Uniform Convergence of Weighted Sums of Non- and Semi-parametric Residuals for Estimation and Testing*

Juan Carlos Escanciano† Indiana University

David T. Jacho-Chávez‡ Emory University

Arthur Lewbel§ Boston College

Abstract

A new uniform expansion is introduced for sums of weighted kernel-based regression residuals from nonparametric or semiparametric models. This expansion is useful for deriving asymptotic properties of semiparametric estimators and test statistics with data-dependent bandwidths, random trimming, and estimated efficiency weights. Provided examples include a new estimator for a binary choice model with selection and an associated directional test for specification of this model's average structural function. An appendix contains new results on uniform rates for kernel estimators and primitive sufficient conditions for high level assumptions commonly used in semiparametric estimation.

Keywords: Semiparametric regression; Semiparametric residuals; Nonparametric residuals; Uniform-in-bandwidth; Sample selection models; Empirical process theory; Limited dependent variables.

JEL classification: C13; C14; C21; D24

*We would like to thank the Co-Editor, Peter Robinson, the Associate Editor, and two anonymous referees for various helpful suggestions and corrections that greatly improved the readability of the paper. We also thank Xiaohong Chen, Hidehiko Ichimura, Simon Lee, Enno Mammen, Jim Powell, Jean Marc Robin, Christoph Rothe, Ingrid van Keilegom, Adonis Yatchew and participants of many conferences and seminars for helpful suggestions. All remaining errors are our own.

†Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/. Research funded by the Spanish Plan Nacional de I+D+I, reference number SEJ2007-62908.

‡Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: djachocha@emory.edu. Web Page: http://userwww.service.emory.edu/~djachoc/.

§Corresponding Author: Department of Economics, Boston College, 140 Commonwealth Avenue, Chesnut Hill, MA 02467, USA. E-mail: lewbel@bc.edu. Web Page: http://www2.bc.edu/~lewbel/.

1 Introduction

This paper provides a new uniform expansion for a sum of weighted kernel-based regression residuals from nonparametric or semiparametric models, which has a variety of applications in semiparametric estimation and testing. Consider an independent and identically distributed (iid) data set $\{Y_i, X_i^\top\}_{i=1}^n$ drawn from the joint distribution of the vector-valued random variable $(Y, X^\top)$, where $A^\top$ denotes the transpose of A. Let $W(X)$ denote a vector of measurable functions of X, let $\widehat{m}(\cdot|W)$ be a Nadaraya-Watson (NW) kernel estimator of $E[Y|W(X) = \cdot]$ using a possibly data dependent bandwidth $\widehat{h}_n$, let $\widehat{t}_{ni}$ be a data dependent trimming function, and let $\phi(X)$ be a measurable function of X with $E[\phi^2(X)] < \infty$.

We first supply a general representation of the empirical process

$$\hat{\Delta}_n(W, \phi) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W(X_i)|W)\} \widehat{t}_{ni} \phi(X_i) \tag{1}$$

that is uniform in the bandwidth $\widehat{h}_n$, and uniform in both W and ϕ. In addition, for the case where $E[Y|X] = E[Y|W(X)]$ almost surely (a.s.), we provide a uniform representation of the process

$$\hat{\Delta}_n(\widehat{W}, \widehat{\phi}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(\widehat{W}(X_i)|\widehat{W})\} \widehat{t}_{ni} \widehat{\phi}(X_i), \tag{2}$$

which accounts for the effects of estimation errors in $\widehat{W}$ and $\widehat{\phi}$. We show that estimation error in $\widehat{W}$ only affects the limiting distribution through the first argument of $\widehat{m}$, which simplifies derivation of limiting distributions in the presence of complications like data dependent bandwidths. We also show that estimation error in $\widehat{\phi}$ will have no effect on the limiting distribution of $\hat{\Delta}_n(\widehat{W}, \widehat{\phi})$.

We use these uniform representations to show that the asymptotic properties of estimators and tests that might otherwise have been obtained by more standard methods (such as by U-statistics or by empirical processes theory, often with high-level, difficult-to-verify assumptions), can hold under more general conditions, with more easily verified assumptions, than previously recognized. These general conditions include allowing for data dependent bandwidths and random trimming. In addition, we show that $\hat{\Delta}_n$ can be used to develop inference for some other objects of interest, such as semiparametric test statistics.

Standard kernel-based techniques assume a bandwidth h_n that is a deterministic function of the sample size n, and that satisfies conditions such as $a_n \leq h_n \leq b_n$ as $n \to \infty$ for some pre-specified deterministic sequences a_n and b_n. In contrast, our method of proof allow for the smoothing parameter to be any data dependent (and hence random) sequence $\widehat{h}_n$ constrained only by assuming that the probability that $a_n \leq \widehat{h}_n \leq b_n$ goes to one as $n \to \infty$. As a result, our theoretical developments can be used to establish limiting distributions that allows for virtually any suitably rate-constrained data dependent algorithm one might devise for bandwidth selection. This is useful because standard U-processes theory can suffer from technical difficulties when using stochastic bandwidths (see, e.g., Remark 2.13, in Neumeyer, 2004, p. 81). A byproduct of this approach is that one can easily implement data-driven bandwidth selection procedures, both for estimators and test statistics, that are consistent

with their asymptotic properties by construction. We illustrate this with the example of a bandwidth where the rate is set by theory, and the constant is estimated by minimizing the same objective function used to estimate other parameters in a semiparametric model.

