Econometrics and Statistics 17 (2021) 35–63

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Nonparametric Instrumental Variable Estimation of Binary Response Models with Continuous Endogenous Regressors

Samuele Centorrino [a,*], Jean-Pierre Florens [b]

[a] Economics Department, The State University of New York at Stony Brook, 100 Nicolls Rd, Social and Behavioral Sciences Building, 6th Floor, Stony Brook, NY 11794-4384, USA
[b] Toulouse School of Economics, University of Toulouse Capitole, France

ARTICLE INFO

Article history:
Received 12 September 2019
Revised 27 July 2020
Accepted 28 July 2020
Available online 21 August 2020

Keywords:
Nonparametric Methods
Endogeneity
Instrumental Variables
Binary Models
Tikhonov Regularization

ABSTRACT

An instrumental variable approach to the nonparametric estimation of binary response models with endogenous variables is presented. Identification is achieved via a reduced form model constructed from the decomposition of the unobserved dependent variable into the space of instruments. It is further assumed that disturbances in this model are independent of instruments, and their distribution is taken to be known. For estimation purposes, the fully nonparametric model is approximated by a sequence of locally weighted parametric models. Consistency and asymptotic normality of this estimator is proven, and a simulation study is performed to corroborate its small sample properties. Relevant policy parameters are constructed via a simulated nonparametric estimator of choice probabilities.

1. Introduction

Instrumental variables are a standard approach in statistics and econometrics to identify and estimate models with unobservable confounding variables. The underlying rationale for using instrumental variables is that, to uncover the true causal relation, we need a source of exogenous variation that should be informative about the phenomenon under study.

The recent and extensive literature on nonparametric instrumental regressions has grown in an attempt to couple the flexibility of nonparametric methods with the widespread use of instrumental variables in applied studies (see Newey and Powell, 2003; Hall and Horowitz, 2005; Carrasco et al., 2007; Darolles et al., 2011; Horowitz, 2011; Chen and Pouzo, 2012, among others).

In this literature, authors usually consider the following additively separable model

$$Y^* = \varphi(X) - U, \tag{1}$$

where variable X is endogenous (in particular X and U may be dependent), and the researcher is interested in the shape of the regression function φ. The model is completed by a vector of instruments, W, which satisfies $\mathbb{E}(U|W) = 0$. The regression function φ is estimated by solving a regularized version of a functional equation.

The objective of this work is to propose a nonparametric estimation of the function φ in the case in which Y^* is not directly observed. We assume instead to observe a binary transformation of it, i.e., $Y = \mathbb{1}(Y^* \geq 0)$.

* Corresponding author.
E-mail address: samuele.centorrino@stonybrook.edu (S. Centorrino).

https://doi.org/10.1016/j.ecosta.2020.07.003

We exploit the fact that variable Y^* can also be written as

$$Y^* = \mathbb{E}(Y^*|W) - \varepsilon. \tag{2}$$

Moreover, we impose a full independence condition between the error term of this reduced form model, ε, and the instrument W.

If the cumulative distribution function of the residual term ε is taken to be known, this assumption allows us to directly recover the conditional expectation of Y^* given W from the observation of the conditional probability of Y given W. We acknowledge that this is a strong assumption. However, as we discuss in the next section, other strategies that have been successfully exploited when X is exogenous to relax this assumption may yield uninterpretable restrictions in our case.

To obtain a simple estimate of the conditional expectation of the latent dependent variable, we use local likelihood inference (Tibshirani and Hastie, 1987; Fan et al., 1998; Gozalo and Linton, 2000; Frölich, 2006). In other words, we take polynomial approximations of $\mathbb{E}(Y^*|W = w)$ and obtain the coefficients of these approximations to maximize a likelihood function derived from the known cumulative distribution function of ε.

Finally, we obtain φ as the solution to the following functional equation:

$$\mathbb{E}(\varphi(X)|W) = \mathbb{E}(Y^*|W).$$

When the two sides of this equation are estimated using any nonparametric method, the solution is known to be an *ill-posed* inverse problem and requires regularization. Here, we follow the approach of Darolles et al. (2011) and explore the properties of a Tikhonov regularized solution in the case where the dependent variable is binary.

The main advantage of the class of models we consider is that they are single-equation models, insofar as they do not specify the structural equation that determines the endogenous regressor X. The strength of the relationship between the instrument and the endogenous variable is captured here by the so-called *completeness condition*, which is considered a necessary condition for identification in nonparametric instrumental variable models. However, some authors have shown that consistent nonparametric estimation of the function φ is still possible under *local* deviations from completeness (Freyberger, 2017). When completeness fails, and in the case of Tikhonov regularization, the estimator may still converge to the minimal norm solution (Florens et al., 2011; Babii and Florens, 2017b).

A relatively well-established body of literature has examined the estimation of binary regression models with continuous endogenous variables within a semiparametric framework (see Blundell and Powell, 2004; Rothe, 2009, among others). To correct for endogeneity, these authors advocate a control function approach. Identification is achieved by specifying a parametric form for the function φ and an index sufficiency restriction (see, e.g., Cosslett, 1983; Klein and Spady, 1993; Ahn et al., 2004). Finally, an exclusion restriction common to the control function literature allows one to obtain a nonparametric estimator of the error term's distribution. This approach and ours are not nested, as their identification strategies rely on markedly different assumptions. However, the relative advantage of instrumental variables over control functions is that the former can allow for simultaneity without explicitly modeling the dependence between the endogenous variable and the instrument (Imbens and Newey, 2009; Blundell et al., 2013).

Dong (2010) and Dong and Lewbel (2015) provide a semiparametric framework in which they relax some of the distributional assumptions on the error term ε. They do not require instruments. To achieve identification, they use the properties of a *special regressor*, as defined in Lewbel (1998). Notice, however, that such a special regressor should enter the conditional expectation function linearly. Although this identification strategy could be successfully exploited in a semiparametric specification of φ, it does not appear to yield any identifying power in a fully nonparametric context.

Moreover, all these papers restrict function φ to be parametric so that it is possible to relax the assumptions on the conditional distribution of the error term. Here, we adopt a different approach: relax the restrictions on the function φ, by making some additional assumptions about the conditional distribution of the error term, i.e., impose a full independence condition between the error term ε and the instrument. While both frameworks are subject to different forms of misspecification, there are several instances in applied work where our approach could complement existing ones. For instance, labor supply decisions are often taken based on unearned sources of income (for instance, the income provided by a spouse). These sources of income are not exogenous to the labor supply decision. Endogeneity could be caused by an unobservable and persistent taste component, correlated with both the available unearned sources of income and the labor supply decision. When using a general quasi-concave specification of the utility function, the labor supply decision is written as a nonlinear function of the unearned sources of income. Therefore, the index function may exhibit nonlinearities that an empirical researcher may want to capture.

We also establish the finite sample properties of the proposed estimator in an extensive Monte-Carlo simulation study.

Notation: We denote as $\|\varphi\|^2$ the $\mathbb{L}^2$-norm for a real-valued function φ. If a and b are scalars, we denote as $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. For two sequences a_n and b_n, we use the notation $a_n \approx b_n$ to signify that the ratio a_n/b_n is bounded away from zero and infinity. For a triplet of random variables, $\{X_1, X_2, X_3\}$, we use the notation $X_1 \perp\!\!\!\perp X_2$ to signify that X_1 is independent of X_2. Similarly, we write $X_1 \perp\!\!\!\perp X_2|X_3$, if X_1 is independent of X_2, conditionally on X_3.

2. The Model: Identification

Let (Y^*, X, W) be a random vector in $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$, with $q \geq p$, generated by a distribution F, which is absolutely continuous with respect to the Lebesgue measure, and such that:

$$Y^* = \varphi(X) - U, \quad \text{with} \quad \mathbb{E}(U|W) = 0, \tag{3}$$

where $\varphi(\cdot)$ is an unknown element of $\mathbb{L}^2_X$, the space of square-integrable functions of X with respect to the generating distribution F. We further assume that X and W take values in a compact set that, without loss of generality, is taken to be the hypercube $[0,1]^{p+q}$.

Model (3) is equivalent to:

$$\mathbb{E}(\varphi(X)|W) = r, \tag{4}$$

where $r = \mathbb{E}(Y^*|W)$, assuming Y^* is square-integrable.

When Y^* is observed, one can estimate r using any nonparametric technique and finally solve the inverse problem in (4) to obtain an estimator of φ (Newey and Powell, 2003; Hall and Horowitz, 2005; Darolles et al., 2011; Horowitz, 2011; Chen and Pouzo, 2012).

In this paper, we consider the estimation of φ in the case in which dependent variable Y^* is not observable. Instead, we assume that we have available a binary transformation of that variable, $Y = \mathbb{1}(Y^* \geq 0)$. The additional difficulty, in this case, is to obtain an estimation of r from the binary response variable Y^*.

To propose a strategy for the identification of the function φ when Y^* is not directly observable, we construct the model in an alternative way. For a given random variable Y^*, we can suppose that there exists a function $r(W)$ and a noise ε, with $\mathbb{E}[\varepsilon|W] = 0$, such that

$$Y^* = r(W) - \varepsilon.$$

This decomposition gives $r(W) = \mathbb{E}[Y^*|W]$, the conditional expectation of Y^* given W. Together with (4), this implies that

$$Y^* = \mathbb{E}[\varphi(X)|W] - \varepsilon.$$

The error term U in equation (3) can be defined as

$$U = \varepsilon + (\varphi(X) - \mathbb{E}[\varphi(X)|W]),$$

(see also Chen and Reiss, 2011; Florens and Simoni, 2012).

The main advantage of this construction is to separately consider the identification and estimation issues concerning (a) the conditional expectation of Y^*, given W, arising from the limited information we have on Y^*, and (b) the function of the endogenous variable φ in model (3). The function φ is defined by the usual integral equation of the first kind, and it is, therefore, a solution to an *ill-posed* inverse problem.

We start by discussing the identification of the latter. Notice that it remains unchanged also in this limited information case.

Define $T\varphi = \mathbb{E}[\varphi(X)|W]$, where $T : \mathbb{L}^2_X \to \mathbb{L}^2_W$ is the conditional expectation operator, and $T^* : \mathbb{L}^2_W \to \mathbb{L}^2_X$ denotes its adjoint. When r is known, the function φ is still uniquely determined by equation (4) if T is one to one, or, equivalently, if:

$$T\varphi \overset{a.s.}{=} 0 \quad \Rightarrow \quad \varphi \overset{a.s.}{=} 0, \tag{5}$$

(Newey and Powell, 2003; Darolles et al., 2011). We assume the following.

Assumption 2.1. The *completeness condition* stated in equation (5) holds.

We now turn the discussion to the identification in the binary model. Denote as $G_{\cdot|W=w}$ the conditional cumulative distribution function of the reduced form error ε with respect to the instrument W.

We, therefore, have

$$\begin{aligned} p(w) &= \mathbb{P}(Y = 1|W = w) = \mathbb{P}(Y^* \geq 0|W = w) = \mathbb{P}(r(W) - \varepsilon \geq 0|W = w) \\ &= G_{\varepsilon|W=w}(r(W)|W = w). \end{aligned} \tag{6}$$

This conditional probability is identified directly from the joint distribution of the observed random variables. However, in binary response models, it is not feasible to jointly nonparametrically identify the conditional expectation function, r, and the conditional distribution of the error term, ε, without additional assumptions (Manski, 1988).

A viable approach would be to replace the unknown conditional expectation function r with some finite parametric approximation. One could then estimate the parameters vector, and $G_{\varepsilon|W=w}$ nonparametrically, using either an index sufficiency condition or restrictions based on conditional quantiles of ε, given W (see Manski, 1985; Horowitz, 1992; Klein and Spady, 1993; Ichimura, 1993, among others). An alternative approach is to suppose that the error term ε is independent of the instrumental variable W.

The former approach has the advantage of not imposing any restriction on the distribution of the error term. It, therefore, avoids misspecification of the conditional distribution in the reduced form model. We are, however, interested in the

nonparametric estimation of the regression function φ, and a finite-dimensional parametric approximation may lead to erroneous conclusions regarding the shape of φ.

Moreover, although the maximum score and the smoothed maximum score estimators of Manski (1985) and Horowitz (1992) allow for any form of potential heteroskedasticity, the index sufficiency restriction requires the variance component to be a function of the index function.

Here, we pursue the identification and estimation of the structural function φ by restricting the dependence between the instrument and the reduced form residual, ε. We start with the following assumption.

Assumption 2.2. $\varepsilon \perp\!\!\!\perp W$ and $\mathbb{E}(\varepsilon) = 0$.

This assumption places a direct independence restriction on the reduced form model.

Our construction of the model, together with Assumption 2.2, induces some restrictions on the structural error term U, particularly on the form of heteroskedasticity that is allowed. At this level of generality, we are not able to derive more primitive conditions on the structural model that would translate into our independence assumption.

Assumption 2.2 trivially holds when U can be written as the sum of a random component, independent of W, and $\mathbb{E}[\varphi(X)|W] - \varphi(X)$. Moreover, if we directly assume that

$$(U, \varphi(X) - \mathbb{E}[\varphi(X)|W]) \perp\!\!\!\perp W,$$

Assumption 2.2 is satisfied. This restriction is easily interpretable when the function φ is linear, and X can be written as an additively separable function of W and $\varphi(X) - \mathbb{E}[\varphi(X)|W]$. Here below, we detail such an example. In the linear case, we also show that our assumption is amenable to the exclusion restriction used in the control function approach literature (see Newey et al., 1999, among others).

Example 1 (Linear Model)**.** Suppose data are generated by the following triangular model

$$\begin{aligned} Y^* &= \phi_0 + X\phi_1 - U, \\ X &= m(W) - \eta, \end{aligned}$$

where φ is linear in X, and $(U, \eta) \perp\!\!\!\perp W$, with $\mathbb{E}(U) = \mathbb{E}(\eta) = 0$. We can then define

$$\varepsilon = U - X\phi_1 + m(W)\phi_1 = U + \eta\phi_1 \perp\!\!\!\perp W.$$

We could take, for instance,

$$\begin{pmatrix} U \\ \eta \end{pmatrix} | W = w \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_\eta \tau \\ \sigma_\eta \tau & \sigma_\eta^2 \end{pmatrix} \right),$$

where τ is a constant in $(-1, 1)$. If we further take the function $m(W)$ to be linear, the model becomes the same as that studied in Rivers and Vuong (1988). However, notice that our framework allows for any (known) distribution of ε. □

By Assumption 2.2, from equation (6), we can write that

$$p(w) = G_\varepsilon(r(w)), \tag{7}$$

Finally, we impose the following condition.

Assumption 2.3. G_ε is a known, monotone increasing, continuous, and strictly increasing function of $r(W)$.

We acknowledge that Assumption 2.3 is somewhat restrictive. We do not know whether our identification result holds more generally under milder conditions on the CDF of ε. Other approaches in latent variable models restrict the functional form of r to allow for nonparametric identification of the CDF (see Matzkin, 1992, Lewbel, 2000, Lewbel and Linton, 2007, and Jacho-Chávez et al., 2010, among others). It is not evident to us that similar identification strategies could be meaningfully employed in our case, without further arbitrary constraints on the index function φ. For instance, Lewbel and Linton (2007) restrict r to belong to the class of homogeneous functions of degree one. If G_ε is continuous and strictly monotone in r, they show that both r and G_ε are nonparametrically identified, up to location and scale. However, taking r to be homogeneous of degree one in our framework does not lead to any interpretable restrictions on φ. A viable approach would be to impose this restriction directly on the function φ. The latter does not seem to yield any identifying power in our case beyond the simple linear framework, i.e. $\varphi(X) = X\phi$, with ϕ being a p dimensional parameter. Similarly, Jacho-Chávez et al. (2010) restrict the conditional expectation to be additive in at least one component, say W_1, and assume that $\varepsilon \perp\!\!\!\perp W_1 | W_{-1}$, which is a special regressor restriction (Lewbel, 1998). There is no particular reason to assume that $r(W)$ is additive in at least one component in our framework.

Assumption 2.3 could be relaxed at the cost of losing point identification. We defer such a discussion for future research. Because of Assumption 2.3,

$$r(w) = G_{\varepsilon}^{-1}(p(w)). \tag{8}$$

Finally, φ is identified from the completeness condition in (5).

The main identification result is summarized in the following proposition, whose proof is straightforward and therefore omitted.

Proposition 2.1. *Under Assumption 2.2 and 2.3, and the completeness condition in Assumption 2.1, the regression function, φ, is identified.*

Notice that the distribution of ε is usually known up to a scale factor. A further scale normalization should be imposed by taking either $\|\varphi\| = 1$ or $Var(\varepsilon) = 1$.

3. The Model: Estimation

We presume that one observes an i.i.d. random sample from the distribution of (Y, X, W), which we denote (Y_i, X_i, W_i), for $i = 1, \dots, n$. All the high-level assumptions are standard in the nonparametric IV literature, and we refer the interested reader to Hall and Horowitz (2005), Darolles et al. (2011), and Horowitz (2011) for an extensive review of them.

We consider univariate generalized kernel functions K_{h_n} of order $\kappa \geq 2$, where h_n is a bandwidth parameter that goes to 0 as $n \to \infty$. We further take the class of joint probability density functions of the pair (X, W) to be in $\mathbb{C}^d$. We denote $\rho = \kappa \wedge d$. The bandwidth parameter is taken to be the same across all components of W and X, although this could be easily relaxed.

The main issue arising from the nonparametric approach concerns the *ill-posedness* of the inverse problem defined by the functional equation

$$T\varphi = r. \tag{9}$$

This problem is now well understood and has been extensively analyzed in the econometrics literature (Carrasco et al., 2007). To address the inverse problem, we must apply a regularization method. In particular, we decide to use the so-called *Tikhonov* regularization approach, advocated in Hall and Horowitz (2005), and Darolles et al. (2011). However, any other regularization method could have been equivalently applied (see, e.g., Horowitz, 2011; Florens et al., 2018; Johannes et al., 2013; Chen and Pouzo, 2012).

