0} \frac{w(x)}{E[S(Y,x)]^2} \left\{ \hat{E}[\hat{S}(Y,x)] - E[S(Y,x)] \right\} dx = \frac{1}{n} \sum_{i=1}^n \psi^{\theta^2}_i (y) + o_P \left( 1/\sqrt{n} \right) \]

where \( \psi^{\theta^2}_i (y) \) is defined in Lemma 3, and the weight function \( \bar{w}_2 \) is given by \( \bar{w}_2 (x) := w(x) / E[S(Y,x)]^2 \). Finally, by Lemmas 1 and 4, \( \left\| \hat{S} - S \right\|_\infty^2 = o_P \left( 1/\sqrt{n} \right) \)

and \( \left\| \hat{E}[\hat{S}] - E[S] \right\|_\infty^2 = o_P \left( 1/\sqrt{n} \right) \). In total,

\[ \sqrt{n}(\hat{\Theta}(y) - \Theta(y)) = \frac{1}{n} \sum_{i=1}^n \delta_i (y) + o_P \left( 1/\sqrt{n} \right), \]

where \( \delta_i (y) \) is defined in the theorem. The theorem now follows by the CLT for i.i.d. sequences with \( \Omega(y) = E[\delta_i^2 (y)] \).
We define the following functionals for any functions \( dg(y,x) \) and \( df(x) \):

\[
\nabla g S(y,x) [dg] := \int_0^y D_{g,0}(u,x) \, dg(u,x) \, du + \int_0^y D_{g,y}(u,x) \, dg_y(u,x) \, du,
\]

\[
\nabla f S(y,x) [df] := \int_0^y D_{f,0}(u,x) \, df(u,x) \, du + \int_0^y D_{f,1}(u,x) \, df_1(u,x) \, du,
\]

where

\[
D_{g,0}(y,x) := \frac{\Phi_{0,y}(y,x) \, f_{0,1}(x)}{\Phi_{0,1}(y,x) \, f_0^2(x)}, \quad D_{g,y}(y,x) := \frac{1}{f_0(x) \, \Phi_{0,1}(y,x)},
\]

\[
D_{f,0}(y,x) := \frac{g_{0,y}(y,x)}{f_0(x)} \, \Phi_{0,1}(y,x) - \frac{\Phi_{0,y}(y,x)}{\Phi_{0,1}^2(y,x)} \left( \frac{2g_0(y,x)}{f_0^2(x)} \, f_{0,1}(x) + \frac{g_{0,1}(y,x)}{f_0(x)} \right),
\]

\[
D_{f,1}(y,x) := \frac{\Phi_{0,y}(y,x) \, g_0(y,x)}{\Phi_{0,1}(y,x) \, f_0^2(x)}, \quad D_{g,1}(y,x) := -\frac{\Phi_{0,y}(y,x)}{f_0(x) \, \Phi_{0,1}^2(y,x)}.
\]

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

**Lemma 1.** Under Assumptions A1-A9: With \( \nabla g S(y,x) [dg] \) and \( \nabla f S(y,x) [df] \) defined in eqs. (15)-(16), the resulting expansion holds:

\[
\hat{S}(y,x) - S(y,x) = \nabla g S(y,x) [\hat{g} - g] + \nabla f S(y,x) [\hat{f} - f] + o_P \left( \frac{1}{\sqrt{n}} \right),
\]

uniformly over \((x,y)\) in any compact interval.

**Proof of Lemma 1.** First, by a standard Taylor expansion (where we suppress dependence on \( y \) and \( x \))

\[
\frac{\Phi_y}{\Phi_1} - \frac{\Phi_{0,y}}{\Phi_{0,1}} = \frac{1}{\Phi_{0,1}} \{ \Phi_y - \Phi_{0,y} \} - \frac{\Phi_{0,y}}{\Phi_{0,1}^2} \{ \Phi_1 - \Phi_{0,1} \}
\]

\[
+ O \left( \| \Phi_y - \Phi_{0,y} \|_\infty^2 \right) + O \left( \| \Phi_1 - \Phi_{0,1} \|_\infty^2 \right),
\]
Next, recall that $\Phi = g/f$. Thus, its derivatives w.r.t. $y$ and $x_1$ respectively are on the form

$$\Phi_y = \frac{g_y}{f}, \quad \Phi_1 = \frac{g_1}{f} - \frac{gf_1}{f^2}.$$  

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

$$\Phi_y - \Phi_{0,y} = \frac{1}{f_0} \{g_y - g_{0,y}\} + \frac{g^2_{0,y}}{f_0} \{f - f_0\} + O \left( \|g_y - g_{0,y}\|_\infty^2 \right) + O \left( \|f - f_0\|_\infty^2 \right),$$

and

$$\Phi_1 - \Phi_{0,1}$$

$$= \frac{1}{f_0} \{g_1 - g_{0,1}\} + \left( \frac{2g_{0,f_0,1}}{f_0^3} + \frac{g^2_{0,1}}{f_0} \right) \{f - f_0\} - \frac{g_0}{f_0^2} \{f_1 - f_{0,1}\} - \frac{f_{0,1}}{f_0^2} \{g - g_0\}$$

$$+ O \left( \|g - g_0\|_\infty^2 \right) + O \left( \|g_1 - g_{0,1}\|_\infty^2 \right) + O \left( \|f - f_0\|_\infty^2 \right) + O \left( \|f_1 - f_{0,1}\|_\infty^2 \right).$$