Our theorems similarly allow for general data dependent trimming $\widehat{t}_{ni}$, and data dependent observation weights $\widehat{\phi}(X_i)$ that one might construct to improve efficiency of estimators for example. Standard methods for dealing with estimated nuisance functions in semiparametric models entail U-statistic expansions, and require $n^{-1/4}$-rates as in Newey and McFadden (1994), or require functional derivative techniques like in Chen, Linton, and van Keilegom (2003). In contrast, our results employ stochastic equicontinuity arguments, and standard bias calculations. We show that whenever certain equicontinuity conditions hold, rate restrictions on nuisance functions are not needed and estimation error does not affect the limiting distributions. An example of where this occurs is in the efficiency weights $\widehat{\phi}(X_i)$.[1] This is useful because efficiency weights can involve estimated variance calculations including nonparametric derivatives. These components may cause $\widehat{\phi}(X_i)$ to converge slowly or have a convergence rate that is difficult to establish. Similar conditions on rates for estimated weights could have been obtained by exploiting the fact that $\widehat{\phi}$ enters linearly in the sample mean, as suggested in Newey (1994) and Chen (2007). However, random trimming implies a fundamental discontinuity in estimated nuisance parameters that would make the application of functional derivative techniques difficult, but which is readily amenable to our stochastic equicontinuity arguments.

To illustrate how our results may be used for both estimation and testing, we apply them to a semiparametric binary threshold crossing model with sample selection. Let D be a binary variable that indicates if an individual is selected, and define $g_0(X) = E[D|X]$. Then $D = \mathbb{I}[g_0(X) - u \geq 0]$ where u is uniform on $[0,1]$, independent of X, and $\mathbb{I}(\cdot)$ represents the indicator function that equals one if its argument is true and zero otherwise. Let a binary outcome Y be given by $Y = \mathbb{I}\left(X^\top \theta_0 - e \geq 0\right) D$, so an individual who is not selected has $Y = D = 0$, while selected individuals have $D = 1$ and choose an outcome Y to be either zero or one based on a threshold crossing model. The function $g_0(X) = E[D|X]$ and the distribution of the errors e given u are nonparametric, with the motivation that economic theory drives model specification for the outcome Y, but relatively less is typically known about the selection mechanism. We propose a semiparametric maximum likelihood estimator $\widehat{\theta}$ for θ_0 similar to Klein and Spady (1993) or Klein, Shen and Vella (2012), but with a nonparametric generated regressor estimated in a first stage and with estimated weights to improve efficiency.

The corresponding asymptotic limiting distribution for this estimator could have been obtained by more standard U-statistic based methods such as Newey and McFadden (1994), or by functional derivative techniques like Chen, Linton, and van Keilegom (2003), but we obtain this same limiting distribution under alternative, more general, conditions that can be verified using simpler arguments than in typical applications of these methods. In particular, we obtain the $\sqrt{n}$-limit normal distribution for $\widehat{\theta}$ allowing for general forms of estimated weights, data dependent asymptotic trimming, and data dependent bandwidth choice.

[1] In contrast to $\widehat{\phi}$, in our example applications these equicontinuity conditions do not hold for the nuisance functions that appear in $\widehat{W}$, and so those nuisance functions will require $n^{-1/4}$-rate and expansion arguments. However, our equicontinuity results will still be useful for dealing with data dependent bandwidths in $\widehat{m}$ as a function of $\widehat{W}$.

For a second application, we construct a directional test for the correct specification of a policy parameter in the above model. More precisely, we consider a researcher who is concerned about misspecification of the semiparametric model only to the extent that the misspecification may lead to inconsistent estimates of an average structural function (ASF) parameter. We show how a directional test can be developed for this situation using our uniform expansions. These expansions permit the use of a data-driven bandwidth choice procedure that leads to a test with better power properties than alternatives that use bandwidths chosen for estimation rather than for testing. In this example, the typical approach to avoid random denominators in testing, as used in, e.g., Delgado and González Manteiga (2001), cannot be applied. As with our estimation examples, it would be difficult to establish the asymptotic properties of our test, including data dependent bandwidth choice, using methods other than ours that are not uniform in the bandwidth.

The paper is organized as follows: After this introduction and a short literature review, Section 3 provides our main uniform expansion results, including the extension allowing for generated regressors. Section 4 illustrates the utility of these results by applying them to the new estimator and new test statistic for the binary threshold crossing model with sample selection described above. Section 5 concludes, and the main proofs of Sections 3 and 4 are gathered into an Appendix A.