The regularized solution of the inverse problem in (9) is taken to minimize the following penalized criterion:

$$\varphi^{\alpha} = \arg\min_{\varphi \in \mathbb{L}^2_X} \|T\varphi - r\|^2 + \alpha\|\varphi\|^2, \tag{10}$$

where α is the regularization parameter which should be chosen appropriately (see Fève and Florens, 2010; Centorrino, 2016; Centorrino et al., 2017, on the choice of the tuning parameter for Tikhonov regularization).

The minimization problem in (10), for a given value of the parameter α, yields the following solution:

$$\varphi^{\alpha} = (\alpha I + T^*T)^{-1}T^*r, \tag{11}$$

where I denotes the identity operator. Hence, the parameter α serves to bound away from 0 the eigenvalues of the operator T^*T. This generates a *regularization bias* in finite samples that should disappear as the sample size increases. To control the convergence to zero of this regularization bias, we make the following additional assumption.

Assumption 3.1. The unknown function φ function has regularity $\beta > 0$. That is,

$$\varphi \in \mathcal{R}\left[(T^*T)^{\frac{\beta}{2}}\right],$$

where $\mathcal{R}$ denotes the range of the operator.

Assumption 3.1 is the so-called *source condition* (Engl et al., 2000). This high-level assumption relates the properties of the function φ to those of the conditional expectation operator T. We refer the interested reader to Carrasco et al. (2007), Chen and Reiss (2011), Darolles et al. (2011), and Centorrino (2016), for a more detailed discussion. Notice that our source condition is *strong*, in the sense that we allow only for mildly ill-posed inverse problems (or for severely ill-posed problems only when the φ is sufficiently smooth). However, Assumption 3.1 can easily be extended to encompass also a *weak* source condition, which would also allow us to consider the severely ill-posed case in its full generality. As we do not intend to make a direct contribution to the rich literature on ill-posed inverse problems in econometrics, we have decided to maintain this strong source condition.

We state the following result regarding the $\mathbb{L}^2$ convergence of the regularization bias.

Proposition 3.1. *When Assumption 3.1 is satisfied, we obtain:*

$$\|\varphi^{\alpha} - \varphi\|^2 = O\left(\alpha^{\beta \wedge 2}\right),$$

where φ^α is defined in (11).

A proof of this proposition is given in Carrasco et al. (2007), and we refer the interested reader to their work. The minimum between β and 2 arises because of the so-called *qualification* of the Tikhonov regularization approach, which is equal to 2. Loosely speaking, Tikhonov regularization cannot take advantage of additional smoothness of the function φ, for a given convergence rate of the eigenvalues of T^*T to zero. This limitation can be overcome with an iterated Tikhonov approach or by penalization in Hilbert scales (Florens et al., 2011; Gagliardini and Scaillet, 2012; Centorrino, 2016).

To estimate the regularized solution in (11), we must first construct an estimator of r, the conditional mean of Y^* given W, the conditional expectation operator T, and its adjoint T^* (Centorrino et al., 2017).

Denote as $f_{X,W}$, f_X and f_W, the joint and the marginal pdfs of X and W, respectively, and as $K_{W,h}$, and $K_{X,h}$ the multivariate kernel functions of order κ and dimension q and p, respectively. For any pair of functions, φ and ψ, the estimators of T and T^* are defined as follows:

$$\left(\hat{T}\varphi\right)(w) = \int \varphi(x) \frac{\hat{f}_{X,W}(x,w)}{\hat{f}_W(w)} dx, \tag{12}$$

$$\left(\hat{T}^*\psi\right)(x) = \int \psi(w) \frac{\hat{f}_{X,W}(x,w)}{\hat{f}_X(x)} dw, \tag{13}$$

where $\hat{f}_{X,W}$, $\hat{f}_X$, and $\hat{f}_W$ are the usual nonparametric kernel estimators of the joint and marginal pdfs.

The most important part of our procedure is, however, to obtain an estimator of the conditional expectation function, $r(w)$.

We follow the local likelihood approach developed in Tibshirani and Hastie (1987). For a given G_ε, we take a local polynomial approximation of the conditional expectation function $r(W_i)$ around a point $w \in [0,1]^q$. For instance, depending on the choice of G_ε, we can estimate a local Probit or Logit model. This local approximation may also be more robust to certain types of misspecification (Gozalo and Linton, 2000; Lewbel, 2007).

To this end, we need to assume that the function r is sufficiently smooth, in the sense that it possesses at least ρ continuous derivatives. We can then approximate $r(W_i)$ around w as

$$r(W_i) \approx \sum_{0 \leq |\mathbf{j}| \leq \rho - 1} \frac{1}{\mathbf{j}!} (D^{(\mathbf{j})} r)(w) (W_i - w)^{\mathbf{j}}, \tag{14}$$

where

$$\mathbf{j} = (j_1, \ldots, j_q), \qquad \mathbf{j}! = j_1! \times \cdots \times j_q!, \qquad |\mathbf{j}| = \sum_{k=1}^{q} j_k,$$

$$(W_i - w)^{\mathbf{j}} = (W_{1i} - w_1)^{j_1} \times \cdots \times (W_{qi} - w_q)^{j_q},$$

$$\sum_{0 \leq |\mathbf{j}| \leq \rho - 1} = \sum_{k=0}^{\rho-1} \sum_{\substack{j_1=0 \\ j_1 + \cdots + j_q = |j|}}^{k} \cdots \sum_{j_q=0}^{k},$$

and

$$(D^{(\mathbf{j})} r)(w) = \frac{\partial^{\mathbf{j}} r(w)}{\partial w_1^{j_1} \ldots \partial w_q^{j_q}}.$$

Finally, we define

$$\gamma_{0\mathbf{j}}(w) = \frac{(D^{(\mathbf{j})} r)(w)}{\mathbf{j}!}.$$

Further, define $r(w)$ as the solution to the following conditional moment restriction

$$\mathbb{E}[\ell_1(r(W), Y) | W = w] = 0,$$

where ℓ_1 is the first derivative of the log-likelihood function,

$$\ell(r(W), Y) = Y \log(G_\varepsilon(r(W))) + (1 - Y) \log(1 - G_\varepsilon(r(W))),$$

with respect to its first argument. The function G_ε can take known forms. For instance, it could be either the normal or the logistic cdf (Fan et al., 1998). Denote by $\gamma(w)$ the vector of coefficients arranged in a lexicographical order (Masry, 1996). The sample (leave-one-out) counterpart of the local log-likelihood function can be written as

$$\mathcal{L}_{n,\rho}^{(-i)}(\gamma(W_i)) = \frac{1}{(n-1)h^q} \sum_{i'=1, i' \neq i}^{n} \ell\left(\sum_{0 \leq |\mathbf{j}| \leq \rho - 1} \gamma_{\mathbf{j}}(w) (W_{i'} - W_i)^{\mathbf{j}}, Y_i \right) K_h(W_{i'} - W_i), \tag{15}$$

using the polynomial approximation in equation (14). All the appropriate regularity conditions that the likelihood function should satisfy are given in the Appendix.

Our leave-one-out estimator of γ, denoted $\hat{\gamma}$, satisfies:

$$\hat{\gamma}_{(-i)}(w) = \arg\min_{\gamma\in\mathcal{G}} \mathcal{L}_{n,\rho}^{(-i)}(\gamma(w)), \quad \forall w \in [0,1]^q,$$

where $\mathcal{G}$ is an appropriate parameter space. Notice that we can consistently estimate the function r and all its partial derivatives up to the order $\rho - 1$. Finally, we have

$$\hat{r}_{(-i)}(w) = e_1' \hat{\gamma}_{(-i)}(w),$$

where $e_1 = (1, 0, \dots)'$. Thus, $\hat{r}_{(-i)}(w)$ is the first component of the vector $\hat{\gamma}_{(-i)}(w)$, for every $w \in [0,1]^q$. The leave-one-out estimation is used for theoretical convenience, as it is shown later.

Remark 1. Our restriction on the reduced form residual can be translated into the following system of conditional moments

$$\mathbb{E}[\mathbb{1}(Y=1)|W] = p(W),$$
$$\mathbb{E}[\varphi(X)|W] = G_\varepsilon^{-1}(p(W)).$$

An alternative approach to our is to use the methodology developed in Chen and Pouzo (2012,2015) and obtain an estimator of φ through the method of sieves.

3.1. Consistency

The $\mathbb{L}^2$ convergence rate of $\hat{r}_{(-i)}(w)$ and $\left(\hat{T}^*\hat{r}_{(-i)}\right)(x)$ are given in the following lemma:

Lemma 3.2. *Under the regularity conditions in Assumption A.1-A.6, there exists, for all $w \in [0,1]^q$, a sequence of solutions $\{\hat{\gamma}\}$ to the likelihood equations*

$$\frac{\partial}{\partial \gamma_{\mathbf{j}}} \mathcal{L}_{n,\rho}^{(-i)}\left(\hat{\gamma}\right) = 0, \quad \forall \mathbf{j} = 0, 1, \dots, \rho - 1.$$

Furthermore, the nonparametric estimator of r, $\hat{r}_{(-i)} = e_1'\hat{\gamma}_{(-i)}$, is such that:

$$\|\hat{r}_{(-i)} - r\|^2 = O_P\left(\frac{1}{nh^q} + h^{2\tilde{\rho}}\right),$$
$$\|\hat{T}^*\left(\hat{r}_{(-i)} - r\right)\|^2 = O_P\left(\frac{1}{n} + h^{2\tilde{\rho}}\right),$$

where q is the dimension of the instrumental vector used to identify r, $\tilde{\rho} = [(\rho-1)/2] + 1$ depends on the order of the polynomial ρ, and $[\cdot]$ denotes the integer part of a positive real number. □

This order of convergence of the local log-likelihood estimator is standard in the literature (Fan et al., 1995; 1997; Claeskens and Van Keilegom, 2003), and holds under mild continuity conditions on the log-likelihood function $\ell(\cdot,\cdot)$. The term in $2([(\rho-1)/2]+1)$ arises because it is necessary to distinguish cases in which the polynomial is of odd or even degree (Gu et al., 2015). In the following, to simplify notations, we take $\tilde{\rho} = \rho$. The second part of the Lemma implies that we can get a $\sqrt{n}$ convergence by applying an additional smoothing step through $\hat{T}^*$ as shown in Darolles et al. (2011). The result of this Lemma is proven in Appendix.

Finally, we define the estimator of φ for our binary nonparametric regression function as

$$\hat{\varphi}^{\alpha_n} = \left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\hat{r}_{(-i)}, \tag{16}$$

where α_n is a sequence of regularization parameters that goes to 0 as $n \to \infty$.

The following result provides the convergence rate of the nonparametric binary instrumental variable estimator of the regression function φ.

Theorem 3.3. *Let Assumption 2.3, 3.1, and A.1-A.7, the result, and the conditions of Proposition 3.1 and Lemma 3.2 hold. We have that*

$$\|\hat{\varphi}^{\alpha_n} - \varphi\|^2 = O_P\left[\left(\frac{1}{nh_n^q\alpha_n} + \frac{h_n^{2\rho}}{\alpha_n}\right) \vee \left(\frac{1}{n\alpha_n^2} + \frac{h_n^{2\rho}}{\alpha_n^2}\right) + \left(\frac{1}{nh_n^{p+q}} + h_n^{2\rho}\right)\alpha_n^{(\beta-1)\wedge 0} + \alpha_n^{\beta\wedge 2}\right]. \tag{17}$$

□

This bound is equivalent to the one in Darolles et al. (2011). The only minor difference is that Darolles et al. (2011) characterize the rate of convergence of the variance component of $\hat{\varphi}^{\alpha_n}$ by directly considering the squared norm of $\hat{T}^*\left(\hat{r} - r\right)$. Here, we take the worst-case upper bound, by also considering directly the bound as a function of the squared norm of $\hat{r} - r$.

The following corollary specifies the conditions on the bandwidth and the regularization parameter such that the upper bound in Theorem 3.3 goes to zero.

Corollary 3.4. *Let the result of Theorem 3.3 hold. As $\alpha_n \to 0$ and $h_n \to 0$ with $h_n^{\rho}\alpha_n^{-1} \to 0$ as $n \to \infty$, if*

(i) $\beta \geq 1$, *and* $n\alpha_n\left(h_n^q \wedge \alpha_n\right) \to \infty$ *with*

$$\frac{\alpha_n\left(h_n^q \wedge \alpha_n\right)}{h_n^{p+q}} = o(1); \text{ or}$$

(ii) $n\alpha_n\left(h_n^q \wedge \alpha_n \wedge \alpha_n^{-\beta} h_n^{p+q}\right) \to \infty$,

then,

$$\|\hat{\varphi}^{\alpha_n} - \varphi\|^2 \xrightarrow{p} 0.$$

□

The proof of this Corollary is straightforward and provided in the Appendix. The conditions under which the estimator is consistent depend on the joint choice of the smoothing and regularization parameter in a nontrivial way, and we refer to Centorrino (2016) for a detailed discussion. We simply note that when $\beta > 1$, $4\rho > (\beta \wedge 2)(p+q)$ and the bandwidth is chosen to be proportional to $n^{-1/(2\rho+p+q)}$, then:

$$\alpha_n \propto n^{-\frac{1}{\beta\wedge 2+1}\frac{2\rho}{2\rho+p+q}},$$

satisfies the conditions of our Corollary 3.4.

3.2. Asymptotic normality

In this section, we consider conditions under which the Tikhonov regularized estimator $\hat{\varphi}^{\alpha_n}$ is asymptotically normal when $\alpha_n \to 0$ with n. Carrasco and Florens (2011) and Horowitz (2007) look at the pointwise asymptotic normality of this estimator. Here, we follow Carrasco et al. (2013) and we consider the asymptotic normality of inner products of the type

$$\langle \hat{\varphi}^{\alpha_n} - \varphi^{\alpha_n}, \delta \rangle,$$

where δ is a square-integrable function, $\varphi^{\alpha_n} = (\alpha_n I + T^*T)^{-1}T^*T\varphi$ and α_n vanishes as the sample size increases. We start by making the following Assumption.

Assumption 3.2. There exists $\nu \geq 0$ such that $\delta = (T^*T)^{\frac{\nu}{2}}v$, for a v with $\|v\| < \infty$.

This Assumption is similar to the one in Carrasco et al. (2013), and it imposes some regularity conditions on the function δ. Let $\ell_2(\cdot, \cdot)$ be the second derivative of the log-likelihood function with respect to its first argument. A preliminary result is given in the following Proposition.

Proposition 3.5. *Under Assumption A.1-A.7, and provided that $h_n = o_P(n^{-1/2\rho})$, we have*

$$\sqrt{n}\left(\hat{T}^*\left(\hat{r}_{(-i)} - r\right)\right) \Rightarrow N(0, \Omega),$$

with $\Omega = T^*I(r)^{-1}T$, where $I(r(w)) = -\mathbb{E}[\ell_2(r(w), Y)|W = w]$ is the local Fisher information.

This proposition shows that we can apply a functional central limit theorem to the term $\hat{T}^*\left(\hat{r}_{(-i)} - r\right)$. Such a result allows us to extend the asymptotic normality of inner products presented in Carrasco et al. (2013) to our case.

Theorem 3.6. *Let the conditions of Theorem 3.3, Assumption 3.2, and Proposition 3.5 hold. Further let $M = (\alpha_n I + T^*T)^{-1}$, such that*

$$\frac{\|\hat{T}^*\hat{T} - T^*T\|\alpha_n^{\frac{\nu-1}{2}\wedge 0}}{\|\Omega^{1/2}M\delta\|} \to 0,$$
$$\frac{\sqrt{n}\|\hat{T}^*\hat{T} - T^*T\|\alpha_n^{\frac{\beta-1}{2}\wedge 0}}{\|\Omega^{1/2}M\delta\|} \to 0.$$

Then we have,

$$\sqrt{n}\frac{\langle \hat{\varphi}^{\alpha_n} - \varphi^{\alpha_n}, \delta\rangle}{\|\Omega^{1/2}M\delta\|} \xrightarrow{d} N(0, 1),$$

where $\xrightarrow{d}$ denotes convergence in distribution.

This theorem provides asymptotic normality for linear functionals of the nonparametric instrumental variable estimator given in this paper. For our result to hold, we need to have some undersmoothing in the estimation of conditional expectations, i.e., h_n needs to go to zero fast enough. Despite the result being expressed as $\sqrt{n}$ convergence, the norm of the variance $\|\Omega^{1/2}M\delta\|$ may not necessarily be bounded away from infinity when $n\to\infty$, and thus we often have a rate that is slower than $\sqrt{n}$. The regularity properties of the functions δ and φ are another important element in determining the speed of convergence of the inner product. These regularities are expressed, as above, in terms of a strong source condition. Higher regularity of these functions allows us to ignore terms that are due to the estimation of the conditional expectation operators. The saturation effect of Tikhonov regularization plays an important role, as it effectively bounds above the level of regularity of the function that can be exploited.

This point is going to be extremely important when we discuss the estimation of the Average Structural Function in Section 3.4. The latter is a nonlinear functional of the function φ. However, upon a linearization argument, we reach a similar conclusion.

3.3. Implementation

Our estimator of regression function φ is obtained as follows:

(i) Using local likelihood methods, we estimate conditional mean function $r(w)$ directly. For the local model, we use a Logit specification (Frölich, 2006). We also use the bandwidth parameter associated with component W of the instrument to obtain an estimator of T.
(ii) We estimate the adjoint operator T^*, as the conditional expectation of the function $r(W)$, given X.
(iii) We find the Tikhonov regularized solution $\hat{\varphi}^{\alpha_n}$.