Combining these expressions we now obtain

$$\frac{\Phi_y - \Phi_{0,y}}{\Phi_1 - \Phi_{0,1}} = \frac{1}{f_0 \Phi_{0,1}} \{g_y - g_{0,y}\} + \frac{g^2_{0,y}}{f_0 \Phi_{0,1}} \{f - f_0\} - \frac{\Phi_{0,y}}{f_0 \Phi_{0,1}^2} \{g_1 - g_{0,1}\}$$

$$- \frac{\Phi_{0,1}}{\Phi_{0,1}^2} \left( \frac{2g_{0,f_0,1}}{f_0^3} + \frac{g^2_{0,1}}{f_0} \right) \{f - f_0\}$$

$$+ \frac{\Phi_{0,1} g_0}{\Phi_{0,1}^2 f_0} \{f_1 - f_{0,1}\} + \frac{\Phi_{0,y} f_{0,1}}{\Phi_{0,1}^2 f_0^2} \{g - g_0\} + R$$

$$= D_{g,0} \{g - g_0\} + D_{g,y} \{g_y - g_{0,y}\} + D_{g,1} \{g_1 - g_{0,1}\}$$

$$+ D_{f,0} \{f - f_0\} + D_{f,1} \{f_1 - f_{0,1}\} + R,$$

where $R$ is the remainder term satisfying

$$R = O \left( \|g - g_0\|_\infty^2 \right) + O \left( \|g_1 - g_{0,1}\|_\infty^2 \right) + O \left( \|g_y - g_{0,y}\|_\infty^2 \right)$$

$$+ O \left( \|f - f_0\|_\infty^2 \right) + O \left( \|f_1 - f_{0,1}\|_\infty^2 \right)$$

and $D_{g,0}, D_{g,y}, D_{g,1}, D_{f,0}$ and $D_{f,1}$ are defined in eqs. (17). Given the definitions of $\nabla g S (y, x) [dg]$ and $\nabla f S (y, x) [df]$, we now obtain

$$\hat{S} (y, x) - S (y, x) = \nabla g S (y, x) [\hat{g} - g] + \nabla f S (y, x) \left[ \hat{f} - f \right] + R,$$
and what remains to be shown is that the remainder term $R = o_P(1/\sqrt{n})$. 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 A7 and A8:

\[
\|\hat{g} - g_0\|_\infty = O_P(\max(h_x, h_y)^m) + O_P\left(\frac{\log(n)}{nh_x^{d_x}}\right),
\]
\[
\|g_1 - g_{0,1}\|_\infty = O_P(\max(h_x, h_y)^m) + O_P\left(\frac{\log(n)}{nh_x^{d_x+1}}\right),
\]
\[
\|g_y - g_{0,y}\|_\infty = O_P(\max(h_x, h_y)^m) + O_P\left(\frac{\log(n)}{nh_y^{d_x}}\right),
\]
\[
\|f - f_0\|_\infty = O_P(h_x^m) + O_P\left(\frac{\log(n)}{nh_x^{d_x}}\right),
\]
\[
\|f_1 - f_{0,1}\|_\infty = O_P(h_x^m) + O_P\left(\frac{\log(n)}{nh_x^{d_x+1}}\right),
\]
(18)

Now, under Assumption A9, we see that the squared estimation error of the kernel estimators $\hat{g}$ and $\hat{f}$ and their relevant derivatives all are of order $o_P(1/\sqrt{n})$. In particular, $R = o_P(1/\sqrt{n})$ which completes the proof. □