Appendix B for this paper contains new results on uniform rates of convergence of kernel estimators. These rate results extend previous important work by Einmahl and Mason (2005), and can be used to provide uniform-in-bandwidth inference for general two-step estimators with kernel nuisance estimates. Appendix C contains examples of primitive conditions that suffice to satisfy some high level assumptions. A supplemental appendix to this paper offers more example applications of our results, including a description of how they may be generically applied to derive the asymptotic properties of semiparametric estimators such as Ichimura (1993), Klein and Spady (1993) and Rothe (2009), while allowing for data-driven bandwidths, data-driven asymptotic trimming, and estimated weights.

2 Literature Review

Our equation (1) has the form of typical terms that show up in expansions of semiparametric estimators and test statistics. For example, by defining ϕ accordingly, if $W(X) = (X^\top\theta_1, \ldots, X^\top\theta_J)$ for a collection of J-finite dimensional unknown parameters $\theta_1, \ldots, \theta_J$, $\hat{\Delta}_n$ could be the first order conditions for a semiparametric weighted least squares estimator of index parameters as in Ichimura and Lee (1991) or when $J = 1$, $\hat{\Delta}_n$ could be the first order conditions for semiparametric weighted least squares or maximum likelihood estimators as those in Ichimura (1993) and Klein and Spady (1993), respectively. Similarly, if $X := (X_1^\top, X_2^\top, Z_1^\top, Z_2^\top)^\top$ and $W(X) = (Z_1^\top\theta_1 + X_2^\top\theta_2, X_2 - g(Z_1, Z_2))$, then $\hat{\Delta}_n$ could be the first order conditions for semiparametric weighted least squares or maximum likelihood estimators that uses 'control function' approaches as in Escanciano, Jacho-Chávez and Lewbel (2012) and Rothe (2009), respectively. Alternatively, if $W(X) = X_1 \subset X$ and $\phi(X) = \mathbb{I}[X \leq x]$, $x \in \mathbb{R}^p$, $\hat{\Delta}_n$ also has the form of test statistics designed to test nonparametrically the significance of a subset of covariates as in Delgado and González Manteiga (2001).

When $\widehat{W}$ replaces W in (1), we have a generated regressors model, as (parametrically) described by

Pagan (1984). Semiparametric models with generated regressors abound in the literature and include Ichimura and Lee (1991), Ichimura (1993), Ahn and Powell (1993), Ahn and Manski (1993), Olley and Pakes (1996), Ahn (1997), Heckman, Ichimura and Todd (1998), Newey, Powell and Vella (1999), Pinkse (2001), Li and Wooldridge (2002), Das, Newey, and Vella (2003), Blundell and Powell (2004), Heckman and Vytlacil (2005), Lewbel and Linton (2007), Imbens and Newey (2009), Rothe (2009), and Mammen, Rothe and Schienle (2013), among others. The asymptotic variance of general estimators within this class of models is studied by Hahn and Ridder (2013). Analyses of the properties of generic nonparametric two step estimators with nonparametric generated regressors, include Andrews (1995), Song (2008), Sperlich (2009), and Mammen, Rothe and Schienle (2012).

We contribute to these literatures in several ways. First, we derive our results allowing for stochastic bandwidths. Second, we show how straightforward stochastic equicontinuity arguments can be used to derive the impact of generated regressors on inference. Third, we propose a unified method for inference in semiparametric models with generated regressors, including estimation, testing, and bandwidth choice. Fourth, we contribute to the literature on nonparametric two step estimation with nonparametric generated regressors by providing uniform rates for kernel estimators that are uniform in the bandwidth. In particular, the Appendix C to this paper shows how our new results can be used to prove that, under simple primitive conditions, the infinite-dimensional nuisance parameter estimator belongs to a certain class of smooth functions with probability tending to one. This then provides primitive conditions for a high level assumption that is routinely employed in the semiparametric estimation literature (see e.g, Chen, Linton, and van Keilegom, 2003 and Ichimura and Lee, 2010).

Works devoted to estimation of general semiparametric models include Bickel, Klaassen, Ritov and Wellner (1993), Andrews (1994), Newey (1994), Newey and McFadden (1994), Ai and Chen (2003), Chen, Linton, and van Keilegom (2003), Chen (2007), Ichimura and Lee (2010) and references therein. These works aim at very general models, and applications in specific settings might require considerable work, e.g. an investigation of pathwise functional derivatives (often up to a second order) for implicitly defined nuisance parameters. Our uniform representation of the process $\hat{\Delta}_n(\widehat{W}, \widehat{\phi})$ accounts for the estimation effects of $\widehat{W}$ and $\widehat{\phi}$ without requiring all of the machinery involved in the explicit calculation of pathwise functional derivatives. This is made possible here using an approach based on stochastic equicontinuity arguments. Andrews (1994) also used stochastic equicontinuity for estimating semiparametric models, but he relies on an asymptotic orthogonality condition that does not always hold in our setting.