Step (i)

Recall that $p(w)=\mathbb{P}(Y=1|W=w)$, the conditional probability in our binary nonparametric regression model.

Frölich (2006) extensively compares several semi/nonparametric methods to the local linear Logit regression in the class of binary models. He concludes that a local linear Logit regression is preferable to other specifications, as it exhibits greater precision in estimating both conditional means and marginal effects, especially in models with many predictors. He also extensively discusses some of the implementation and computational issues that may arise when using local likelihood methods and the data-driven selection of the smoothing parameter in such a model. We refer interested readers to his paper. For an earlier comparison of nonparametric estimators for binary models, see also Signorini and Jones (2004).

The choice of the Logit specification is made for convenience, as the logistic cdf is analytically tractable and strictly concave. However, it can be replaced by any other cdf, such as the normal, to estimate a local Probit model (Gozalo and Linton, 2000). Smoothing parameters are selected using least-squares cross-validation. Least-square cross-validation can be computationally very expensive in this class of models. At each sample point, we need to implement a numerical search for the coefficients that minimize the likelihood function and iterate to optimize the cross-validation function over a grid of plausible bandwidth values. We, therefore, use a Fisher scoring method when computing the value of the bandwidth parameter.

Finally, we use bandwidth parameter h_W to obtain an estimator of the conditional expectation operator T as:

$$\hat{T}=\left[\frac{K_{h_W}\left(W_i-W_j,W_j\right)}{\sum_{i=1}^{n}K_{h_W}\left(W_i-W_j,W_j\right)}\right]_{i,j=1}^{n},$$

that is, by using the $n\times n$ matrix of kernel weights. This estimation is consistent as T is a Hilbert-Schmidt operator and can, therefore, be approximated by a sequence of finite-dimensional operators (Carrasco et al., 2007). We further point out that this estimator of $\hat{T}$, and the one defined in equation (12) are equivalent up to an approximation error that goes to zero with the bandwidth. For $\varphi\in\mathbb{L}_X^2$, we have

$$\left(\hat{T}\varphi\right)(w)=\int\varphi(x)\frac{\frac{1}{nh_n^{p+q}}\sum_{i=1}^{n}K_{W,h}(w-W_i,w)K_{X,h}(x-X_i,x)}{\frac{1}{nh_n^{q}}\sum_{i=1}^{n}K_{W,h}(w-W_i,w)}dx.$$

By rearranging terms, we obtain

$$\left(\hat{T}\varphi\right)(w)=\frac{\frac{1}{nh_n^{q}}\sum_{i=1}^{n}K_{W,h}(w-W_i,w)\int\frac{1}{h_n^{p}}K_{X,h}(x-X_i,x)\varphi(x)dx}{\frac{1}{nh_n^{q}}\sum_{i=1}^{n}K_{W,h}(w-W_i,w)}.$$

The change of variable $x=X_i+uh_n$ gives $\int h_n^{-p}K_{X,h}(x-X_i,x)\varphi(x)dx=\varphi(X_i)+O_P(h_n)$. This entails the desired nonparametric estimation of the conditional expectation operator. Similar reasoning applies to the estimator of $\hat{T}^*$ discussed below.

Step (ii)

Adjoint operator T^* defines the conditional expectation of all square-integrable functions of W given X. Therefore, a natural nonparametric estimator is

$$(\hat{T}^*\hat{r})(x) = \frac{\sum_{i=1}^{n} \hat{r}_i K_{h_X}(X_i - x, x)}{\sum_{i=1}^{n} K_{h_X}(X_i - x, x)},$$

with bandwidth parameter, h_X.

Step (iii)

We finally obtain the nonparametric instrumental regression function by solving (9), using the Tikhonov regularization method (see equation 16).

To compute a data-driven value of the regularization parameter, we adopt the cross-validation approach developed in Centorrino (2016). It consists of finding the parameter α_n as the minimizer of the following criterion:

$$CV_n(\alpha) = \sum_{i=1}^{n} \left((\hat{T}\hat{\varphi}^{\alpha}_{(-i)})(W_i) - \hat{r}(W_i) \right)^2 \tag{18}$$

where $\hat{\varphi}^{\alpha}_{(-i)}$ is the estimator of φ in which the i^{th} observation has been removed. This function corresponds to the minimization of the residuals norm from the integral equation (9).

Using this selection criterion, we obtain our Tikhonov regularized nonparametric estimator of the regression function as described in (16).

3.4. Policy Parameters

In this paper, we have so far focused on the estimation of the *index* function φ. However, the estimation of the function alone does not allow us to consider potential counterfactuals in our specification. r and G_ε are not structural parameters and thus are not directly useful for policy evaluation.

One central policy parameter is the Average Structural Function (ASF), as defined in Blundell and Powell (2004). This is the choice probability evaluated at some fixed value of X. In our case, the choice probability can be defined by the value of the distribution of structural error U, evaluated at $\varphi(x) = \phi$.

Let $\eta = \varphi(X) - r(W)$. This random variable can be consistently estimated using our estimators of the functions φ and r. We then have

$$F_U(\phi) = P(U \leq \phi) = P(\varepsilon + \eta \leq \phi) = \int_{S_\eta} \int_{-\infty}^{\phi - \eta} f_{\varepsilon\eta}(\varepsilon, \eta) d\varepsilon d\eta.$$

Hence, knowledge of the conditional density of ε, given η, is sufficient to identify the marginal distribution of the structural error term U. This density cannot be evaluated directly, as the dependent variable is latent in this framework. However, as the marginal cdf of ε is taken to be known, it is possible to draw random realizations from it directly. Thus, one way to obtain the ASF is through simulation directly from G_ε.

Denote as ε^* a random realization from G_ε. Knowledge of ε^* would allow one to compute

$$U^* = \varepsilon^* + \eta,$$

and thus retrieve

$$F^*_U(\phi) = P(U^* \leq \phi).$$

All the random variables above can be replaced by consistent estimators to construct a simulated nonparametric estimator of the ASF. Notice, however, that when r and φ are replaced by consistent estimators, we cannot directly simulate ε from its known distribution because the estimator of φ depends on our sample, and therefore, on a specific realization of ε. Thus, we proceed as follows

(i) We take the (leave-one-out) estimator of r and let $\{\xi^*_i, i = 1, 2, \ldots, n\}$ be a random sample from the uniform distribution on [0,1]. The ξ^*_i-quantile of the distribution of ε, which we denote as Q^*_i, satisfies,

$$G_\varepsilon\left(Q^*_i\right) = \xi^*_i, \forall i = 1, \ldots, n. \tag{19}$$

Under Assumption 2.3, the solution to this problem is unique for every $\xi^*_i \in [0, 1]$.

(ii) Let $\tilde{Y}_i = \hat{r}_{(-i)}(W_i) - Q_i^*$, and $\hat{r}^*(W_i)$ be the estimator of the conditional expectation of $\tilde{Y}_i$ given W_i. We obtain

$$\hat{\varphi}^{*,\alpha_n} = \left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\hat{r}^*.$$

(iii) Finally, we compute

$$\hat{U}_i^* = Q_i^* + \hat{\varphi}^{*,\alpha_n}_{(-i)}(X_i) - \hat{r}_{(-i)}(W_i),$$

where $\hat{\varphi}^{*,\alpha_n}_{(-i)}(X_i)$ is the leave-one-out version of our estimator of φ in step (*ii*).

This simulation procedure is consistent, as it is based on the estimation of φ that would be obtained if the true ε_i had been equal to Q_i^*, for all $i = 1, \dots, n$. This procedure can be repeated $S \geq 1$ times to smooth the estimator by integration. However, its consistency does not depend on S. As simulated samples are not independent, it is not possible to reduce the variance by letting S diverge to infinity. Hence, to simplify notations and the derivation of the asymptotic properties, we take $S = 1$. The estimator of the ASF is thus given by

$$\widehat{ASF}(\phi) = \hat{F}^*_U(\phi) = \frac{1}{n}\sum_{i=1}^{n} \bar{K}_{b_F}\left(\hat{U}_i^* - \phi\right),$$

where $\bar{K}_{b_F}$ is the integral of the kernel function $K_{b_F}\left(\hat{U}_i^* - \phi, 1\right)$, with respect to its first argument, as defined in Assumption A.2, and b_F is a bandwidth parameter. The ASF is a nonlinear functional of the nonparametric IV estimator (Santos, 2011; Severini and Tripathi, 2012; Chen and Christensen, 2018). It is not apparent that we can achieve $\sqrt{n}-$consistency for this functional. The latter result usually depends on the smoothness of the function φ, i.e., the source condition in Assumption 3.1 and the smoothness of the true functional to be estimated (see also Assumption 3.2). In the following, we use a similar source condition to summarize the smoothness properties of the conditional density of U given the endogenous regressor X. See Babii and Florens (2017a), for a similar assumption.

Assumption 3.3. There exists a constant $\nu > 1 \vee \frac{q}{4\rho}$ such that,

$$f_{U|X}(\phi|x) = (T^*T)^{\frac{\nu}{2}} v(\phi, \cdot),$$

with $\sup_\phi |v(\phi, \cdot)| < \infty$.

However, there is an additional issue which is related to the qualification of the Tikhonov regularization scheme. We show below that, without sufficient qualification, it is not possible to obtain a $\sqrt{n}-$rate, independently of the smoothness of the underlying components involved. To increase the qualification, we use an iterated Tikhonov approach (see, e.g., Carrasco et al., 2007; Fève and Florens, 2010), with a given number of iterations, m.

As the support of the residuals is potentially unbounded, we look at the consistency of this estimator over a $\mathbb{L}^2-$norm, which is weighted by a positive function $\omega(\phi)$. Several choices of this function are possible. For instance, we could take $\omega(\phi) = \mathbb{1}(|\phi| \leq C)$, for a finite positive constant C, which would correspond to restricting our estimator over a compact set.

The main result is given in the following Theorem.

Theorem 3.7. *Let Assumption 2.2, 2.3, 3.1, 3.3, and A.1-A.8 hold, with* $b_F \asymp n^{-1/2\rho}$, *and* $2m > \beta \vee \nu + 1$. *Then*

$$\|\widehat{ASF}(\phi) - ASF(\phi)\|^2_\omega = o_P(1).$$

We do not give the precise upper bound in the statement of the Theorem simply for ease of presentation. The estimator of the ASF is not a sum of independent terms, and therefore we need to account for the various covariances between these terms. The lower bound on m, the number of iterations for the Tikhonov scheme, is not essential for the result of the Theorem and could be dropped without affecting it. It becomes, however, necessary, as we aim at achieving $\sqrt{n}-$convergence of this estimator.

From standard results in nonparametric estimation, and under our condition on the bandwidth parameter b_F, the convergence rate would be of order n^{-1}. In Theorem 3.7, the bound on the estimator of the ASF depends on the properties of the preliminary estimator of φ that must be plugged in. One would expect to undersmooth the first step estimator insofar as this dampens the impact of the bias on the second step estimator. The additional smoothing then reduces the variance in the second step (Mammen et al., 2012). This is not enough in this case unless we impose sufficient regularity conditions on the ill-posed inverse problem.

Corollary 3.8. *Let us suppose that* $\beta > 2$, $\alpha_n \asymp n^{-1/\beta}$, *and* $h_n \asymp n^{-1/(2\rho)}$, *with* $\rho \geq q(\beta/(\beta-2))$. *Then, under the Assumptions of Theorem 3.7, we obtain*

$$\|\widehat{ASF}(\phi) - ASF(\phi)\|^2_\omega = O_P\left(n^{-1}\right).$$

The condition on the parameter β serves to impose enough regularity on the ill-posed inverse problem. When β is lower than two, the bias converges to zero too slowly to allow for a choice of the regularization parameter, α_n, which controls the variance. Similarly, the condition on the bandwidth parameter h_n allows us to control the bias arising from the nonparametric estimation. Finally, the lower bound on the order of the kernel ρ comes from a *variance* component that needs to be appropriately controlled.

In practice, it is not simple to determine the values of these tuning parameters. In the simulation study, we provide some heuristics for their choice, and we defer a theoretical approach to the matter to future research.

4. Monte-Carlo Simulations

In this section, we provide a Monte-Carlo simulation to explore the finite sample properties of our estimator. We consider two settings. In Setting 1, both X and W are scalar. In Setting 2, both X and W are bivariate to assess the curse of dimensionality in our setting. We work with a mild sample size of $n = 1000$.

The reduced form residual ε is generated according to a normal distribution, and a mixture of normal distributions, with mixing coefficients 0.8 and 0.2. The latter simulation scheme, adapted from Rothe (2009), is employed to assess our estimator's performance under an asymmetric distribution of the error term. The standard deviation of ε is set to be equal to $\sigma_\varepsilon = 0.5$. We estimate the binary choice model conditional on the instruments using a local Logit regression (Logit-type specification). Notice that the model is, therefore, always globally misspecified. We run 1000 replications for each simulation exercise.

We use standard Gaussian kernels. The regularization parameter is computed as explained in Section 3.3. The bandwidth parameters are obtained through least-squares cross-validation. Codes, in MatLab and R, are available upon request from the authors.

In each replication, we compute an estimator of φ as detailed in Section 3.3.

For a clear assessment of its finite sample properties, we compare it with the nonparametric IV estimator that would be obtained if dependent variable Y^* were observed. For simplicity, we refer to the latter as the oracle estimator of φ, and we denote it as $\hat{\varphi}_o^\alpha$. The latter is computed using the same bandwidth and regularization parameters employed for the binary response estimator. For each data generating process, we compute the Mean Integrated Squared Error (MISE) and the Mean Integrated Absolute Error (MIAE).

Finally, we assess the sample behavior of the simulated estimator of the ASF. One key element in this estimation step is the selection of the tuning parameters. For each simulated sample of residuals, we take

$$b_F = c_F n^{-1/4},$$

where c_F is a positive constant. We use a similar adjustment to the bandwidth parameter h_n. As the cross-validated selection of this parameter should lead to $h_n \asymp n^{-1/5}$, we multiply it by a factor of $n^{-1/20}$, to obtain $h_n \asymp n^{-1/4}$. Finally, we use an iterated Tikhonov approach in the regularization step, with the number of iterations equal to 2, so that the qualification of the Tikhonov method increases from 2 to 4. We also take $\alpha = c_\alpha n^{-1/4}$, with c_α a positive constant. For this simulation study, we fix $c_F = c_\alpha = 0.01$.

4.1. Scalar endogenous variable and instrument

The data-generating process is as follows:

$$\begin{aligned} Y^* &= \mathbb{E}(\varphi(X)|W) - \sigma_\varepsilon \varepsilon \\ X &= \exp\left[-(0.15W + \eta)\right] \end{aligned}$$

where

$$\begin{aligned} W &\sim \mathcal{U}[-1.5, 1.5], \\ \eta &\sim \mathcal{N}\left(0, (0.17)^2\right). \end{aligned}$$

Conditionally on η, we generate $\varepsilon|\eta \sim \mathcal{N}\left(\frac{\tau}{0.17}\eta, (1-\tau^2)\right)$, and $\varepsilon|\eta \sim 0.8\mathcal{N}\left(-0.25 + \frac{\tau}{0.17}\eta, 0.5(1-\tau^2)\right) + 0.2\mathcal{N}\left(1 + \frac{\tau}{0.17}\eta, 1.75(1-\tau^2)\right)$, respectively, with $\tau = -0.3$.

We employ two specifications for function φ:

$$\varphi(x) = 0.25x^2, \tag{S$_1$}$$

$$\varphi(x) = -0.25 + \sin(8\log(x)). \tag{S$_2$}$$

The functional forms for φ are employed, as we can analytically compute the corresponding conditional expectation functions. We have

$$\mathbb{E}\left(0.25X^2|W = w\right) = 0.25\exp\left[-0.30w + (0.17)^2\right]$$

and

$$\mathbb{E}(-0.25 + 8\sin(\log(X))|W = w) = -0.25 + \sin(-1.2w)\exp\left[-32(0.17)^2\right].$$

Table 1 reports the results regarding the finite sample performance of our estimator of the regression function φ.

As might be expected, in both simulation studies, the value of the loss function increases compared to the oracle estimator as we have limited information about the dependent variable. As the smoothness of the joint distribution is the same across all data-generating processes, the properties of the Tikhonov estimator generally deteriorate when the smoothness of

Table 1
MISE and MIAE of the binary estimator versus the Oracle (scalar case).

		S_1			S_2		
		$\hat{\varphi}^{\alpha}$	$\hat{\varphi}_o^{\alpha}$	Ratio	$\hat{\varphi}^{\alpha}$	$\hat{\varphi}_o^{\alpha}$	Ratio
Normal	MISE	2.8504	1.9211	1.4837	10.8359	8.8420	1.2255
	MIAE	35.3517	28.3483	1.2471	79.5597	70.2471	1.1326
Mixture	MISE	2.1213	1.7196	1.2336	9.9759	8.9906	1.1096
	MIAE	30.0025	26.8543	1.1172	75.3413	71.2708	1.0571

Table 2
MISE and MIAE of the estimator of r versus the Oracle (scalar case).

		S_1			S_2		
		$\hat{r}$	$\hat{r}_o$	Ratio	$\hat{r}$	$\hat{r}_o$	Ratio
Normal	MISE	0.3239	0.1641	1.9756	0.4047	0.2617	1.5465
	MIAE	21.7435	15.7547	1.3801	27.0162	21.6823	1.2460
Mixture	MISE	0.4052	0.1797	2.2547	0.8445	0.2181	3.8722
	MIAE	26.9166	16.2703	1.6543	41.792	19.489	2.1444

Table 3
MISE and MIAE of the estimator of the Average Structural Function (ASF) versus the Oracle (scalar case).