**Lemma 2.** Under Assumptions A1-A9: For any weighting function $\bar{w}$ with compact support, the functionals $\nabla g S(y, x)[dg]$ and $\nabla f S(y, x)[df]$ defined in eqs. (15)-(16) satisfy:

\[
\int_X \bar{w}(x) \left\{ \nabla g S(y, x)[\hat{g} - g] + \nabla f S(y, x)[\hat{f} - f] \right\} dx = \frac{1}{n} \sum_{i=1}^{n} \psi^{\bar{w}}_i(y) + o_P(1/\sqrt{n}),
\]

where

\[
\psi^{\bar{w}}_i(y) := \bar{w}(X_i) \int_{Y_i}^y D_{g,0}(u, X_i) du + \mathbb{1}{Y_i \leq y} \bar{w}(X_i) D_{g,y}(Y_i, X_i)
\]
\[+ \int_{Y_i}^y \frac{\partial \left[ \bar{w}(X_i) D_{g,1}(u, X_i) \right]}{\partial x_1} du.
\]
Proof of Lemma 2. By definition,

\[ \nabla g \cdot S(y, x) \, [\hat{g}] = \int_0^y D_{g,0}(u, x) \hat{g}(u, x) \, du + \int_0^y D_{g,y}(u, x) \hat{g}_y(u, x) \, du + \int_0^y D_{g,1}(u, x) \hat{g}_1(u, x) \, du \]

\[ =: \nabla_g^{(1)} S(y, x) \, [\hat{g}] + \nabla_g^{(2)} S(y, x) \, [\hat{g}] + \nabla_g^{(3)} S(y, x) \, [\hat{g}]. \]

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

\[ \nabla_g^{(1)} S(y, x) \, [\hat{g}] = \frac{1}{n} \sum_{i=1}^n K_{h_x} (X_i - x) \int_0^y D_{g,0}(u, x) K_{h_y} \{Y_i - u\} \, du \]

\[ = \frac{1}{n} \sum_{i=1}^n K_{h_x} (X_i - x) \int_0^y D_{g,0}(u, x) \mathbb{I} \{Y_i \leq u\} \, du + O_P(h_{y}^m) \]

\[ = \frac{1}{n} \sum_{i=1}^n K_{h_x} (X_i - x) \int_{Y_i}^y D_{g,0}(u, x) \, du + O_P(h_{y}^m). \]

Similarly,

\[ \nabla_g^{(2)} S(y, x) \, [\hat{g}] = \frac{1}{n} \sum_{i=1}^n K_{h_x} (X_i - x) \int_0^y D_{g,y}(u, x) K_{h_y} \{Y_i - u\} \, du \]

\[ = \frac{1}{n} \sum_{i=1}^n K_{h_x} (X_i - x) \mathbb{I} \{Y_i \leq y\} D_{g,y}(Y_i, x) + O_P(h_{y}^m) \]

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

\[ \nabla_g^{(3)} S(y, x) \, [\hat{g}] \]

\[ = \frac{1}{n} \sum_{i=1}^n K'_{h_x} (X_{1,i} - x_1) K_{-1,h_x} (X_{-1,i} - x_{-1}) \int_0^y D_{g,1}(u, x) K_{h_y} \{Y_i - u\} \, du \]

\[ = \frac{1}{n} \sum_{i=1}^n K_{h_x} (X_{1,i} - x_1) K_{-1,h_x} (X_{-1,i} - x_{-1}) \int_{Y_i}^y D_{g,1}(u, x) \, du + O_P(h_{y}^m). \]
Thus,
\[
\int_X \bar{w}(x) \nabla_g^{(1)} S(y, x) \hat{g} \, dx
= \frac{1}{n} \sum_{i=1}^{n} \int_X \bar{w}(x) K_{h_y} (X_i - x) \int_0^y D_{g,0}(u, x) K_{h_y} \{Y_i - u\} \, du
= \frac{1}{n} \sum_{i=1}^{n} \int_{Y_i}^y \bar{w}(x) D_{g,0}(u, x) K_{h_y} (X_i - x) \, dx \, du \times \left[ 1 + O_P \left( h_y^m \right) \right]
= \frac{1}{n} \sum_{i=1}^{n} \bar{w}(X_i) \int_{Y_i}^y D_{g,0}(u, X_i) \, du \times \left[ 1 + O_P \left( h_y^m \right) + O_P \left( h_x^m \right) \right].
\]

By similar arguments,
\[
\int_X \bar{w}(x) \nabla_g^{(2)} S(y, x) \hat{g} \, dx
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \{ Y_i \leq y \} \int_X w(x) K_{h_x} (X_i - x) D_{g,y}(Y_i, x) \, dx + O_P \left( h_y^m \right)
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \{ Y_i \leq y \} w(X_i) D_{g,y}(Y_i, X_i) \left[ 1 + O_P \left( h_y^m \right) + O_P \left( h_x^m \right) \right],
\]