Related to our derivation is work on nonparametric and semiparametric estimation with possibly parametric or nonparametric generated covariates. In particular, Mammen, Rothe and Schienle (2012, 2013) study these problems using kernel estimators, and characterize the asymptotic contribution of generated regressors to the pointwise distribution of their local linear estimator, as well as to the distribution of optimization estimators. Unlike Mammen, Rothe and Schienle (2012, 2013), our results permit data dependent bandwidths and random trimming.

Also related is a recent paper by Li and Li (2010) which provides sufficient conditions for the first-order asymptotic properties of a larger class of kernel-based semiparametric estimators and test statistics to hold with data dependent bandwidths. Their method of proof requires one to use an

estimated bandwidth first with a 'rule-of-thumb' asymptotic representation, i.e. a constant term times a known power of the sample size, and then establish the stochastic equicontinuity of these generic estimators and test statistics with respect to this constant term. Our development does not require this last step. Instead, our results are shown to hold uniformly over sets of admissible bandwidths which include estimated bandwidths with 'rule-of-thumb' asymptotic representations as a special case. Essentially, our results show that any data dependent procedure one might propose will yield the same asymptotic limiting distribution, as long as the procedure shrinks the bandwidth within a certain range of rates.

3 A Uniform Expansion

Let $\mathcal{Z}_n := \{Y_i, X_i^\top\}_{i=1}^n$ represent a sample of size n from the joint distribution of $(Y, X^\top)$ taking values in $\mathcal{X}_Y \times \mathcal{X}_X \subset \mathbb{R}^{1+p}$. We assume $(Y, X^\top)$ is independent of $\mathcal{Z}_n$. Let $(\Omega, \mathcal{F}, P)$ be the probability space in which all the variables of this paper are defined. Henceforth, $\mathcal{X}_\xi$ denotes the support of the generic random vector ξ. Let $\mathcal{W}$ be a class of measurable functions of X with values in $\mathbb{R}^d$, and let $f(w|W)$ denote the Lebesgue density of $W(X)$ evaluated at $w \in \mathcal{X}_W \equiv \mathcal{X}_{W(X)}$. Define $\mathcal{Q}_\mathcal{W} := \{W(x) \in \mathbb{R}^d : W \in \mathcal{W} \text{ and } x \in \mathcal{X}_X\}$. We assume that $E|Y| < \infty$, so that the regression function

$$m(w|W) := E[Y|W(X) = w], \qquad w \in \mathcal{X}_W \subset \mathbb{R}^d,$$

is well defined a.s., for each $W \in \mathcal{W}$. Under standard regularity conditions, the function $m(w|W)$ can be consistently estimated by the nonparametric NW kernel estimator

$$\begin{aligned}
\widehat{m}(w|W) &:= \widehat{T}(w|W)/\widehat{f}(w|W), \\
\widehat{T}(w|W) &:= \frac{1}{n}\sum_{i=1}^n Y_i K_{\widehat{h}_n}(w - W(X_i)), \\
\widehat{f}(w|W) &:= \frac{1}{n}\sum_{i=1}^n K_{\widehat{h}_n}(w - W(X_i)),
\end{aligned}$$

where $K_h(w) = \prod_{l=1}^d k_h(w_l)$, $k_h(w_l) = h^{-1}k(w_l/h)$, $k(\cdot)$ is a kernel function, $w = (w_1, \ldots, w_d)^\top$ and $\widehat{h}_n$ denotes a possibly data dependent bandwidth parameter satisfying regularity conditions described in Assumption 5 below. Appendix B provides sufficient conditions for the uniform (in w, W and $\widehat{h}_n$) consistency of $\widehat{m}$ and related quantities. These new uniform-in-bandwidth convergence results should be of some independent interest.

Let $f_X(x|w, W)$ be the density, with respect to a σ-finite measure $\mu_W(\cdot)$, of X conditional on $W(X) = w$, and evaluated at $x \in \mathcal{X}_X$. Note that X need not be absolutely continuous as we do not require $\mu_W(\cdot)$ to be the Lebesgue measure. To measure the complexity of the class $\mathcal{W}$, we employ bracketing numbers. For a measurable class of functions $\mathcal{G}$ from $\mathbb{R}^p$ to $\mathbb{R}$, let $\|\cdot\|$ be a generic pseudo-norm on $\mathcal{G}$, defined as a norm except for the property that $\|f\| = 0$ does not necessarily imply that $f \equiv 0$. Given two functions l, u, a bracket $[l, u]$ is the set of functions $f \in \mathcal{G}$ such that $l \leq f \leq u$. An ε-bracket with respect to $\|\cdot\|$ is a bracket $[l, u]$ with $\|l - u\| \leq \varepsilon$, $\|l\| < \infty$ and $\|u\| < \infty$ (note