		S_1			S_2		
		$\hat{F}_U^*$	$\hat{F}_U^o$	Ratio	$\hat{F}_U^*$	$\hat{F}_U^o$	Ratio
Normal	MISE	2.7156	0.2080	13.0564	8.8091	0.1991	44.2367
	MIAE	85.5553	26.0825	3.2802	157.7756	24.6267	6.4067
Mixture	MISE	7.4759	0.2715	27.5403	8.1545	0.2227	36.6105
	MIAE	169.4248	30.1775	5.6143	147.3285	26.7361	5.5105

the function increases. This is because Tikhonov regularization *saturates* at a certain degree of smoothness of the regression function, which may explain the difference between the MISE and the MIAE when we move from a quadratic specification to a sine function. Although the estimator obtained from the limited observation of the dependent variable suffers less from the increase in smoothness than the oracle.

To further analyze how the limited information about the dependent variable affects our final estimator, we also report the MISE and MIAE for the estimator of the conditional expectation of Y^* given W, r. In this case, the oracle is defined from the local linear nonparametric regression of Y^* on W, using the same bandwidth parameter. The more the model is misspecified, the more our estimation of r deviates from the Oracle.

Finally, we report the results for the estimation of the ASF in Table 3 below. The MISE and MIAE of this estimator are given in comparison to an Oracle estimator, which treats the error term as if it were observed. We argue that the differences between the various regularization schemes are mostly due to the precision in the estimation of φ and r (see Tables 1 and 2). Moving from a normal to a mixture of normal distributions increases the misspecification bias, and therefore the properties of the estimator deteriorate in both schemes.

4.2. Bivariate endogenous variable and instrument

The data-generating process is as follows:

$$Y^* = \mathbb{E}(\varphi(X)|W) - \sigma_\varepsilon \varepsilon$$
$$X = \exp\left[-\left(\begin{pmatrix} 0.25 & -0.1 \\ -0.1 & 0.25 \end{pmatrix} W + \eta\right)\right],$$

where:

$$W = \begin{pmatrix} -1.5 + 3\Phi(\omega_1) \\ -1.5 + 3\Phi(\omega_2) \end{pmatrix},$$
$$\eta \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, (0.17)^2 \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right),$$

where ω_1 and ω_2 are normal random variables with correlation coefficient equal to $\pi/12$, and $\Phi(\cdot)$ is the cdf of the standard normal distribution.

Conditionally on η, we generate $\varepsilon|\eta \sim \mathcal{N}\left(\frac{\tau}{0.17}\eta, (1-\tau^2)\right)$, and $\varepsilon|\eta \sim 0.8\mathcal{N}\left(-0.25 + \frac{\tau}{0.17}\eta, 0.5(1-\tau^2)\right) + 0.2\mathcal{N}\left(1 + \frac{\tau}{0.17}\eta, 1.75(1-\tau^2)\right)$, respectively.

Table 4
MISE and MIAE of the binary estimator versus the Oracle (bivariate case).

		S_1			S_2		
		$\hat{\varphi}^{\alpha}$	$\hat{\varphi}^{\alpha}_{o}$	Ratio	$\hat{\varphi}^{\alpha}$	$\hat{\varphi}^{\alpha}_{o}$	Ratio
Normal	MISE	27.3905	23.5766	1.1618	25.3568	24.2381	1.0462
	MIAE	164.3545	150.776	1.0901	156.0822	152.4496	1.0238
Mixture	MISE	21.7446	23.7202	0.9167	24.9813	24.4463	1.0219
	MIAE	141.4864	149.8554	0.9442	154.3329	153.91	1.0027

Table 5
MISE and MIAE of the estimator of the Average Structural Function (ASF) versus the Oracle (bivariate case).

		S_1			S_2		
		$\hat{F}^{*}_{U}$	$\hat{F}^{o}_{U}$	Ratio	$\hat{F}^{*}_{U}$	$\hat{F}^{o}_{U}$	Ratio
Normal	MISE	6.9273	0.2067	33.5128	12.6501	0.2079	60.8591
	MIAE	140.9542	25.9888	5.4237	208.8321	25.1442	8.3054
Mixture	MISE	9.1481	0.2934	31.1781	11.0131	0.2293	48.0241
	MIAE	186.4958	31.0786	6.0008	182.6682	27.2302	6.7083

We employ two specifications for function φ:

$$\varphi(x_1, x_2) = 0.25(x_1 - x_2)^2, \tag{S_1}$$

$$\varphi(x_1, x_2) = -0.25 + \sin(4(\log(x_1) + \log(x_2))). \tag{S_2}$$

As before, we can easily compute

$$\mathbb{E}\big(0.25(X_1 - X_2)^2 | W = w\big) = 0.25\Big\{\exp\left(-2(0.25w_1 - 0.1w_2)\right) + \exp\left(-2(-0.1w_1 + 0.25w_2)\right)\exp(0.17^2)\Big\} - 0.5 * \exp\left(-0.15(w_1 + w_2)\right)\exp(1.5(0.17)^2),$$

and

$$\mathbb{E}(-0.25 + \sin(4(\log(X_1) + \log(X_2)))|W = w) = -0.25 + sin(-0.6(w_1 + w_2))\exp\left[-24(0.17)^2\right].$$

Table 4 reports the results regarding the finite sample performance of our estimator of the regression function φ.

Compared to the scalar case, the properties of both our estimator and the oracle deteriorate, consistently with the usual curse of dimensionality problem in nonparametric estimation. However, our estimator's performance relative to the oracle improves in this setting, most likely because the choice of the tuning parameters for the latter may not be optimal.

Regarding the estimation of the ASF, it is interesting to notice that, while its performance worsens relative to the scalar case, it does so to a lesser extent than the index function's estimator. This supports our theoretical result that faster rates of convergence can be reached for this functional upon an appropriate choice of the tuning parameters.

5. Conclusions

We propose a nonparametric instrumental variable estimator of binary outcome regression models that feature continuous endogenous regressors. We prove the estimator's consistency and show its finite sample properties via a simulation study. Our simulation study shows that our estimator is computationally feasible, easy to implement, and can complement existing semiparametric and parametric models when one wishes to assess potential nonlinearities in the regression function.

Declaration of Competing Interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgements

The authors wish to thank Andrii Babii, Stephane Bonhomme, Andrew Chesher, Patrick Gagliardini, Xavier d'Haultfoeuille, Frank Kleibergen, Blaise Melly, Jeffrey S. Racine, Eric Renault, Peter Robinson, Christoph Rothe, Anna Simoni, seminar participants at Brown University, CREST-ENSAE, McMaster University and the anonymous referees for their useful comments and

remarks. We also thank Stony Brook Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science at Stony Brook University for access to the high-performance SeaWulf computing system, which was made possible by an NSF grant (#1531492). All remaining errors are ours.

Appendix A. Appendix

A1. Additional Assumptions

To obtain uniform consistency of the nonparametric estimators, we must impose some additional assumptions. These are listed below:

Assumption A.1. The probability density function $f_{X,W}(x,w)$ is $d \geq \rho$ times continuously differentiable and uniformly bounded away from 0 and ∞ on its support.

Assumption A.2. The multivariate kernels $K_{W,h}$ and $K_{X,h}$ are product kernels generated by the univariate generalized kernel function K_h satisfying the following properties:

(i) Kernel function $K_h(\cdot,\cdot)$ has order $k \geq \rho \wedge 2$.
(ii) For each $t \in [0,1]$, the function $K_h(h\cdot, t)$ is supported on $[(t-1)/h, t/h] \cap \mathcal{K}$, where $\mathcal{K}$ is a compact interval that does not depend on t and

$$\sup_{h>0, t\in[0,1], u\in\mathcal{K}} |K_h(hu,t)| < \infty$$

(iii) $K_h(\cdot,1) = K_h(\cdot) = K(\cdot/h)$ is a symmetric Lipschitz continuous kernel function with compact support, such that:

$$\mu_{i+j}(K) = \int u^{i+j} K(u)du < \infty, \text{ and}$$
$$\mu_{i+j}(K^2) = \int u^{i+j} K^2(u)du < \infty, \forall j = 1,\dots,\rho.$$

Assumption A.3. The smoothing parameter satisfies $h_n \to 0$ as $n \to \infty$. Additionally

(i) There exists $\lambda \in (2,\infty]$, such that $h_n^q \geq (\log n/n)^{1-2/\lambda}$.
(ii) $(nh_n^{p+q})^{-1}(\log n)^3 \to 0$, and $nh_n^{4[(\rho-j)/2]+2j+5} \log n \to 0$, for $j = 0,\dots,\rho$.

Assumption A.4. The conditional variance of U, given instrument W, is uniformly bounded on $[0,1]^q$.

Assumption A.5. $r(w)$ has at least $\tilde{\rho} = [(\rho-1)/2]+1$ continuous derivatives on $[0,1]^q$, with $(D^{(\tilde{\rho})}r)(w)/\tilde{\rho}!$ uniformly bounded in $[0,1]^q$.

Assumption A.6. The log-likelihood function $\ell(r(w),y)$ satisfies the following properties:

(i) For every y, the function $\ell(r(w),y)$ is at least three-times continuously differentiable with respect to r.
(ii) The following conditional moment holds for all $w \in [0,1]^q$:

$$\mathbb{E}[\ell_1(r(w),Y)|W=w] = 0.$$

(iii) The local Fisher information matrix, defined as

$$I(r(w)) = -\mathbb{E}[\ell_2(r(w),Y)|W=w] = \mathbb{E}\left[(\ell_1(r(w),Y))^2|W=w\right]$$

possesses at least one continuous derivative, and

$$\inf_{w\in[0,1]^q} I(r(w)) > 0.$$

(iv) There exists a neighborhood $\mathcal{B}(r(w))$, such that,

$$\max_{k=1,2} \sup_{w\in[0,1]^q} \left\| \sup_{r\in\mathcal{B}(r(w))} \left| \frac{\partial^k}{\partial r^k} \ell(r(W),Y) \right| \right\|_{\lambda,w} < \infty,$$

for some $\lambda \in (2,\infty]$, where $\|\cdot\|_{\lambda,w}$ denotes the $\mathbb{L}^\lambda$-norm conditional of $W=w$. Furthermore,

$$\sup_{w\in[0,1]^q} \mathbb{E}\left[\sup_{r\in\mathcal{B}(r(w))} \left| \frac{\partial^3}{\partial r^3} \ell(r(W),Y) \right| \right] < \infty.$$

Assumption A.7. Darolles et al. (2011, Assumption A.3 p. 1553) Under the regularity conditions in Assumption A.1, A.2, and A.3, we have

$$\|\hat{T} - T\|^2 = O_P\left(\frac{1}{nh^{p+q}} + h^{2\rho}\right),$$

and

$$\|\hat{T}^* - T^*\|^2 = O_P\left(\frac{1}{nh^{p+q}} + h^{2\rho}\right).$$

Assumption A.1, A.2, A.3, and A.4 are needed to obtain the uniform convergence of the nonparametric estimators of the joint and marginal densities. Assumption A.4, A.5, and A.6 are used in Lemma 3.2 to claim the existence and global consistency of the local likelihood estimator of r.

Assumption A.7 is needed to prove the consistency result of Theorem 3.3 and gives the order of convergence for the estimators of the conditional expectation operators. This assumption is proven in Darolles et al. (2011), and its proof is not reproduced here.

In the proofs below, we make extensive use of the following lemma (Engl et al., 2000, Th. 4.3, p. 74):

Lemma A.1. *For a given $v \in \mathbb{L}_X^2$, such that $\|v\| < \infty$:*

$$\alpha^{2-\nu}\|(\alpha I + T^*T)^{-1}(T^*T)^{\frac{\nu}{2}}v\|^2 = O_P(1), \tag{22}$$

where $\nu \leq 2$.

A2. Proof of Lemma 3.2

Since we are using local polynomials for estimation, we do not require correction at the boundary points of the support of W. Thus, for simplicity, in this proof, we use the usual notation

$$K_h(W_i - w) = K_h(W_i - w, 1) = K\left(\frac{W_i - w}{h}\right),$$

for a symmetric q-variate product kernel with compact support. Therefore, W_i is the q-dimensional vector of sample observations, and w is the q dimensional vector of evaluation points, respectively.

To prove the main result of the Lemma, we require some additional notations, and some background results about the asymptotic decomposition of the estimator of r. For simplicity, we take $\tilde{\rho} = \rho$.

Let $N_k = \binom{k+\rho-2}{\rho-2}$ be the number of distinct $(\rho-1)$–tuples $\mathbf{j}$ with $|j| = k$. Arrange these N_k $(\rho-1)$–tuples as a sequence in a lexicographical order (with the highest priority given to the last position so that $(0,\dots,0,k)$ is the first element in the sequence, and $(k,0,\dots,0)$ is the last element). Let τ_k denote this one-to-one mapping; that is, $\tau_k(1) = (0,\dots,0,k),\dots,\tau_k(N_k) = (k,0,\dots,0)$. Finally, $N = \sum_{k=1}^{\rho-1} N_k$. We then follow the notation in Assumption A.2 and let $\mu_k(K) = \int u^k K(u)du$ and define $\mu_{nk}(K,w) = \int u^k K(u)\mathbb{E}[\ell_2(\mathbf{W}_i\gamma_0, Y_i)|W_i = w + hu]f_W(w+hu)du$, where

$$\underbrace{\mathbf{W}_i}_{1\times N} = \{(W_i - w)^{\mathbf{j}}\}_{0\leq|\mathbf{j}|\leq\rho-1}.$$

For $0 \leq j, k \leq \rho-1$, let $M_{j,k}$ and $M_{n,j,k}(w)$ be two $N_j \times N_k$ matrices with their (l, m) elements, respectively, given by

$$\left[M_{j,k}\right]_{l,m} = \mu_{\tau_j(l)+\tau_k(m)}(K), \qquad \left[M_{n,j,k}(w)\right]_{l,m} = \mu_{n,\tau_j(l)+\tau_k(m)}(K,w).$$

We define the $N \times N$ matrices, $\mathbf{M}_{\rho-1}$ and $\mathbf{M}_{n,\rho-1}(w)$, as follows,

$$\mathbf{M}_{\rho-1} = \begin{bmatrix} M_{0,0} & M_{0,1} & \dots & M_{0,\rho-1} \\ M_{1,0} & M_{1,1} & \dots & M_{1,\rho-1} \\ \vdots & \vdots & \ddots & \vdots \\ M_{\rho-1,0} & M_{\rho-1,1} & \dots & M_{\rho-1,\rho-1} \end{bmatrix}$$

$$\mathbf{M}_{n,\rho-1}(w) = \begin{bmatrix} M_{n,0,0}(w) & M_{n,0,1}(w) & \dots & M_{n,0,\rho-1}(w) \\ M_{n,1,0}(w) & M_{n,1,1}(w) & \dots & M_{n,1,\rho-1}(w) \\ \vdots & \vdots & \ddots & \vdots \\ M_{n,\rho-1,0}(w) & M_{n,\rho-1,1}(w) & \dots & M_{n,\rho-1,\rho-1}(w) \end{bmatrix}.$$

It follows from our definitions that

$$\mathbf{M}_{n,\rho-1}(w) - f_W(w)I(r(w))\mathbf{M}_{\rho-1} = O(h_n),$$

where $I(r(w))$ is the local information matrix, as defined in Assumption A.6(iii). Further, denote

$$\underbrace{\mathbf{W}_{li}}_{1\times N} = \{(W_l - W_i)^{\mathbf{j}}\}_{0\leq|\mathbf{j}|\leq\rho-1}, \text{ and } \mathbf{W}_{li}\mathbf{H}_n^{-1} = \left\{\left(\frac{W_l - W_i}{h}\right)^{\mathbf{j}}\right\}_{0\leq|\mathbf{j}|\leq\rho-1},$$

where $\mathbf{H}_n = Diag\left(1,\dots,h_n^{\rho-1}\right)$ is a $N \times N$ diagonal matrix of bandwidth parameters.