and
\[
\int_X \bar{w}(x) \nabla_g^{(3)} S(y, x) \hat{g} \, dx
= \frac{1}{n} \sum_{i=1}^{n} \int_{Y_i}^y \int_X \bar{w}(x) K_{h_x} \prime (X_{1,i} - x_1) K_{h_x} \left( X_{-1,i} - x_{-1} \right) D_{g,1}(u, x) \, dx \, du \times \left[ 1 + O_P \left( h_y^m \right) \right]
= -\frac{1}{n} \sum_{i=1}^{n} \int_{Y_i}^y K_{h_x} (X_{1,i} - x_1) K_{h_x} \left( X_{-1,i} - x_{-1} \right) \frac{\partial}{\partial x_1} \left[ \bar{w}(x) D_{g,1}(u, x) \right] \, dx \, du
\times \left[ 1 + O_P \left( h_y^m \right) \right]
= -\frac{1}{n} \sum_{i=1}^{n} \int_{Y_i}^y \frac{\partial}{\partial x_1} \left[ \bar{w}(X_i) D_{g,1}(u, X_i) \right] du \left[ 1 + O_P \left( h_y^m \right) + O_P \left( h_x^m \right) \right].
\]

Since \(\sqrt{n} \left[ h_x^m + h_y^m \right] = o(1)\), the claimed result now holds. \(\square\)
Lemma 3. Under Assumptions A1-A9: For any weighting function $\bar{w}$ with compact support,

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

where

$$
\phi_i^{\bar{w}} := \bar{\psi}_i^{\bar{w}} + \int_{X} \bar{w}(x) (S(Y_i, x) - E[S(Y, x)]) dx,
$$

$$
\bar{\psi}_i^{\bar{w}} := E \left[\psi_i^{\bar{w}}(Y_j) | X_i, Y_i\right],
$$

and $\psi_i^{\bar{w}}(y)$ is defined in Lemma 2.

Proof of Lemma 3. Applying Lemma 2,

$$
\int_{X} \bar{w}(x) \hat{E}[\hat{S}(Y, x)] - E[S(Y, x)] dx
\begin{eqnarray*}
\quad = & \int_{X} \bar{w}(x) \left\{ \hat{E}[\hat{S}(Y, x)] - E[S(Y, x)] \right\} dx + \int_{X} \bar{w}(x) \left\{ E[S(Y, x)] - E[S(Y, x)] \right\} dx \\
\quad = & \frac{1}{n} \sum_{i=1}^{n} \int_{X} \bar{w}(x) \left\{ \hat{S}(Y_i, x) - S(Y_i, x) \right\} dx \\
\quad & + \frac{1}{n} \sum_{i=1}^{n} \int_{X} \bar{w}(x) \left\{ S(Y_i, x) - E[S(Y, x)] \right\} dx \\
\quad = & \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{i,j}^{\bar{w}} + \frac{1}{n} \sum_{i=1}^{n} \int_{X} \bar{w}(x) \left\{ S(Y_i, x) - E[S(Y, x)] \right\} dx + o_P \left(1/\sqrt{n}\right),
\end{eqnarray*}
$$

where $\psi_{i,j}^{\bar{w}} = \psi_i^{\bar{w}}(Y_j)$. The first term is a $U$-statistic, and by appealing to standard results (see e.g. Newey and McFadden (1994), Lemma 8.4),

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

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

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

where $\phi_i^{\bar{w}}$ is defined in the lemma. \qed
Lemma 4. Under Assumptions A1-A9:

$$\sup_{(x,y) \in C} |\nabla g S(y,x)(\hat{g} - g)|^2 = o_P\left(\frac{1}{\sqrt{n}}\right), \quad \sup_{x,y} |\nabla f S(y,x)(\hat{f} - f)|^2 = o_P\left(\frac{1}{\sqrt{n}}\right),$$

for any compact set $C \subseteq \mathcal{X} \times \mathcal{Y}$.

Proof of Lemma 4. From the definition of $\nabla g S(y,x)(dg)$,

$$\sup_{(x,y) \in C} \left| \nabla g S(y,x)(dg) \right| \leq \sup_{(x,y) \in C} |D_{g,0}(y,x)| \sup_{(x,y) \in C} |dg(y,x)| + \sup_{(x,y) \in C} |D_{g,y}(y,x)| \sup_{(x,y) \in C} |dg_y(y,x)| + \sup_{(x,y) \in C} |D_{g,1}(y,x)| \sup_{(x,y) \in C} |dg_1(y,x)|,$$

where $\sup_{(x,y) \in C} |D_{g,a}(y,x)| < \infty$, $a = 0, y, 1$, given the continuity requirements imposed in Assumption A8 and the fact that $C$ is compact. Next, with $dg = \hat{g} - g$, it follows from the convergence rate results in eqs. (18) together with the bandwidth requirement in Assumption A9 that $\sup_{(x,y) \in C} |dg(y,x)| = o_P\left(\frac{1}{n^{1/4}}\right)$ and similarly for its partial derivatives w.r.t. $y, x$. This proves the first claim. The proof of the second claim follows along the same lines and so is left out. \qed
REFERENCES