that u and l not need to be in $\mathcal{G}$). The *covering number with bracketing* $N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|)$ is the minimal number of ε-brackets with respect to $\|\cdot\|$ needed to cover $\mathcal{G}$. These definitions are extended to classes taking values in $\mathbb{R}^d$, with $d > 1$, by taking the maximum of the bracketing numbers of the coordinate classes. Let $\|\cdot\|_{2,P}$ be the $L_2(P)$ norm, i.e. $\|f\|_{2,P}^2 = \int f^2 dP$. When P is clear from the context, we simply write $\|\cdot\|_2 \equiv \|\cdot\|_{2,P}$. Let $|\cdot|$ denote the Euclidean norm, i.e. $|A|^2 = A^\top A$. Let $\|\cdot\|_\infty$ and $\|\cdot\|_{\mathcal{W},\infty}$ denote the *sup*-norms $\|f\|_\infty := \sup_{x \in \mathcal{X}_X} |f(x)|$ and $\|q\|_{\mathcal{W},\infty} := \sup_{W \in \mathcal{W}, w \in \mathcal{X}_W} |q(w|W)|$, respectively. Finally, throughout C denotes a positive constant that may change from expression to expression. We consider the following regularity conditions.

Assumption 1 *The sample observations* $\{Y_i, X_i^\top\}_{i=1}^n$ *are a sequence of iid variables, distributed as* $(Y, X^\top)$*, satisfying* $E[|Y|^s | X = x] < C$ *a.s., for some* $s > 2$*.*

Assumption 2 *The class* $\mathcal{W}$ *is such that* $\log N_{[\cdot]}(\varepsilon, \mathcal{W}, \|\cdot\|_\infty) \leq C\varepsilon^{-v_w}$ *for some* $v_w < 1$ *and all* $\varepsilon > 0$*.*

Assumption 3 *For all* $W \in \mathcal{W}$ *and* $x \in \mathcal{X}_X$ *:* $f(w|W)$*,* $m(w|W)$ *and* $f_X(x|w, W)$ *are* r*-times continuously differentiable in* w*, with uniformly (in* w*,* W *and* x*) bounded derivatives (including the functions themselves), where* r *is as in Assumption 4 below.*

Assumption 4 *The kernel function* $k(t) : \mathbb{R} \to \mathbb{R}$ *is bounded, symmetric, continuously differentiable, and satisfies the following conditions:* $\int k(t)\, dt = 1$*,* $\int t^l k(t)\, dt = 0$ *for* $0 < l < r$*, and* $\int |t^r k(t)|\, dt < \infty$*, for some* $r \geq 2$*;* $|\partial k(t)/\partial t| \leq C$ *and for some* $v > 1$*,* $|\partial k(t)/\partial t| \leq C|t|^{-v}$ *for* $|t| > L$*,* $0 < L < \infty$*.*

Assumption 5 *The possibly data dependent bandwidth* $\widehat{h}_n$ *satisfies* $P(a_n \leq \widehat{h}_n \leq b_n) \to 1$ *as* $n \to \infty$*, for deterministic sequences of positive numbers* a_n *and* b_n *such that: (i)* $b_n \to 0$ *and* $a_n^d n / \log n \to \infty$*; (ii)* $n b_n^{2r} \to 0$*.*

The conditional bounded moment of Assumption 1 can be relaxed to $E[|Y|^s] < C$ by working with bracketing entropies of weighted L_2-norms instead. Assumption 2 restricts the "size" of the class $\mathcal{W}$ with respect to $\|\cdot\|_\infty$. Similarly to Assumption 1, Assumption 2 could be relaxed to bracketing conditions with weighted L_2-norms, which are weaker than the sup-norm $\|\cdot\|_\infty$, at the cost of longer proofs. For simplicity, we focus on the sup-norm $\|\cdot\|_\infty$. van der Vaart and Wellner (1996) contains numerous examples of classes $\mathcal{W}$ satisfying Assumption 2. To give an example, define for any vector a of p integers the differential operator $\partial_x^a := \partial^{|a|_1} / \partial x_1^{a_1} \dots \partial x_p^{a_p}$, where $|a|_1 := \sum_{i=1}^p a_i$. Assume that $\mathcal{X}$ is the finite union of convex, bounded subsets of $\mathbb{R}^p$, with non-empty interior. For any smooth function $h : \mathcal{X} \subset \mathbb{R}^p \to \mathbb{R}$ and some $\eta > 0$, let $\underline{\eta}$ be the largest integer smaller than η, and

$$\|h\|_{\infty,\eta} := \max_{|a|_1 \leq \underline{\eta}} \sup_{x \in \mathcal{X}} |\partial_x^a h(x)| + \max_{|a|_1 = \underline{\eta}} \sup_{x \neq x'} \frac{|\partial_x^a h(x) - \partial_x^a h(x')|}{|x - x'|^{\eta - \underline{\eta}}}.$$