In the following, we also let

$$\mathcal{J}(w) = f_W(w) I(r(w)) M_{\rho-1}.$$

A.2[Asymptotic expansion of the local likelihood estimator] Define

$$\hat{\gamma}_{-i}(W_i) = \arg\max_{\gamma} \frac{1}{(n-1)h_n^q} \sum_{l=1, l\neq i}^{n} \ell\left(\mathbf{W}_{li}\gamma_{\mathbf{j}}(W_i), Y_l\right) K_{h_n}(W_l - W_i),$$

as the leave-one-out local likelihood estimator. Under Assumption A.1, A.2(iii), and A.3-A.6, we obtain

$$\begin{aligned}\hat{\gamma}_{-i}(W_i) - \gamma_0(W_i) = \mathcal{J}(W_i)^{-1} &\left[\frac{\mathbf{H}_n^{-1}}{(n-1)h_n^q} \sum_{l=1, l\neq i}^{n} \left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_l), Y_l) K_{h_n}(W_l - W_i) \right] \\ &+ h_n^{\rho}\left(\mathbf{M}_{\rho-1}\mathbf{H}_n\right)^{-1} \sum_{0\leq|\mathbf{k}|\leq\rho} \sum_{|\mathbf{j}|=\rho} \frac{D^{(\mathbf{j})} r(W_i)}{\mathbf{j}!} M_{\mathbf{k},\mathbf{j}} + o_P((nh_n^q \log n)^{-1/2}).\end{aligned}$$

Proof. From Claeskens and Van Keilegom (2003, Corollary 2.1, p. 1856), upon Assumption A.1, A.2(iii), and A.3-A.6, we have that

$$\begin{aligned}\hat{\gamma}_{-i}(W_i) - \gamma_0(W_i) = \mathcal{J}(W_i)^{-1} &\left[\frac{\mathbf{H}_n^{-1}}{(n-1)h^q} \sum_{l=1, l\neq i}^{n} \left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)' \ell_1(\mathbf{W}_{li}\gamma_0(W_i), Y_l) K_h(W_l - W_i) \right] \\ &+ o_P((nh_n^q \log n)^{-1/2}).\end{aligned}$$

We can further take a Taylor expansion of the score function around $r(W_l)$, which gives

$$\begin{aligned}\ell_1(\mathbf{W}_{li}\gamma_0, Y_l) &= \ell_1(r(W_l), Y_l) + \ell_2(r(W_l), Y_l)(\mathbf{W}_{li}\gamma_0 - r(W_l)) + R_2(\mathbf{W}_{li}\gamma_0) \\ &= \ell_1(r(W_l), Y_l) + \ell_2(r(W_l), Y_l) \sum_{|\mathbf{j}|=\rho} \frac{D^{(\mathbf{j})} r(W_i)}{\mathbf{j}!} (W_l - W_i)^{\mathbf{j}} + R_2(\mathbf{W}_{li}\gamma_0),\end{aligned}$$

where $R_2(\mathbf{W}_{li}\gamma_0)$ is a remainder term. Finally, by a change of variable,

$$\begin{aligned}&\mathcal{J}(W_i)^{-1}\mathbf{H}_n^{-1}\mathbb{E}\left[\left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)' \ell_2(r(W_l), Y_l) \sum_{|\mathbf{j}|=\rho} \frac{D^{(\mathbf{j})} r(W_i)}{\mathbf{j}!} (W_l - W_i)^{\mathbf{j}} \frac{1}{h^q} K_h(W_l - W_i) | W_i \right] \\ &= \mathcal{J}(W_i)^{-1}\mathbf{H}_n^{-1}\mathbb{E}\Bigg[\left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)' \mathbb{E}[\ell_2(r(W_l), Y_l)|W_l] \times \\ &\quad \sum_{|\mathbf{j}|=\rho} \frac{D^{(\mathbf{j})} r(W_i)}{\mathbf{j}!} (W_l - W_i)^{\mathbf{j}} \frac{1}{h^q} K_h(W_l - W_i) | W_i \Bigg] \\ &= h_n^{\rho}\left(\mathbf{M}_{\rho-1}\mathbf{H}_n\right)^{-1} \sum_{0\leq|\mathbf{k}|\leq\rho} \sum_{|\mathbf{j}|=\rho} \frac{D^{(\mathbf{j})} r(W_i)}{\mathbf{j}!} M_{\mathbf{k},\mathbf{j}} (1 + o_P(1)),\end{aligned}$$

where the last line follows from the uniform boundedness of the third derivative of the log-likelihood function and the derivatives of the function r up to the order ρ. □

The existence and uniform consistency of the local likelihood estimator follows directly from Assumption A.3 and A.6 using arguments similar to those of Zhao (1994) and Claeskens and Van Keilegom (2003).

Given the result of Lemma A.2, to prove $\mathbb{L}^2$ consistency of the local likelihood estimator, we only need to obtain the rates of each of the terms of the stochastic decomposition. That is

$$\begin{aligned}&\mathbb{E}\|e'_{j+1}\left(\hat{\gamma}_{-i} - \gamma_0\right)\|^2 = \mathbb{E}\int_{[0,1]^q} \left[e'_{j+1}\left(\hat{\gamma}_{-i}(w) - \gamma_0(w)\right)\right]^2 f_W(w) dw \\ \leq &2\mathbb{E}\int_{[0,1]^q} \left\{ e'_{j+1}\left(\mathcal{J}(w)^{-1}\left[\frac{\mathbf{H}_n^{-1}}{(n-1)h_n^q} \sum_{l=1, l\neq i}^{n} \left(\mathbf{W}_l\mathbf{H}_n^{-1}\right)' \ell_1(r(W_l), Y_l) K_{h_n}(W_l - w) \right] \right) \right\}^2 f_W(w) dw \\ &+ 2h_n^{2\rho} \int_{[0,1]^q} \left[e'_{j+1}\left(\mathbf{M}_{\rho-1}\mathbf{H}_n\right)^{-1} \sum_{0\leq|\mathbf{k}|\leq\rho} \sum_{|\mathbf{j}|=\rho} \frac{D^{(\mathbf{j})} r(w)}{\mathbf{j}!} M_{\mathbf{k},\mathbf{j}} \right]^2 f_W(w) dw.\end{aligned}$$

Because of the uniform boundedness of the ρ^{th} derivatives of the function r in Assumption A.4, we directly have

$$e'_{j+1} \left\| \left(\mathbf{M}_{\rho-1}\mathbf{H}_n\right)^{-1} \sum_{0\leq|\mathbf{k}|\leq\rho} \sum_{|\mathbf{j}|=\rho} \frac{D^{(\mathbf{j})} r(W_i)}{\mathbf{j}!} M_{\mathbf{k},\mathbf{j}} \right\|^2 e_{j+1} = O\left(h_n^{-2j}\right),$$

which directly implies that the second term is of the order $h_n^{2(\rho-j)}$.

Similarly, we notice that

$$\begin{aligned}
&\mathbb{E}\left[e_{j+1}'\left(\mathcal{J}(w)^{-1}\left[\frac{\mathbf{H}_n^{-1}}{(n-1)h_n^q}\sum_{l=1,l\neq i}^{n}\left(\mathbf{W}_l\mathbf{H}_n^{-1}\right)'\ell_1(r(W_l),Y_l)K_{h_n}(W_l-w)\right]\right)\right]^2\\
&\leq\frac{1}{(n-1)h_n^{2q}}\mathbb{E}\left[e_{j+1}'\left((\mathbf{H}_n\mathcal{J}(w))^{-1}\left(\mathbf{W}_l\mathbf{H}_n^{-1}\right)'\ell_1(r(W_l),Y_l)K_{h_n}(W_l-w)\right)\right]^2\\
&=\frac{1}{(n-1)h_n^{2q}}e_{j+1}'(\mathbf{H}_n\mathcal{J}(w))^{-1}\times\\
&\quad\mathbb{E}\left[\left(\mathbf{W}_l\mathbf{H}_n^{-1}\right)'I(r(W_l))\left(\mathbf{W}_l\mathbf{H}_n^{-1}\right)K_{h_n}^2(W_l-w)\right]\left(f_W(w)I(r(w))M_{\rho-1}\mathbf{H}_n\right)^{-1}e_{j+1}\\
&=\frac{1}{(n-1)h_n^{q+2j}}e_{j+1}'\mathcal{J}(w)^{-1}\mu_{2j}(K^2)\left(M_\rho\right)^{-1}e_{j+1}(1+o(1)),
\end{aligned}$$

where $\mu_j(K^2)$ is defined in Assumption A.2(iii). We finally have

$$\begin{aligned}
&\mathbb{E}\int\left[e_{j+1}'\left(\mathcal{J}(w)^{-1}\left[\frac{\mathbf{H}_n^{-1}}{(n-1)h_n^q}\sum_{l=1,l\neq i}^{n}\left(\mathbf{W}_l\mathbf{H}_n^{-1}\right)'\ell_1(r(W_l),Y_l)K_{h_n}(W_l-w)\right]\right)\right]^2 f_W(w)dw\\
&=\frac{1}{(n-1)h_n^{q+2j}}e_{j+1}'\mathcal{J}(w)^{-1}\mu_{2j}(K^2)\left(M_{\rho-1}\right)^{-1}e_{j+1}=O\left(\frac{1}{nh_n^{q+2j}}\right).
\end{aligned}$$

The first part of the Lemma follows from Markov's inequality by taking $j=0$.

To prove the second part of the Lemma, we proceed as follows. Let us write

$$\begin{aligned}
\hat{T}\left(\hat{r}_{(-i)}-r\right)(x)&=\frac{\frac{1}{nh_n^p}\sum_{i=1}^n K_{h_n}(X_i-x,x)\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)}{\hat{f}_X(x)}\\
&=\frac{\frac{1}{nh_n^p}\sum_{i=1}^n K_{h_n}(X_i-x,x)\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)}{f_X(x)}(1+o_P(1))\\
&=\frac{\frac{1}{nh_n^p}\sum_{i=1}^n K_{h_n}(X_i-x,x)e_1'\mathcal{J}(W_i)^{-1}\left[\frac{\mathbf{H}_n^{-1}}{(n-1)h_n^q}\sum_{l=1,l\neq i}^n\left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)'\ell_1(r(W_l),Y_l)K_{h_n}(W_l-W_i)\right]}{f_X(x)}(1+o_P(1))\\
&\quad+\frac{\frac{h_n^\rho}{nh_n^p}\sum_{i=1}^n K_{h_n}(X_i-x,x)e_1'\left(\mathbf{M}_{\rho-1}\mathbf{H}_n\right)^{-1}\sum_{0\leq|\mathbf{k}|\leq\rho}\sum_{|\mathbf{j}|=\rho}\frac{D^{(\mathbf{j})}r(W_i)}{\mathbf{j}!}M_{\mathbf{k},\mathbf{j}}}{f_X(x)}(1+o_P(1)),
\end{aligned}$$

where the first equality follows from the uniform consistency of the kernel density estimator, and the second equality follows from Lemma A.2, where we have omitted negligible terms.

It is immediate from this decomposition that the latter term is of the order h_n^ρ, directly by Assumption A.4. We, therefore, focus our proof on the first component.

We let

$$g(X_i,W_i,Y_l,W_l)=\frac{1}{h_n^p}K_{h_n}(X_i-x,x)e_1'\mathcal{J}(W_i)^{-1}\left[\frac{\mathbf{H}_n^{-1}}{h_n^q}\left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)'\ell_1(r(W_l),Y_l)K_{h_n}(W_l-W_i)\right],$$

in a way that one can write

$$\mathbf{R}_n=\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{l=1,l\neq i}^n g(X_i,W_i,Y_l,W_l)=\frac{2}{n(n-1)}\sum_{1\leq i<l\leq n}R(Z_i,Z_l),$$

which is a second-order U-statistics, where $Z_i=(Y_i,X_i,W_i)$, and

$$R(Z_i,Z_l)=\frac{1}{2}[g(X_i,W_i,Y_l,W_l)+g(X_l,W_l,Y_i,W_i)].$$

Notice that the function R depends on the score function $\ell_1(r(W_i),Y_i)$, which satisfies $\mathbb{E}[\ell_1(r(W_i),Y_i)|W_i]=0$, so that

$$\mathbb{E}[R(Z_i,Z_l)]=0.$$

Similarly,

$$\mathbb{E}[R(Z_i,Z_l)|Z_i]=\frac{1}{2}\mathbb{E}[g(X_l,W_l,Y_i,W_i)|Z_i]$$

$$= \frac{1}{2h_n^{p+q}}\mathbb{E}\Big[K_{h_n}(X_l - x, x)e_1'\mathcal{J}(W_i)^{-1}\mathbf{H}_n^{-1}\Big(\mathbf{W}_{il}\mathbf{H}_n^{-1}\Big)'\ell_1(r(W_i), Y_i)K_{h_n}(W_l - W_i)|Z_i\Big]$$
$$= \frac{1}{2}e_1'\mathcal{J}(W_i)^{-1}\mathbf{H}_n^{-1}\ell_1(r(W_i), Y_i)\frac{f_{XW}(x, W_i)}{f_X(x)} + O_P(h_n^{\rho}),$$

directly from a change of variable, the kernel properties in Assumption A.2, and the uniform differentiability of the joint density of (X, W) from Assumption A.1. We ignore the second term of the variance of the U-statistics, which is negligible as long as $nh_n^{p+q} = o(1)$, granted by Assumption A.3(i). Therefore, we finally have

$$Var(\mathbf{R}_n) = \frac{4(n-2)}{n(n-1)}Var\left(\frac{1}{2}e_1'\mathcal{J}(W_i)^{-1}\mathbf{H}_n^{-1}\ell_1(r(W_i), Y_i)\frac{f_{XW}(x, W_i)}{f_X(x)}\right)$$
$$= \frac{n-2}{n(n-1)}\mathbb{E}\left[e_1'\Big(I(r(W_i))M_{\rho-1}\Big)^{-1}\mathbf{H}_n^{-1}I(r(W_i))\mathbf{H}_n^{-1}\Big(M_{\rho-1}I(r(W_i))\Big)^{-1}e_1\frac{f_{XW}^2(x, W_i)}{f_X^2(x)f_W^2(W_i)}\right]$$
$$= \frac{n-2}{n(n-1)}\mathbb{E}\left[I(r(W_i))^{-1}\frac{f_{XW}^2(x, W_i)}{f_X^2(x)f_W^2(W_i)}\right],$$

where the second equality follows from the fact that

$$\mathbb{E}\Big[\ell_1(r(W_i), Y_i)\ell_1'(r(W_i), Y_i)|W_i\Big] = I(r(W_i)),$$

and the latter term is well defined because of Assumption A.6(iii). Therefore, we conclude that

$$Var(\mathbf{R}_n) = O(n^{-1}),$$

and the second part of the Lemma follows from Markov's inequality.

A3. Proof of Theorem 3.3

First, notice that, under Assumption 3.1, we have $r = T\varphi \in \mathcal{R}\Big[(TT^*)^{\frac{\beta+1}{2}}\Big]$. Therefore, following the same line of proof used for Proposition 3.1 and the result in Lemma A.1, it can be shown that

$$\|(\alpha I + TT^*)^{-1}r\|^2 = O\Big(\alpha^{(\beta+1)\wedge 2}\Big). \tag{23}$$

Recall that

$$\varphi^{\alpha_n} = (\alpha_n I + T^*T)^{-1}T^*T\varphi$$

We can write the decomposition of the regularized estimator as

$$\hat{\varphi}^{\alpha_n} - \varphi = \hat{\varphi}^{\alpha_n} - \varphi^{\alpha_n} + \varphi^{\alpha_n} - \varphi$$
$$= \Big(\alpha_n I + \hat{T}^*\hat{T}\Big)^{-1}\hat{T}^*\hat{r} - \Big(\alpha_n I + \hat{T}^*\hat{T}\Big)^{-1}\hat{T}^*r \tag{A_1}$$
$$+ \Big(\alpha_n I + \hat{T}^*\hat{T}\Big)^{-1}\hat{T}^*T\varphi - (\alpha_n I + T^*T)^{-1}T^*T\varphi \tag{A_2}$$
$$- \alpha_n(\alpha_n I + T^*T)^{-1}\varphi. \tag{A_3}$$

From Proposition 3.1, we directly have

$$\|A_3\|^2 = O_P\Big(\alpha_n^{\beta\wedge 2}\Big).$$

We now focus on the term A_2. Notice that it can be written as follows:

$$A_2 = \hat{T}^*\Big(\alpha_n I + \hat{T}\hat{T}^*\Big)^{-1}T\varphi - T^*(\alpha_n I + TT^*)^{-1}T\varphi$$
$$= \hat{T}^*\left[\Big(\alpha_n I + \hat{T}\hat{T}^*\Big)^{-1} - (\alpha_n I + TT^*)^{-1}\right]T\varphi + \Big(\hat{T}^* - T^*\Big)(\alpha_n I + TT^*)^{-1}T\varphi$$
$$= \hat{T}^*\left[\Big(\alpha_n I + \hat{T}\hat{T}^*\Big)^{-1}\hat{T}\Big(\hat{T}^* - T^*\Big)(\alpha_n I + TT^*)^{-1}\right]T\varphi \tag{A_{21}}$$
$$+ \hat{T}^*\left[\Big(\alpha_n I + \hat{T}\hat{T}^*\Big)^{-1}\Big(\hat{T} - T\Big)T^*(\alpha_n I + TT^*)^{-1}\right]T\varphi \tag{A_{22}}$$
$$+ \Big(\hat{T}^* - T^*\Big)(\alpha_n I + TT^*)^{-1}T\varphi \tag{A_{23}}.$$

We start by showing that

$$\|A_{23}\|^2 \leq \|\hat{T}^* - T^*\|^2 \|(\alpha_n I + TT^*)^{-1} T\varphi\|^2.$$

From lemma (A.1) and equation (23), we have

$$\|(\alpha_n I + TT^*)^{-1} T\varphi\|^2 = \frac{\alpha_n^{1-\beta}}{\alpha_n^{1-\beta}} \|(\alpha_n I + TT^*)^{-1} T\varphi\|^2 = O\left(\alpha_n^{(\beta-1)\wedge 0}\right).$$

This is because, by spectral theory, the operator T's singular values are the same as the eigenvalues of the operators $(T^*T)^{1/2}$ and $(TT^*)^{1/2}$. Therefore

$$\|(\alpha_n I + TT^*)^{-1} T\varphi\|^2 = \|(\alpha_n I + TT^*)^{-1} (TT^*)^{(\beta+1)/2} v\|^2,$$

in a way that lemma (A.1) can be applied with $\nu = \beta + 1$ (see also Darolles et al., 2011, p. 1563, and Carrasco et al., 2007, Proposition 3.8, p. 5674). Therefore, it is straightforward to show that

$$\|A_{23}\|^2 = O_P\left[\left(\frac{1}{nh^{p+q}} + h^{2\rho}\right)\alpha_n^{(\beta-1)\wedge 0}\right],$$

where the bound on $\hat{T}^* - T^*$ comes from the properties of the nonparametric estimation of the conditional expectation operator given in Assumption A.7.