Further, let $C_M^\eta(\mathcal{X})$ be the set of all continuous functions $h : \mathcal{X} \subset \mathbb{R}^p \to \mathbb{R}$ with $\|h\|_{\infty,\eta} \leq M$. Since the constant M is irrelevant for our results, we drop the dependence on M and denote $C^\eta(\mathcal{X})$. Then, it is known that $\log N_{[\cdot]}(\varepsilon, C^\eta(\mathcal{X}), \|\cdot\|_\infty) \leq C\varepsilon^{-v_w}$, $v_w = p/\eta$, so if $\mathcal{W} \subset C^\eta(\mathcal{X}_X)$, then $p < \eta$ suffices

for our Assumption 2 to hold in this example. For extensions to unbounded $\mathcal{X}$ see Nickl and Pötscher (2007).

Assumption 3 is used for controlling the bias of $\widehat{m}$ and related quantities. Assumption 4 is standard in the nonparametric kernel estimation literature, while Assumption 5 permits data dependent bandwidths, as in e.g. Andrews (1995). In particular, our theory allows for plug-in bandwidths of the form $\widehat{h}_n = \widehat{c}h_n$ with $\widehat{c}$ stochastic and h_n a suitable deterministic sequence converging to zero as $n \to \infty$. Andrews (1995) points out that this condition holds in many common data dependent bandwidth selection procedures, such as cross-validation and generalized cross-validation. Similarly, our results also apply to deterministic sequences. In particular if $\widehat{h}_n$ is of the form $\widehat{h}_n = cn^{-\delta}$, for some constant $c > 0$, then Assumption 5 requires that $1/2r < \delta < 1/d$, so r needs to be greater than $d/2$. That is, a simple second-order Gaussian kernel can be used when $d < 4$, in view of Assumption 5.

We now introduce a class of functions that will serve as a parameter space for m and $\widehat{m}$. We assume that for each $W \in \mathcal{W}$, $\mathcal{X}_W$ is a finite union of convex, bounded subsets of $\mathbb{R}^d$, with non-empty interior. Let $\mathcal{T}^\eta$ be a class of measurable functions on $\mathcal{X}_X$, $q\left(W(x)\middle| W\right)$ say, such that $W \in \mathcal{W}$ and q satisfies for a universal constant C_L and each $W_j \in \mathcal{W}$, $j = 1, 2$,

$$\left\| q\left(W_1(\cdot)\middle| W_1\right) - q\left(W_2(\cdot)\middle| W_2\right) \right\|_\infty \leq C_L \left\| W_1 - W_2 \right\|_\infty . \tag{3}$$

Moreover, assume that for each $W \in \mathcal{W}$, $q\left(\cdot\middle| W\right) \in C^\eta(\mathcal{X}_W)$, for some $\eta > \max(1, d/2)$, and $\|q\|_{\mathcal{W},\infty} < \infty$.

Assumption 6 *(i) $m \in \mathcal{T}^\eta$; and (ii) $P(\widehat{m} \in \mathcal{T}^\eta) \to 1$.*

Assumption 6 is a high level condition, some version of which is commonly required in the literature of semiparametric estimation. See e.g. Assumption 2.4 in Chen, Linton, and van Keilegom (2003, p. 1594) and Assumption 3.4(b) in Ichimura and Lee (2010, p. 255). Even for simple cases such as standard kernel estimators, the verification of assumptions like 6(ii) is rather involved, see e.g. Akritas and van Keilegom (2001) and Neumeyer and van Keilegom (2010). Appendix C provides primitive conditions for 6(ii) and similar assumptions, showing how they can be used in complex settings (including ours) that can include possibly data dependent bandwidths and generated regressors.

We next introduce some technical conditions to handle the random trimming factor

$$\widehat{t}_{ni} := \mathbb{I}(X_i \in \widehat{\mathcal{X}}_n),$$

where $\widehat{\mathcal{X}}_n$ is a possibly estimated subset of $\mathcal{X}_X$. For instance, $\widehat{\mathcal{X}}_n$ may depend on estimates of the density of X. We shall assume that $\widehat{\mathcal{X}}_n$ converges to a deterministic set $\mathcal{X}_n \subset \mathcal{X}_X$ in a suitable sense, see Assumption 7 below. Define $t_{ni} := \mathbb{I}(X_i \in \mathcal{X}_n)$, $\Delta t_{ni} := \widehat{t}_{ni} - t_{ni}$ and the rate

$$d_n := \sqrt{\frac{\log a_n^{-d} \vee \log\log n}{n a_n^d}} + b_n^r,$$

where $a \vee b = \max(a, b)$. Note that d_n is the rate of convergence of quantities like $\| \widehat{f} - f \|_{\mathcal{W},\infty}$, where a_n and b_n are as in Assumption 5.

Assumption 7 *(i) There is a sequence τ_n of positive numbers satisfying $\tau_n \leq \inf_{W\in\mathcal{W},x\in\mathcal{X}_n} f\left(W(x)|\,W\right)$, $n\tau_n^{-6}d_n^4 \to 0$ and $\tau_n^{-1}d_n \to 0$; and (ii) $P(X_i \in \mathcal{X}_n) \to 1$ and $E\left[|\Delta t_{ni}|\right] = o(n^{-1/2})$ as $n \to \infty$.*

Assumption 7 allows for densities arbitrarily close to zero. We show below that Assumption 7(ii) holds for typical trimming sequences $\widehat{t}_{ni} = \mathbb{I}(\widehat{f}_i \geq \tau_n)$ under mild regularity conditions, where $\widehat{f}_i$ is a kernel density estimator and τ_n is a sequence going to zero at a suitable rate, see Assumption 11 below. Of course, our results can be easily extended to asymptotically non-vanishing trimming.

We assume that the weight function ϕ lies in a class Φ of real-valued measurable functions of X satisfying the following regularity condition:

Assumption 8 *The class Φ is a class of uniformly bounded functions such that $\log N_{[\cdot]}(\varepsilon, \Phi, \|\cdot\|_2) \leq C\varepsilon^{-v_\phi}$ for some $v_\phi < 2$.*

Assumption 8 restricts the size of the class Φ. This condition is satisfied, for instance, for the indicator class $\Phi = \{\phi(X) = \mathbb{I}\left[X \leq x\right]\colon\, x \in \mathbb{R}^p\}$, which has been extensively used in the nonparametric testing literature. The boundedness restriction in Assumption 8 can be relaxed by requiring instead suitable bounded moments for errors and weights.

Define the parameter space $\mathcal{A} := \mathcal{W} \times \Phi$ and a generic element $\alpha := (W, \phi) \in \mathcal{A}$. We are interested in the asymptotic representation of the process (1), i.e.

$$\hat{\Delta}_n(\alpha) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - \widehat{m}(W(X_i)|W)\}\widehat{t}_{ni}\phi(X_i),$$

that is uniform over $\alpha \in \mathcal{A}$.

Define the error-weighted empirical process $\Delta_n(\alpha)$ as

$$\Delta_n(\alpha) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - m(W(X_i)|W)\}\phi_W^{\perp}(X_i),$$

where henceforth for any generic measurable and integrable function $\phi(\cdot)$ we define

$$\phi_W^{\perp}(X_i) := \phi(X_i) - E[\phi(X_i)|W(X_i)].$$

We prove below that $\hat{\Delta}_n$ and Δ_n are asymptotically uniformly equivalent. This provides a general uniform representation for $\hat{\Delta}_n(\alpha)$ in terms of iid variables. This uniform expansion quantifies the asymptotic effect from estimating true errors by nonparametric kernel regression residuals. To give some informal intuition for the asymptotic equivalence between $\hat{\Delta}_n$ and Δ_n, ignore trimming effects for now and write, for each $\alpha \in \mathcal{A}$,[2]

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - \widehat{m}(W(X_i)|W)\}\phi(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - m(W(X_i)|W)\}\phi(X_i) \\ + \sqrt{n}E_{\mathcal{Z}_n}\left[\left(m(W(X)|W) - \widehat{m}(W(X)|W)\right)\phi(X)\right] \\ + \mathbb{G}_n\left(m(W(\cdot)|W)\phi(\cdot)\right) - \mathbb{G}_n\left(\widehat{m}(W(\cdot)|W)\phi(\cdot)\right),\end{aligned} \tag{4}$$

[2]We thank an anonymous referee for this suggestion.

where $\mathbb{G}_n(\hat{g}) := n^{-1/2}\sum_{i=1}^n\{\hat{g}(X_i) - E_{\mathcal{Z}_n}[\hat{g}(X)]\}$ is the so-called empirical process at $\hat{g}$ and, henceforth, $E_{\mathcal{Z}_n}[\hat{g}(X)] := E[\hat{g}(X)|\,\mathcal{Z}_n]$ denotes the mean operator conditional on the original sample $\mathcal{Z}_n$. By consistency of $\widehat{m}$ and stochastic equicontinuity[3]

$$\mathbb{G}_n(m(W(\cdot)|W)\phi(\cdot)) - \mathbb{G}_n(\widehat{m}(W(\cdot)|W)\phi(\cdot)) = o_P(1),$$

uniformly in $\alpha \in \mathcal{A}$ and $a_n \leq \widehat{h}_n \leq b_n$. The estimation effect in $\widehat{m}(\cdot|W)$ presents the second term in the sum (4), which can be shown to satisfy, by standard bias calculations,

$$\begin{aligned}&\sqrt{n}E_{\mathcal{Z}_n}\left[(m(W(X)|W) - \widehat{m}(W(X)|W))\,\phi(X)\right]\\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^n\{Y_i - m(W(X_i)|W)\}E[\phi(X_i)|W(X_i)] + o_P(1),\end{aligned}$$

uniformly in $\alpha \in \mathcal{A}$ and $a_n \leq \widehat{h}_n \leq b_n$. The following theorem formalizes these arguments. Henceforth, we abstract from measurability problems that may arise in $\hat{\Delta}_n$ when viewed as a process indexed by $\alpha \in \mathcal{A}$, see van der Vaart and Wellner (1996) for details.