In the same way, we can write:

$$\begin{aligned}\|A_{21}\|^2 &\leq \|\hat{T}^*\left(\alpha_n I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\|^2 \|\hat{T}^* - T^*\|^2 \|(\alpha_n I + TT^*)^{-1} T\varphi\|^2 \\ &= O_P\left[\left(\frac{1}{nh^{p+q}} + h^{2\rho}\right)\alpha_n^{(\beta-1)\wedge 0}\right],\end{aligned}$$

as

$$\|\hat{T}^*\left(\alpha_n I + \hat{T}\hat{T}^*\right)^{-1}\hat{T}\|^2 = O_P(1),$$

Finally,

$$\begin{aligned}\|A_{22}\|^2 \leq& \|\hat{T}^*\left(\alpha_n I + \hat{T}\hat{T}^*\right)^{-1}\|^2 \|\hat{T} - T\|^2 \|(\alpha_n I + T^*T)^{-1} T^*T\varphi\|^2 \\ &= O_P\left[\left(\frac{1}{nh^{p+q}} + h^{2\rho}\right)\alpha_n^{(\beta-1)\wedge 0}\right],\end{aligned}$$

as

$$\|\hat{T}^*\left(\alpha_n I + \hat{T}\hat{T}^*\right)^{-1}\|^2 = O_P\left(\frac{1}{\alpha_n}\right), \tag{24}$$

and

$$\frac{1}{\alpha_n}\|(\alpha_n I + T^*T)^{-1} T^*T\varphi\|^2 = \frac{\alpha_n^{-\beta}}{\alpha_n^{1-\beta}} \|(\alpha_n I + T^*T)^{-1} T^*T\varphi\|^2 = O\left(\alpha_n^{(\beta-1)\wedge 0}\right).$$

We can now move to the first term of the expansion, A_1. We can either write

$$\|A_1\|^2 \leq \|\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\|^2 \|\hat{r} - r\|^2 = O_P\left[\frac{1}{\alpha_n}\left(\frac{1}{nh^q} + h^{2\rho}\right)\right],$$

or

$$\|A_1\|^2 \leq \|\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\|^2 \|\hat{T}^*\left(\hat{r} - r\right)\|^2 = O_P\left[\frac{1}{\alpha_n^2}\left(\frac{1}{n} + h_n^{2\rho}\right)\right],$$

by equation (A.24) and Lemma 3.2.

A4. Proof of Corollary 3.4

The regularization bias $\alpha_n^{\beta\wedge 2}$ converges to zero for any $\beta > 0$, and $\alpha_n \to 0$. The condition $h_n^\rho \alpha_n^{-1} = o(1)$ also guarantees that bandwidth and regularization parameters are chosen so that the bias goes to zero as $n \to \infty$. We, therefore, only need to control the variance components in the decomposition.

When $\beta \geq 1$, the middle term in the upper bound only depends on the nonparametric estimation error. Therefore, as long as h_n^{p+q} converges to zero slower than $\alpha_n\left(h_n^q \wedge \alpha_n\right)$, the middle term is negligible, and the first term dominates. Otherwise, if $\beta < 1$, the conclusion of the Corollary is reached as long as the largest variance term converges to zero. This concludes the proof.

A5. Proof of Proposition 3.5

From the proof of Lemma A.2 and provided that $h_n = o_P(n^{-1/2\rho})$, we have that

$$\hat{T}^*\left(\hat{r}_{(-i)} - r\right)(x) = \frac{1}{n}\sum_{i=1}^{n} e_1' \mathcal{J}(W_i)^{-1}\mathbf{H}_n^{-1}\ell_1(r(W_i), Y_i)\frac{f_{XW}(x, W_i)}{f_X(x)}(1 + o_P(1)) = \frac{1}{n}\sum_{i=1}^{n}\Upsilon_i(x).$$

Because of Assumption A.1 and A.6(iii), it is immediate that $\mathbb{E}\|\Upsilon_i\|^2 < \infty$. This condition is sufficient to apply a functional version of the central limit theorem (van der Vaart and Wellner, 1996, Theorem 1.8.4, p. 50). For functions $\psi_1, \psi_2 \in \mathbb{L}^2(X)$, the variance is given by an operator Ω defined as

$$\begin{aligned}\langle \Omega\psi_1, \psi_2\rangle &= \int\int\int (I(r(W_i)))^{-1}\frac{f_{XW}(x_1, W_i)f_{XW}(x_2, W_i)}{f_X(x_1)f_X(x_2)f_W(W_i)}\psi_1(x_1)\psi_2(x_2)dx_1dx_2dW_i \\ &= \int\left\{\int (I(r(W_i)))^{-1}\left[\int \psi_1(x_1)f_{X|W}(x_1|W_i)dx_1\right]f_{W_i|X}(W_i|x_2)dW_i\right\}\psi_2(x_2)f_X(x_2)dx_2 \\ &= \langle T^*I(r)T\psi_1, \psi_2\rangle,\end{aligned}$$

from which the result of the Proposition follows.

A6. Proof of Theorem 3.6

We use a similar decomposition as in Theorem 3.3. We let

$$\begin{aligned}&\sqrt{n}\langle\hat{\varphi}^{\alpha_n} - \varphi^{\alpha_n}, \delta\rangle \\ &= \sqrt{n}\langle M\hat{T}^*\left(\hat{r}_{(-i)} - r\right), \delta\rangle &(A_1)\\ &\quad + \sqrt{n}\langle\left(\hat{M} - M\right)\hat{T}^*\left(\hat{r}_{(-i)} - r\right), \delta\rangle &(A_2)\\ &\quad + \sqrt{n}\langle\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*T\varphi - (\alpha_n I + T^*T)^{-1}T^*T\varphi, \delta\rangle. &(A_3)\end{aligned}$$

The first term is the leading term of our decomposition. Using the result of Proposition 3.5, we directly have

$$\sqrt{n}M\hat{T}^*\left(\hat{r}_{(-i)} - r\right) \Rightarrow N(0, M^*\Omega M),$$

in a way that

$$\langle M^*\Omega M\delta, \delta\rangle = \|\Omega^{1/2}M\delta\|^2,$$

and

$$\sqrt{n}\frac{\langle M\hat{T}^*\left(\hat{r}_{(-i)} - r\right), \delta\rangle}{\|\Omega^{1/2}M\delta\|} \xrightarrow{d} N(0, 1).$$

The rest of the proof is about showing that the remaining terms are negligible under the conditions given in the statement of the Theorem. Recall from Assumption A.7 that

$$\|\hat{T}^* - T^*\|^2 \asymp \|\hat{T} - T\|^2 \asymp \|\hat{T}^*\hat{T} - T^*T\|^2 = O_P\left(\frac{1}{nh_n^{p+q}} + h_n^{2\rho}\right).$$

From this fact, $\|\delta\| < \infty$, and the proof of Theorem 3.3, we have

$$\begin{aligned}\|A_3\|^2 \leq& n\|\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*T\varphi - (\alpha_n I + T^*T)^{-1}T^*T\varphi\|^2\|\delta\|^2 \\ &= O_P\left(n\|\hat{T}^*\hat{T} - T^*T\|^2\alpha_n^{\beta-1\wedge 0}\right).\end{aligned}$$

Therefore, as long as

$$\frac{\sqrt{n}\|\hat{T}^*\hat{T} - T^*T\|\alpha_n^{\frac{\beta-1}{2}\wedge 0}}{\|\Omega^{1/2}M\delta\|} = o_P(1),$$

the term in A_3 is negligible.

We now focus on the term A_2. Notice that it can be written as follows:

$$\begin{aligned}\|A_2\|^2 &= n\langle\left(\hat{M} - M\right)\hat{T}^*\left(\hat{r}_{(-i)} - r\right), \delta\rangle^2 \\ &= n\langle\hat{T}^*\left(\hat{r}_{(-i)} - r\right), \left(\hat{M} - M\right)\delta\rangle^2 \\ &\leq \left\|\sqrt{n}\hat{T}^*\left(\hat{r}_{(-i)} - r\right)\right\|^2\left\{\|\hat{M}\hat{T}^*\left(T - \hat{T}\right)M\delta\|^2 + \|\hat{M}\left(T^* - \hat{T}^*\right)TM\delta\|^2\right\}.\end{aligned}$$

From the proof of Proposition 3.5, we have that

$$\left\| \sqrt{n}\hat{T}^* \left(\hat{r}_{(-i)} - r \right) \right\|^2 = O_P(1).$$

Finally, reasoning as in the proof of Theorem 3.3 and using Assumption 3.2, we also obtain that

$$\|\hat{M}\hat{T}^* \left(T - \hat{T} \right) M\delta\|^2 + \|\hat{M} \left(T^* - \hat{T}^* \right) TM\delta\|^2 = O_P \left(\|\hat{T}^*\hat{T} - T^*T\| \alpha_n^{\nu - 1 \wedge 0} \right)$$

Therefore, as long as

$$\frac{\|\hat{T}^*\hat{T} - T^*T\| \alpha_n^{\frac{\nu-1}{2} \wedge 0}}{\|\Omega^{1/2} M\delta\|} = o_P(1),$$

the term in A_2 is negligible.

A7. Proof of Theorem 3.7

Assumption A.8. The joint distribution of U^*, X, and W is uniformly bounded away from infinity and at least ρ times continuously differentiable with respect to any combination of its arguments. Its derivatives up to the order ρ are uniformly bounded by a constant.

The latter Assumption imposes a smoothness restriction on the joint distribution of the simulated error term, U^*, and the regressor X. This assumption is necessary in order to overcome the ill-posedness common to nonparametric instrumental regressions (see Babii and Florens, 2017a, for a similar assumption).

Let us write

$$\begin{aligned} &\hat{F}^*_U(\phi) - F_U(\phi) \\ &= \hat{F}^*_U(\phi) - \hat{F}_{U^*}(\phi) && (A_1) \\ &+ \hat{F}_{U^*}(\phi) - F_U(\phi) && (A_2) \end{aligned}$$

where we denote

$$\hat{F}_{U^*}(\phi) = \frac{1}{n} \sum_{i=1}^n \bar{K}_{b_F} \left(U_i^* - \phi \right),$$

with $U_i^* = Q_i^* + \varphi^*(X_i) - r(W_i)$.

Given Assumption 2.2 and 2.3 and existing results in nonparametric estimation, we immediately have

$$\|A_2\|_\omega^2 = O_P \left(\frac{1}{n} + b_F^{2\rho} \right) = O_P(n^{-1}).$$

We then focus on the properties of

$$A_1 = \hat{F}^*_U(\phi) - \hat{F}_{U^*}(\phi) = \frac{1}{n} \sum_{i=1}^n \left[\bar{K}_{b_F} \left(\hat{U}_i^* - \phi \right) - \bar{K}_{b_F} \left(U_i^* - \phi \right) \right].$$

We write

$$\begin{aligned} \|A_1\|_\omega^2 &= \mathbb{E} \int \left\{ \frac{1}{n} \sum_{i=1}^n \left[\bar{K}_{b_F} \left(\hat{U}_i^* - \phi \right) - \bar{K}_{b_F} \left(U_i^* - \phi \right) \right] \right\}^2 \omega(\phi) d\phi \\ &\leq 2\mathbb{E} \int \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{b_F} K_{b_F} \left(U_i^* - \phi \right) \left(\hat{U}_i^* - U_i^* \right) \right\}^2 \omega(\phi) d\phi && (A_{11}) \\ &+ 2\mathbb{E} \int \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{2b_F} K'_{b_F} \left(\tilde{U}_i - \phi \right) \left(\hat{U}_i^* - U_i^* \right)^2 \right\}^2 \omega(\phi) d\phi && (A_{12}), \end{aligned}$$

where the second line follows from Taylor's theorem, with $\tilde{U}_i$ being a point between $\hat{U}_i^*$ and U_i^*, the Cauchy-Schwarz and Young inequality for products.

In the following, we take the support of ε to be the entire real line. We first take care of the term in A_{12}. We have that

$$\left| \frac{1}{n} \sum_{i=1}^n \frac{1}{2b_F} K'_{b_F} \left(\tilde{U}_i - \phi \right) \left(\hat{U}_i^* - U_i^* \right)^2 \right| \leq \frac{1}{n} \sum_{i=1}^n \frac{1}{2b_F} \left| K'_{b_F} \left(\tilde{U}_i - \phi \right) \right| \left(\hat{U}_i^* - U_i^* \right)^2.$$

By Markov inequality

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2b_F}\left|K'_{b_F}\left(\bar{U}_i-\phi\right)\right|\left(\hat{U}_i^*-U_i^*\right)^2\geq t\right)\leq\frac{\mathbb{E}\left[\frac{1}{2b_F}\left|K'_{b_F}\left(\bar{U}_i-\phi\right)\right|\left(\hat{U}_i^*-U_i^*\right)^2\right]}{t}\\
\leq&\frac{\mathbb{E}\left[\frac{1}{b_F}\left|K'_{b_F}\left(\bar{U}_i-\phi\right)\right|\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)^2\right]}{t}+\frac{\mathbb{E}\left[\frac{1}{b_F}\left|K'_{b_F}\left(\bar{U}_i-\phi\right)\right|\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)^2\right]}{t}\\
\leq&\frac{\|\hat{\varphi}_{(-i)}^{\alpha_n,*}-\varphi^*\|^2(1+o_P(1))}{t},
\end{aligned}$$

where the bound follows by a change of variable and Assumption A.2(iii) and 3.3. We thus directly have that

$$\|A_{12}\|^2\leq\|\hat{\varphi}_{(-i)}^{\alpha_n,*}-\varphi^*\|^4=o_P(1).$$

We focus on the first term

$$\begin{aligned}
A_{11}=&\frac{1}{n}\mathbb{E}\left[\frac{1}{b_F^2}K_{b_F}^2\left(U_i^*-\phi\right)\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)-\hat{r}_{(-i)}(W_i)+r(W_i)\right)^2\right] && (A_{11,a})\\
&+\mathbb{E}\left[\frac{1}{b_F^2}K_{b_F}\left(U_i^*-\phi\right)\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)K_{b_F}\left(U_l^*-\phi\right)\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right] && (A_{11,b})\\
&+\mathbb{E}\left[\frac{1}{b_F^2}K_{b_F}\left(U_i^*-\phi\right)\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)K_{b_F}\left(U_l^*-\phi\right)\left(\hat{r}_{(-l)}(W_l)-r(W_l)\right)\right] && (A_{11,c})\\
&-2\mathbb{E}\left[\frac{1}{b_F^2}K_{b_F}\left(U_i^*-\phi\right)\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)K_{b_F}\left(U_l^*-\phi\right)\left(\hat{r}_{(-l)}(W_l)+r(W_l)\right)\right] && (A_{11,d})
\end{aligned}$$

The first term is easily bound by

$$\int A_{11,a}(\phi)\omega(\phi)d\phi=O_P\left(\frac{1}{nb_F}\left[\|\hat{\varphi}^{\alpha_n,*}-\varphi^*\|^2+\|\hat{r}-r\|^2\right]\right),$$

where the result follows by conditioning on the observables X_i and W_i, a change of variable and Assumption A.2 and A.8.

Similarly, as above, we write

$$\begin{aligned}
A_{11,b}=&\mathbb{E}\left[\frac{1}{b_F^2}\mathbb{E}\left[K_{b_F}\left(U_i^*-\phi\right)|X_i\right]\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)\mathbb{E}\left[K_{b_F}\left(U_l^*-\phi\right)|X_l\right]\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right]\\
=&\mathbb{E}\left[\left(\int K_{b_F}(v_i)f_{U^*|X}(\phi+v_ib_F|X_i)dv_i\right)\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)\right.\\
&\left.\left(\int K_{b_F}(v_l)f_{U^*|X}(\phi+v_lb_F|X_l)dv_l\right)\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right]\\
=&\mathbb{E}\left[\left(f_{U^*|X}(\phi|X_i)+\frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}b_F^\rho\mu_\rho(K)\right)\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)\times\right.\\
&\left.\left(f_{U^*|X}(\phi|X_l)+\frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}b_F^\rho\mu_\rho(K)\right)\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right]\\
\leq&\mathbb{E}\left[f_{U^*|X}(\phi|X_i)\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)f_{U^*|X}(\phi|X_l)\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right]\\
&+2b_F^\rho\mathbb{E}\left[\frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)f_{U^*|X}(\phi|X_l)\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right]\\
&+b_F^{2\rho}\mathbb{E}\left[\frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)\frac{d^\rho f_{U^*|X}(\phi|X_l)}{d\phi^\rho}\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right],
\end{aligned}$$

where the second line follows by the IID property of the data and the last line from Assumption A.2.

For the last term, we directly have

$$\begin{aligned}
&b_F^{2\rho}\left|\mathbb{E}\left[\frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)\frac{d^\rho f_{U^*|X}(\phi|X_l)}{d\phi^\rho}\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right]\right|\\
\leq&b_F^{2\rho}\left|\mathbb{E}\left[\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)\left(\hat{\varphi}_{(-l)}^{\alpha_n,*}(X_l)-\varphi^*(X_l)\right)\right]\right|\\
\leq&b_F^{2\rho}\mathbb{E}\left[\left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i)-\varphi^*(X_i)\right)^2\right]=b_F^{2\rho}\|\hat{\varphi}^{\alpha_n,*}-\varphi^*\|^2=o_P(n^{-1}),
\end{aligned}$$

where the conclusion follows from the restriction on the bandwidth parameter b_F and the result of Theorem 3.3.

We now turn to the two remaining terms. Let us decompose

$$\begin{aligned}
&\hat{\varphi}^{*,\alpha_n} - \varphi^{*,\alpha_n} \\
&= (\alpha_n I + T^*T)^{-1}T^*\left(\hat{r}_{(-i)} - T\varphi^*\right) && (B_1)\\
&+ \left[\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^* - (\alpha_n I + T^*T)^{-1}T^*\right]\left(\hat{r} - T\varphi^*\right) && (B_2)\\
&+ \left[\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^* - (\alpha_n I + T^*T)^{-1}T^*\right]T\varphi^* && (B_3)\\
&+ \varphi^{\alpha_n,*} - \varphi && (B_4)
\end{aligned}$$

Using this decomposition, we write

$$\begin{aligned}
&\mathbb{E}\left[f_{U^*|X}(\phi|X_i)\left(\hat{\varphi}^{\alpha_n,*}_{(-i)}(X_i) - \varphi^*(X_i)\right)f_{U^*|X}(\phi|X_l)\left(\hat{\varphi}^{\alpha_n,*}_{(-l)}(X_l) - \varphi^*(X_l)\right)\right] \\
&= \mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{1,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{1,-l}(X_l)\right] && (R_i)\\
&+ \mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{2,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{2,-l}(X_l)\right] && (R_{ii})\\
&+ \mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{3,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{3,-l}(X_l)\right] && (R_{iii})\\
&+ \mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{4,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{4,-l}(X_l)\right] && (R_{iv})\\
&+ 2\mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{1,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{2,-l}(X_l)\right] && (R_v)\\
&+ 2\mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{1,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{3,-l}(X_l)\right] && (R_{vi})\\
&+ 2\mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{1,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{4,-l}(X_l)\right] && (R_{vii})\\
&+ 2\mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{2,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{3,-l}(X_l)\right] && (R_{viii})\\
&+ 2\mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{2,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{4,-l}(X_l)\right] && (R_{ix})\\
&+ 2\mathbb{E}\left[f_{U^*|X}(\phi|X_i)B_{3,-i}(X_i)f_{U^*|X}(\phi|X_l)B_{4,-l}(X_l)\right] && (R_x)
\end{aligned}$$

We treat all these terms in order using the Cauchy-Schwarz inequality and Assumption 3.3. We have

$$\begin{aligned}
|R_i| \leq& \mathbb{E}\left[f^2_{U^*|X}(\phi|X_i)B^2_{1,-i}(X_i)\right] \\
&= O(1)\|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}T^*\left(\hat{r}_{(-i)} - T\varphi^*\right)\|^2 \\
&\leq \|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}T^*\|^2\|\hat{r}_{(-i)} - T\varphi^*\| \\
&= O_P\left(\alpha_n^{\nu-1}\left(\frac{1}{nh_n^q} + h_n^{2\rho}\right)\right).
\end{aligned}$$

Similarly

$$\begin{aligned}
|R_{ii}| \leq& \mathbb{E}\left[f^2_{U^*|X}(\phi|X_i)B^2_{2,-i}(X_i)\right] \\
&= O_P(1)\|(T^*T)^{\frac{\nu}{2}}\left[\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1} - (\alpha_n I + T^*T)^{-1}\right]\hat{T}^*\left(\hat{r} - T\varphi^*\right)\|^2 \\
&+ \|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}\left(\hat{T}^* - T^*\right)\left(\hat{r} - T\varphi^*\right)\|^2 \\
\leq& \|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}\left(T^*T - \hat{T}^*\hat{T}\right)\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\|^2\|\hat{r} - T\varphi^*\|^2 \\
&+ \|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}\left(\hat{T}^* - T^*\right)\left(\hat{r} - T\varphi^*\right)\|^2 \\
\leq& \|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}T^*\|^2\|\hat{T} - T\|^2\|\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\|^2\|\hat{r} - T\varphi^*\|^2 \\
&+ \|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}\|^2\|\hat{T}^* - T^*\|^2\|\hat{T}\left(\alpha_n I + \hat{T}^*\hat{T}\right)^{-1}\hat{T}^*\|^2\|\hat{r} - T\varphi^*\|^2 \\
&+ \|(T^*T)^{\frac{\nu}{2}}(\alpha_n I + T^*T)^{-1}\left(\hat{T}^* - T^*\right)\left(\hat{r} - T\varphi^*\right)\|^2 \\
&= O_P\left(\frac{1}{\alpha_n^{2-\nu}}\left(\frac{1}{nh_n^{p+q}} + h_n^{2\rho}\right)\left(\frac{1}{nh_n^q} + h_n^{2\rho}\right)\right),
\end{aligned}$$

and

$$|R_{iii}| \leq \|B_3\|^2 = O_P\left(\alpha_n^{\beta-1}\left(\frac{1}{nh_n^{p+q}} + h_n^{2\rho}\right)\right) \text{ and } |R_{iv}| \leq \|B_4\|^2 = O_P(\alpha_n^\beta),$$

where the bounds follow directly from the proof of Theorem 3.3 and Assumption A.8.

We now bound the additional terms as above, using the same line of proof, Assumption A.8, 3.3, and the Cauchy-Schwartz inequality.

$$
\begin{aligned}
|R_{vi}| \leq & \|f_{U^*|X}(\phi|\cdot)B_1\| \|f_{U^*|X}(\phi|\cdot)B_2\| = O_P\left(\frac{1}{\alpha_n^{3/2-\nu}}\left(\frac{1}{\sqrt{nh_n^{p+q}}}+h_n^{\rho}\right)\left(\frac{1}{nh_n^{q}}+h_n^{2\rho}\right)\right)\\
|R_{vii}| \leq & \|f_{U^*|X}(\phi|\cdot)B_1\| \|f_{U^*|X}(\phi|\cdot)B_3\| = O_P\left(\alpha_n^{\frac{\beta+\nu-2}{2}}\left(\frac{1}{\sqrt{nh_n^{p+q}}}+h_n^{\rho}\right)\left(\frac{1}{\sqrt{nh_n^{q}}}+h_n^{\rho}\right)\right)\\
|R_{viii}| \leq & \|f_{U^*|X}(\phi|\cdot)B_1\| \|f_{U^*|X}(\phi|\cdot)B_4\| = O_P\left(\alpha_n^{\frac{\beta+\nu-1}{2}}\left(\frac{1}{\sqrt{nh_n^{p+q}}}+h_n^{\rho}\right)\left(\frac{1}{\sqrt{nh_n^{q}}}+h_n^{\rho}\right)\right)\\
|R_{ix}| \leq & \|f_{U^*|X}(\phi|\cdot)B_2\| \|f_{U^*|X}(\phi|\cdot)B_4\| = O_P\left(\alpha_n^{\frac{\beta+\nu-2}{2}}\left(\frac{1}{\sqrt{nh_n^{p+q}}}+h_n^{\rho}\right)\left(\frac{1}{\sqrt{nh_n^{q}}}+h_n^{\rho}\right)\right)\\
|R_{x}| \leq & \|f_{U^*|X}(\phi|\cdot)B_3\| \|f_{U^*|X}(\phi|\cdot)B_4\| = O_P\left(\alpha_n^{\frac{2\beta-1}{2}}\left(\frac{1}{\sqrt{nh_n^{p+q}}}+h_n^{\rho}\right)\right).
\end{aligned}
$$

The middle term can be treated similarly, and it is bounded as long as $b_F^{\rho}\alpha_n^{1/2} = O_P(1)$, that is $n\alpha_n = O_P(1)$, which is always satisfied.

The term in $A_{11,c}$ can be treated as above, taking expectations conditionally on the instrument W, and by a change of variable and Assumption A.2, one would obtain

$$
\begin{aligned}
A_{11,c} \leq & O(1)\mathbb{E}\left[\left(f_{U^*|W}(\phi|W_i)+b_F^{\rho}\frac{d^{\rho}f_{U^*|W}(\phi|W_i)}{d\phi^{\rho}}\right)\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)\times\right.\\
&\left.\left(f_{U^*|W}(\phi|W_l)+b_F^{\rho}\frac{d^{\rho}f_{U^*|W}(\phi|W_l)}{d\phi^{\rho}}\right)\left(\hat{r}_{(-l)}(W_l)-r(W_l)\right)\right]\\
&= \mathbb{E}\left[f_{U^*|W}(\phi|W_i)\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)f_{U^*|W}(\phi|W_l)\left(\hat{r}_{(-l)}(W_l)-r(W_l)\right)\right]\\
&+\ 2b_F^{\rho}\mathbb{E}\left[\frac{d^{\rho}f_{U^*|W}(\phi|W_i)}{d\phi^{\rho}}\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)f_{U^*|W}(\phi|W_l)\left(\hat{r}_{(-l)}(W_l)-r(W_l)\right)\right]\\
&+\ b_F^{2\rho}\mathbb{E}\left[\frac{d^{\rho}f_{U^*|W}(\phi|W_i)}{d\phi^{\rho}}\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)\frac{d^{\rho}f_{U^*|W}(\phi|W_l)}{d\phi^{\rho}}\left(\hat{r}_{(-l)}(W_l)-r(W_l)\right)\right].
\end{aligned}
$$

By Assumption A.8 and the Cauchy-Schwarz inequality, the last term can be bound as follows

$$
\begin{aligned}
&b_F^{2\rho}\mathbb{E}\left[\frac{d^{\rho}f_{U^*|W}(\phi|W_i)}{d\phi^{\rho}}\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)\frac{d^{\rho}f_{U^*|W}(\phi|W_l)}{d\phi^{\rho}}\left(\hat{r}_{(-l)}(W_l)-r(W_l)\right)\right]\\
&\quad = O_P\left(b_F^{2\rho}\|\hat{r}-r\|^2\right) = o_P(n^{-1}).
\end{aligned}
$$

We know focus on the first term. By Lemma A.2, we have that

$$
\begin{aligned}
\left(\hat{r}_{-i}-T\varphi^*\right)(W_i) = & -e_1'\left(\mathbf{M}_{n,\rho-1}(W_i)\right)^{-1}\left[\frac{\mathbf{H}_n^{-1}}{(n-1)h^q}\sum_{l=1,l\neq i}^{n}\left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)'\ell_1(r(W_l),Y_l)K_h(W_l-W_i)\right]\\
&+\ h_n^{\rho}(D^{\rho}r)(W_i),
\end{aligned}
$$

where to simplify notations below, we have defined

$$
D^{\rho}r = e_1'\left(\mathbf{M}_{\rho-1}\mathbf{H}_n\right)^{-1}\sum_{0\leq|\mathbf{k}|\leq\rho}\sum_{|\mathbf{j}|=\rho}\frac{D^{(\mathbf{j})}r(W_i)}{\mathbf{j}!}M_{\mathbf{k},\mathbf{j}}.
$$

Hence, we write

$$
\begin{aligned}
&\mathbb{E}\left[f_{U^*|W}(\phi|W_i)\left(\hat{r}_{(-i)}(W_i)-r(W_i)\right)f_{U^*|W}(\phi|W_l)\left(\hat{r}_{(-l)}(W_l)-r(W_l)\right)\right]\\
&\leq O(1)\mathbb{E}\left[\left(-e_1'\frac{\left(\mathbf{M}_{n,\rho-1}(W_i)\mathbf{H}_n\right)^{-1}}{(n-1)h^q}\sum_{i'=1,i'\neq i}^{n}\left(\mathbf{W}_{i'i}\mathbf{H}_n^{-1}\right)'\ell_1(r(W_{i'}),Y_{i'})K_{h_n}(W_{i'}-W_i)+h_n^{\rho}(D^{\rho}r)(W_i)\right)\right.
\end{aligned}
$$

$$
\begin{aligned}
&\times \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{l'=1,l'\neq l}^{n} \left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{l'}), Y_{l'}) K_{h_n}(W_{l'} - W_l) + h_n^\rho (D^\rho r)(W_l) \right) \Bigg] \\
&= \mathbb{E}\Bigg[\left(e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_i)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{i'=1,i'\neq i}^{n} \left(\mathbf{W}_{i'i}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{i'}), Y_{i'}) K_{h_n}(W_{i'} - W_i) \right) \times \\
&\left(e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{l'=1,l'\neq l}^{n} \left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{l'}), Y_{l'}) K_{h_n}(W_{l'} - W_l) \right) \Bigg] \\
&- 2h_n^\rho \mathbb{E}\Bigg[\left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_i)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{i'=1,i'\neq i}^{n} \left(\mathbf{W}_{i'i}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{i'}), Y_{i'}) K_{h_n}(W_{i'} - W_i) \right) (D^\rho r)(W_l) \Bigg] \\
&+ h_n^{2\rho} \mathbb{E}[(D^\rho r)(W_i)] \mathbb{E}[(D^\rho r)(W_l)] \\
&= O_P\left(\frac{1}{n} + h_n^{2\rho} \right),
\end{aligned}
$$

where the first line follows from Assumption A.8, and the upper bound is a consequence of Assumption A.6(ii). By the same line of proof, the second term is $o_P(n^{-1})$.

Finally, we consider the term

$$
\begin{aligned}
A_{11,d} &= \mathbb{E}\Bigg[\frac{1}{b_F^2} K_{b_F}\left(U_i^* - \phi\right) \left(\hat{\varphi}_{(-i)}^{\alpha_n,*}(X_i) - \varphi^*(X_i) \right) K_{b_F}\left(U_l^* - \phi\right) \left(\hat{r}_{(-l)}(W_l) + r(W_l) \right) \Bigg] \\
&= \mathbb{E}\Bigg[\frac{1}{b_F^2} K_{b_F}\left(U_i^* - \phi\right) (B_{1,-i}(X_i) + B_{2,-i}(X_i) + B_{3,-i}(X_i) + B_{4,-i}(X_i)) K_{b_F}\left(U_l^* - \phi\right) \times \\
&\left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{l'=1,l'\neq l}^{n} \left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{l'}), Y_{l'}) K_{h_n}(W_{l'} - W_l) + h_n^\rho (D^\rho r)(W_l) \right) \Bigg] \\
&= \sum_{j=1}^{4} \mathbb{E}\Bigg[\frac{1}{b_F^2} K_{b_F}\left(U_i^* - \phi\right) B_{j,-i}(X_i) K_{b_F}\left(U_l^* - \phi\right) \times \\
&\left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{l'=1,l'\neq l}^{n} \left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{l'}), Y_{l'}) K_{h_n}(W_{l'} - W_l) \right) \Bigg] \\
&+ h_n^\rho \sum_{j=1}^{4} \mathbb{E}\Bigg[\frac{1}{b_F^2} K_{b_F}\left(U_i^* - \phi\right) B_{j,-i}(X_i) K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg]
\end{aligned}
$$

We again treat these terms in order, starting with the latter bias component.

$$
\begin{aligned}
&h_n^\rho \left| \mathbb{E}\Bigg[\frac{1}{b_F^2} K_{b_F}\left(U_i^* - \phi\right) (\alpha_n I + T^*T)^{-1} T^* \left(\hat{r}_{(-i)} - T\varphi^* \right)(X_i) K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg] \right| \\
&= h_n^\rho \left| \mathbb{E}\Bigg[\frac{1}{b_F^2} \mathbb{E}\left[K_{b_F}\left(U_i^* - \phi\right) | X_i \right] (\alpha_n I + T^*T)^{-1} T^* \left(\hat{r}_{-i} - T\varphi^* \right)(X_i) K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg] \right| \\
&\leq h_n^\rho \left| \mathbb{E}\Bigg[\left(f_{U^*|X}(\phi|X_i) + b_F^\rho \frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho} \right) (\alpha_n I + T^*T)^{-1} T^* \left(\hat{r}_{-i} - T\varphi^* \right)(X_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg] \right| \\
&\leq h_n^\rho \left| \mathbb{E}\Bigg[f_{U^*|X}(\phi|X_i) (\alpha_n I + T^*T)^{-1} T^* \left(\hat{r}_{-i} - T\varphi^* \right)(X_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg] \right| \\
&\quad + (h_n b_F)^\rho \left| \mathbb{E}\Bigg[\frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho} (\alpha_n I + T^*T)^{-1} T^* \left(\hat{r}_{-i} - T\varphi^* \right)(X_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg] \right| \\
&\leq h_n^\rho \left| \mathbb{E}\Big[\left[(\alpha_n I + TT^*)^{-1} T (T^*T)^{\nu/2} v(\phi, \cdot) \right](W_i) \times \right. \\
&\left. \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_i)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_l), Y_l) K_{h_n}(W_l - W_i) \right) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg] \right| \\
&\quad + h_n^{2\rho} \left| \mathbb{E}\Bigg[\left[(\alpha_n I + TT^*)^{-1} T (T^*T)^{\nu/2} v(\phi, \cdot) \right](W_i) (D^\rho r)(W_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) (D^\rho r)(W_l) \Bigg] \right|
\end{aligned}
$$

$$
\begin{aligned}
&+ \ (h_n b_F)^\rho |\mathbb{E}\Bigg[\Bigg[(\alpha_n I + TT^*)^{-1} T \frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}\Bigg](W_i) \times \\
&\left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q}\left(\mathbf{W}_{li}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_l), Y_l) K_{h_n}(W_l - W_i)\right) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right)(D^\rho r)(W_l)\Bigg]| \\
&+ \ (h_n^2 b_F)^\rho |\mathbb{E}\Bigg[\Bigg[(\alpha_n I + TT^*)^{-1} T \frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}\Bigg](W_i)(D^\rho r)(W_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right)(D^\rho r)(W_l)\Bigg]| \\
&= O_P\left(\frac{h_n^\rho \alpha_n^{\frac{\nu-1}{2}}}{n} + h_n^{2\rho}\alpha_n^{\frac{\nu-1}{2}} + \frac{h_n^\rho b_F^\rho}{n\sqrt{\alpha_n}} + \frac{h_n^{2\rho} b_F^\rho}{\sqrt{\alpha_n}}\right).
\end{aligned}
$$

With a change of variable and using the Cauchy-Schwarz inequality, we can show that the remaining term is bounded by

$$
\begin{aligned}
&h_n^\rho \sum_{j=2}^4 \mathbb{E}\left[\frac{1}{b_F^2} K_{b_F}\left(U_i^* - \phi\right) B_{j,-i}(X_i) K_{b_F}\left(U_l^* - \phi\right)(D^\rho r)(W_l)\right] \\
&= h_n^\rho \sum_{j=2}^4 \mathbb{E}\left[f_{U^*|X}(\phi|X_i) B_{j,-i}(X_i) f_{U^*|X,W}(\phi|X_l, W_l)(D^\rho r)(W_l)\right]\left(1 + O_P(b_F^\rho)\right) \\
&= O_P\left(h_n^\rho \alpha_n^{\frac{\nu-1}{2}}\left(\frac{1}{\sqrt{nh_n^{p+q}}} + h_n^\rho\right)\left(\frac{1}{\sqrt{nh_n^q}} + h_n^\rho\right) + h_n^\rho \alpha_n^{\frac{\beta-1}{2}}\left(\frac{1}{\sqrt{nh_n^{p+q}}} + h_n^\rho\right) + h_n^\rho \alpha_n^{\frac{\beta}{2}}\right).
\end{aligned}
$$

We now turn to the remaining terms.

$$
\begin{aligned}
&|\mathbb{E}\Bigg[\frac{1}{b_F^2} K_{b_F}\left(U_i^* - \phi\right)(\alpha_n I + T^*T)^{-1} T^*\left(\hat{r}_{(-i)} - T\varphi^*\right)(X_i) K_{b_F}\left(U_l^* - \phi\right) \\
&\quad \times \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{l'=1, l'\neq l}^n \left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{l'}), Y_{l'}) K_{h_n}(W_{l'} - W_l)\right)\Bigg]| \\
&\leq |\mathbb{E}\Bigg[\frac{1}{b_F} K_{b_F}\left(U_i^* - \phi\right)(\alpha_n I + T^*T)^{-1} T^*\left(\hat{r}_{(-i)} - T\varphi^*\right)(X_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) \\
&\quad \times \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q}\left(\mathbf{W}_{il}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_i), Y_i) K_{h_n}(W_i - W_l)\right)\Bigg]| \\
&\quad + \ |\mathbb{E}\Bigg[\mathbb{E}\Bigg[\frac{1}{b_F} K_{b_F}\left(U_i^* - \phi\right)|X_i\Bigg](\alpha_n I + T^*T)^{-1} T^*\left(\hat{r}_{(-i)} - T\varphi^*\right)(X_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) \\
&\quad \times \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{l'=1, l'\neq l, l'\neq i}^n \left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{l'}), Y_{l'}) K_{h_n}(W_{l'} - W_l)\right)\Bigg]| \\
&\quad = |\mathbb{E}\Bigg[\frac{1}{b_F} K_{b_F}\left(U_i^* - \phi\right)(\alpha_n I + T^*T)^{-1} T^*\left(\hat{r}_{(-i)} - T\varphi^*\right)(X_i) f_{U^*|W}(\phi|W_l) \\
&\quad \times \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q}\left(\mathbf{W}_{il}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_i), Y_i) K_{h_n}(W_i - W_l)\right)\Bigg]|(1 + O_P(b_F^\rho)) \\
&\quad + \ |\mathbb{E}\Bigg[\left(f_{U^*|X}(\phi|X_i) + b_F^\rho \frac{d^\rho f_{U^*|X}(\phi|X_i)}{d\phi^\rho}\right)(\alpha_n I + T^*T)^{-1} T^*\left(\hat{r}_{(-i)} - T\varphi^*\right)(X_i) \frac{1}{b_F} K_{b_F}\left(U_l^* - \phi\right) \\
&\quad \times \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q} \sum_{l'=1, l'\neq l, l'\neq i}^n \left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)' \ell_1(r(W_{l'}), Y_{l'}) K_{h_n}(W_{l'} - W_l)\right)\Bigg]| \\
&\leq |\mathbb{E}\Bigg[(\alpha_n I + T^*T)^{-1} T^*\left(\hat{r}_{(-i)} - T\varphi^*\right)(X_i) f_{U^*|W}(\phi|W_l) \times \left(-e_1' \frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q}\left(\mathbf{W}_{il}\mathbf{H}_n^{-1}\right)'\right. \\
&\quad \left.\mathbb{E}\Bigg[\frac{1}{b_F} K_{b_F}\left(U_i^* - \phi\right)\ell_1(r(W_i), Y_i)|X_i, W_i\Bigg] K_{h_n}(W_i - W_l)\right)\Bigg]|(1 + O_P(b_F^\rho))
\end{aligned}
$$

$$
\begin{aligned}
&+ \left|\mathbb{E}\left[\left[(\alpha_n I + T^*T)^{-1} T (T^*T)^{\frac{\nu}{2}} \nu(\phi,\cdot)\right](W_i)\left(\hat{r}_{(-i)} - T\varphi^*\right)(W_i)\frac{1}{b_F}K_{b_F}\left(U_l^* - \phi\right)\right.\right.\\
&\times \left.\left.\left(-e_1'\frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q}\sum_{l'=1,l'\neq l,l'\neq i}^{n}\left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)'\ell_1(r(W_{l'}),Y_{l'})K_{h_n}(W_{l'}-W_l)\right)\right]\right|\\
&+ b_F^{\rho}\left|\mathbb{E}\left[\left[(\alpha_n I + T^*T)^{-1} T \frac{d^{\rho} f_{U^*|X}(\phi|\cdot)}{d\phi^{\rho}}\right](W_i)\left(\hat{r}_{(-i)} - T\varphi^*\right)(W_i)\frac{1}{b_F}K_{b_F}\left(U_l^* - \phi\right)\right.\right.\\
&\times \left.\left.\left(-e_1'\frac{\left(\mathbf{M}_{n,\rho-1}(W_l)\mathbf{H}_n\right)^{-1}}{(n-1)h^q}\sum_{l'=1,l'\neq l,l'\neq i}^{n}\left(\mathbf{W}_{l'l}\mathbf{H}_n^{-1}\right)'\ell_1(r(W_{l'}),Y_{l'})K_{h_n}(W_{l'}-W_l)\right)\right]\right|\\
&= O_P\left(\frac{1}{nh_n^q\alpha_n^{1/2}}\left(\frac{1}{\sqrt{n}}+h_n^{\rho}\right)+\frac{\alpha_n^{\frac{\nu-1}{2}}}{n}+\frac{b_F^{\rho}}{n\sqrt{\alpha_n}}\right),
\end{aligned}
$$

where the bound on the first term comes from a change of variable, Assumption A.2(iii), A.6(iii), and the Cauchy-Schwarz inequality; while the bound on the second term comes from the decomposition in Lemma A.2, Assumption A.6(ii) and the Cauchy-Schwarz inequality. The conclusion holds given the restrictions on the tuning parameters given in the statement of the Theorem. This concludes the proof.

A8. Proof of Corollary 3.8

Let us use the same notations we have employed in Theorems 3.3 and 3.7. From Theorem 3.3 we know that, for m sufficiently large

$$
\|\hat{\varphi}^{\alpha_n} - \varphi\|^2 = O_P\left[\frac{1}{\alpha_n}\left(\frac{1}{nh_n^q}+h_n^{2\rho}\right)+\left(\frac{1}{nh_n^{p+q}}+h_n^{2\rho}\right)\alpha_n^{\beta-1}+\alpha_n^{\beta}\right],
$$

which, for $\beta > 2$, and $p \leq q$, becomes

$$
\|\hat{\varphi}^{\alpha_n} - \varphi\|^2 = O_P\left[\frac{1}{\alpha_n}\left(\frac{1}{nh_n^q}+h_n^{2\rho}\right)+\alpha_n^{\beta}\right],
$$

as the first term dominates the middle term. From Theorem 3.7, we have obtained that,

$$
\begin{aligned}
\|A_{11,a}\|^2 &= O_P\left(n^{-\frac{2\rho-1}{2\rho}}\left(\frac{1}{\alpha_n}\left(\frac{1}{nh_n^q}+h_n^{2\rho}\right)+\alpha_n^{\beta}\right)\right),\\
\|A_{11,b}\|^2 &= O_P\left(\alpha_n^{\nu-1}\left(\frac{1}{nh_n^q}+h_n^{2\rho}\right)+\frac{1}{\alpha_n^{2-\nu}}\left(\frac{1}{nh_n^{p+q}}+h_n^{2\rho}\right)\left(\frac{1}{nh_n^q}+h_n^{2\rho}\right)+\alpha_n^{\beta-1}\left(\frac{1}{nh_n^{p+q}}+h_n^{2\rho}\right)+\alpha_n^{\beta}\right)\\
\|A_{11,c}\|^2 &= O_P\left(\frac{1}{n}+h_n^{2\rho}\right)\\
\|A_{11,d}\|^2 &= O_P\left(\frac{h_n^{\rho}\alpha_n^{\frac{\nu-1}{2}}}{n}+h_n^{2\rho}\alpha_n^{\frac{\nu-1}{2}}+\frac{h_n^{\rho}b_F^{\rho}}{n\sqrt{\alpha_n}}+\frac{h_n^{2\rho}b_F^{\rho}}{\sqrt{\alpha_n}}+\frac{1}{nh_n^q\alpha_n^{1/2}}\left(\frac{1}{\sqrt{n}}+h_n^{\rho}\right)+\frac{\alpha_n^{\frac{\nu-1}{2}}}{n}+\frac{b_F^{\rho}}{n\sqrt{\alpha_n}}\right)\\
\|A_{12}\|^2 &= O_P\left(\frac{1}{n^2\alpha_n^2h_n^{2q}}+\frac{h_n^{4\rho}}{\alpha_n^2}\right)+o_P(n^{-1})
\end{aligned}
$$

Directly from the restrictions on ρ, and the rates for α_n, and h_n, we obtain that

$$
\|A_{12}\|^2 = O_P(n^{-1}), \quad \|A_{11,a}\|^2 = O_P(n^{-1}), \text{ and } \|A_{11,c}\|^2 = O_P(n^{-1}).
$$

Similarly, as $b_F^{\rho}/\sqrt{\alpha_n} = o(1)$, with $h_n \asymp n^{-1/(2\rho)}$, $n\alpha_n h_n^{2q} = o(1)$, and $\nu > 1$, by Assumption 3.3, then $\|A_{11,d}\|^2 = o_P(n^{-1})$.

References

Ahn, H., Ichimura, H., Powell, J., 2004. Simple Estimators for Monotone Index Models. Manuscript. Department of Economics, UC Berkeley.

Babii, A., Florens, J.-P., 2017. Distribution of Residuals in the Nonparametric IV model with Application to Separability Testing. Mimeo -UNC Chapel Hill.

Babii, A., Florens, J.-P., 2017. Is Completeness Necessary? Estimation in Non-identified Linear Models. Mimeo -UNC Chapel Hill.

Blundell, R., Kristensen, D., Matzkin, R.L., 2013. Control Functions and Simultaneous Equations Methods. American Economic Review 103 (3). 563–69

Blundell, R.W., Powell, J.L., 2004. Endogeneity in Semiparametric Binary Response Models. Review of Economic Studies 71, 655–679.

Carrasco, M., Florens, J.-P., 2011. A Spectral Method for Deconvolving a Density. Econometric Theory 27, 546–581.

Carrasco, M., Florens, J.-P., Renault, E., 2007. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics. Elsevier, pp. 5633–5751.

Carrasco, M., Florens, J.-P., Renault, E., 2013. Asymptotic Normal Inference in Linear Inverse Problems. In: Racine, J.S., Ullah, A., Su, L. (Eds.), Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics.

Centorrino, S., 2016. Data-Driven Selection of the Regularization Parameter in Additive Nonparametric Instrumental Regressions. Mimeo - Stony Brook University.
Centorrino, S., Fève, F., Florens, J.-P., 2017. Additive Nonparametric Instrumental Regressions: a Guide to Implementation. Journal of Econometric Methods 6 (1).
Chen, X., Christensen, T., 2018. Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression. Quantitative Economics 9 (1), 39–84.
Chen, X., Pouzo, D., 2012. Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals. Econometrica 80 (1), 277–321.
Chen, X., Pouzo, D., 2015. Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models. Econometrica 83 (3), 1013–1079.
Chen, X., Reiss, M., 2011. On Rate Optimality for Ill-Posed Inverse Problems in Econometrics. Econometric Theory 27 (3), 497–521.
Claeskens, G., Van Keilegom, I., 2003. Bootstrap confidence bands for regression curves and their derivatives. The Annals of Statistics 31 (6), 1852–1884.
Cosslett, S.R., 1983. Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model. Econometrica 51 (3), 765–782.
Darolles, S., Fan, Y., Florens, J.P., Renault, E., 2011. Nonparametric Instrumental Regression. Econometrica 79 (5), 1541–1565.
Dong, Y., 2010. Endogenous Regressor Binary Choice Models without Instruments, with an Application to Migration. Economics Letters 107 (1), 33–35.
Dong, Y., Lewbel, A., 2015. A Simple Estimator for Binary Choice Models with Endogenous Regressors. Econometric Reviews 34 (1-2), 82–105.
Engl, H.W., Hanke, M., Neubauer, A., 2000. Regularization of Inverse Problems. Mathematics and Its Applications, 375. Kluwer Academic Publishers, Dordrecht.
Fan, J., Farmen, M., Gijbels, I., 1998. Local Maximum Likelihood Estimation and Inference. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 60 (3), pp.591–608.
Fan, J., Gijbels, I., King, M., 1997. Local likelihood and local partial likelihood in hazard regression. The Annals of Statistics 25 (4), 1661–1690.
Fan, J., Heckman, N.E., Wand, M.P., 1995. Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. Journal of the American Statistical Association 90 (429), pp.141–150.
Fève, F., Florens, J.-P., 2010. The Practice of Non-parametric Estimation by Solving Inverse Problems: the Example of Transformation Models. Econometrics Journal 13 (3), S1–S27.
Florens, J.-P., Johannes, J., Van Bellegem, S., 2011. Identification and Estimation by Penalization in Nonparametric Instrumental Regression. Econometric Theory 27 (3), 472–496.
Florens, J.-P., Racine, J., Centorrino, S., 2018. Nonparametric Instrumental Variable Derivative Estimation. Jounal of Nonparametric Statistics 30 (2), 368–391.
Florens, J.-P., Simoni, A., 2012. Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior. Journal of Econometrics 170 (2), 458–475.
Freyberger, J., 2017. On Completeness and Consistency in Nonparametric Instrumental Variable Models. Econometrica 85 (5), 1629–1644.
Frölich, M., 2006. Non-parametric Regression for Binary Dependent Variables. Econometrics Journal 9 (3), 511–540.
Gagliardini, P., Scaillet, O., 2012. Tikhonov Regularization for Nonparametric Instrumental Variable Estimators. Journal of Econometrics 167 (1), 61–75.
Gozalo, P., Linton, O., 2000. Local nonlinear least squares: Using parametric information in nonparametric regression. Journal of Econometrics 99 (1), 63–106.
Gu, J., Li, Q., Yang, J.-C., 2015. Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distribution. Econometric Reviews 34 (6-10), 979–1010.
Hall, P., Horowitz, J.L., 2005. Nonparametric Methods for Inference in the Presence of Instrumental Variables. Annals of Statistics 33 (6), 2904–2929.
Horowitz, J.L., 1992. A Smoothed Maximum Score Estimator for the Binary Response Model. Econometrica 60 (3). 505–31
Horowitz, J.L., 2007. Asymptotic Normality of a Nonparametric Instrumental Variable Estimator. International Economic Review 48 (4), 1329–1349.
Horowitz, J.L., 2011. Applied nonparametric instrumental variables estimation. Econometrica 79 (2), 347–394.
Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics 58 (1–2), 71–120.
Imbens, G.W., Newey, W.K., 2009. Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. Econometrica 77 (5), 1481–1512.
Jacho-Chávez, D., Lewbel, A., Linton, O., 2010. Identification and nonparametric estimation of a transformed additively separable model. Journal of Econometrics 156 (2), 392–407.
Johannes, J., Van Bellegem, S., Vanhems, A., 2013. Iterative Regularization in Nonparametric Instrumental Regression. Journal of Statistical Planning and Inference 143 (1), 24–39.
Klein, R.W., Spady, R.H., 1993. An Efficient Semiparametric Estimator for Binary Response Models. Econometrica 61 (2), 387–421.
Lewbel, A., 1998. Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors. Econometrica 66 (1), 105–121.
Lewbel, A., 2000. Semiparametric Qualitative Response Model Estimation with unknown Heteroscedasticity or Instrumental Variables. Journal of Econometrics 97 (1), 145–177.
Lewbel, A., 2007. A local generalized method of moments estimator. Economics Letters 94 (1), 124–128.
Lewbel, A., Linton, O., 2007. Nonparametric matching and efficient estimators of homothetically separable functions. Econometrica 75 (4), 1209–1227.
Mammen, E., Rothe, C., Schienle, M., 2012. Nonparametric Regression with Nonparametrically Generated Covariates. Annals of Statistics 40 (2), 1132–1170.
Manski, C.F., 1985. Semiparametric Analysis of Discrete Response : Asymptotic Properties of the Maximum Score Estimator. Journal of Econometrics 27 (3), 313–333.
Manski, C.F., 1988. Identification of Binary Response Models. Journal of the American Statistical Association 83 (403), pp.729–738.
Masry, E., 1996. Multivariate Local Polynomial Regression for Time Series: Uniform strong Consistency and Rates. Journal of Time Series Analysis 17 (6), 571–599.
Matzkin, R.L., 1992. Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models. Econometrica 60 (2). 239–70
Newey, W.K., Powell, J.L., 2003. Instrumental Variable Estimation of Nonparametric Models. Econometrica 71 (5), 1565–1578.
Newey, W.K., Powell, J.L., Vella, F., 1999. Nonparametric Estimation of Triangular Simultaneous Equations Models. Econometrica 67 (3), 565–603.
Rivers, D., Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. Journal of Econometrics 39 (3), 347–366.
Rothe, C., 2009. Semiparametric Estimation of Binary Response Models with Endogenous Regressors. Journal of Econometrics 153 (1), 51–64.
Santos, A., 2011. Instrumental Variable Methods for Recovering Continuous Linear Functionals. Journal of Econometrics 161 (2), 129–146.
Severini, T.A., Tripathi, G., 2012. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. Journal of Econometrics 170 (2), 491–498.
Signorini, D.F., Jones, M.C., 2004. Kernel Estimators for Univariate Binary Regression. Journal of the American Statistical Association 99 (465), 119–126.
Tibshirani, R., Hastie, T., 1987. Local Likelihood Estimation. Journal of the American Statistical Association 82 (398), pp.559–567.
van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics. Springer.
Zhao, P.-L., 1994. Asymptotics of Kernel Estimators based on Local Maximum Likelihood. Journal of Nonparametric Statistics 4 (1), 79–90.