Theorem 3.1 *Let Assumptions 1 – 8 hold. Then,*

$$\sup_{a_n \leq \widehat{h}_n \leq b_n}\sup_{\alpha\in\mathcal{A}}|\hat{\Delta}_n(\alpha) - \Delta_n(\alpha)| = o_P(1).$$

Remark 3.1 *If $\phi(X)$ is such that $E[\phi(X)|W(X)] = 0$ a.s., so $\phi_W^\perp \equiv \phi$, then estimation of m has no asymptotic effect in the limit distribution of $\hat{\Delta}_n(\alpha)$. For example, this explains why the limiting distribution of estimators in Ichimura (1993) and Klein and Spady (1993) are not affected by the estimation of nuisance parameters.*

Remark 3.2 *Note that in Theorem 3.1 we do not require an index structure for $E(Y|X)$.*

In the next section we provide one application of Theorem 3.1. Other potential applications for semiparametric inference are discussed in the Introduction and in a supplement to this paper.

3.1 Generated Regressors and Generated Weights

In this section, we apply our uniform expansion to a setting where residuals are from a semiparametric index regression model with nonparametric generated regressors, and possibly nonparametrically generated weights. Specifically, in this section we assume that

$$E[Y|X] = E[Y|v(g_0(X_1), X)] \text{ a.s.},$$

for some d-dimensional known function v of X and a conditional mean function $g_0(x_1) := E[D|X_1 = x_1]$, where D is a random variable and X_1 is a subvector of X, $X_1 \subseteq X$. Thus, the conditioning variable $W_0(X) := v(g_0(X_1), X)$ is known up to the unknown regression g_0. For notational simplicity we only

[3] Appendix A provides a formal definition of stochastic equicontinuity.

consider univariate D but the extension to multivariate D is straightforward. Later in Section 4 we further extend the current setting to $W_0(X) = v(\theta_0, g_0(X_1), X)$ for an unknown finite-dimensional parameter $\theta_0 \in \Theta \subset \mathbb{R}^q$. Similarly, weights are generated as $\phi_0(X) = \psi(X, \gamma_0)$, where ψ is known up to the unknown nuisance parameters γ_0. To simplify the notation, denote $W_{0i} := v(g_0(X_{1i}), X_i)$, $m_{0i} := m(W_{0i}|W_0)$, $g_i := g(X_{1i})$ and $g_{0i} := g_0(X_{1i})$, for $i = 1, ..., n$.

We observe a random sample $\mathcal{Z}_n \equiv \{Y_i, X_i^\top, D_i\}_{i=1}^n$ from the joint distribution of $(Y, X^\top, D)$ and estimate g_0 and γ_0 by some nonparametric estimators $\widehat{g}$ and $\widehat{\gamma}$, respectively, possibly but not necessarily based on kernel methods. We investigate the impact of estimating W_0 by $\widehat{W} = v(\widehat{g}(X_1), X)$ and ϕ_0 by $\widehat{\phi} = \psi(X, \widehat{\gamma})$ in the empirical process $\hat{\Delta}_n(\widehat{W}, \widehat{\phi})$. The goal is to provide an expansion in iid terms for the standardized sample mean of weighted and trimmed residuals

$$\hat{\Delta}_n(\widehat{\alpha}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - \widehat{m}(\widehat{W}_i|\widehat{W})\} \widehat{t}_{ni} \widehat{\phi}(X_i), \tag{5}$$

where $\widehat{\alpha} := (\widehat{W}, \widehat{\phi})$, $\widehat{W}_i := v(\widehat{g}(X_{1i}), X_i)$, $\widehat{\phi}(X_i) = \psi(X_i, \widehat{\gamma})$,

$$\widehat{t}_{ni} := \mathbb{I}(\widehat{f}(\widehat{W}_i|\widehat{W}) \geq \tau_n),$$

and τ_n satisfies Assumption 11 below. Define $\phi_0^\perp(X_i) := \phi_0(X_i) - E[\phi_0(X_i)|W_{0i}]$, $\varepsilon_i := Y_i - m_{0i}$, $u_i := D_i - g_{0i}$ and $f_0 := f(W_0|W_0)$. We show that under regularity conditions

$$\hat{\Delta}_n(\widehat{\alpha}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \varepsilon_i \phi_0^\perp(X_i) - u_i E[\partial_{\bar{g}} m(W_{0i}) \phi_0^\perp(X_i)|X_{1i}] \right\} + o_P(1), \tag{6}$$

where $\partial_{\bar{g}} m(W_{0i}) := \partial m(v(\bar{g}, X_i)|W_0)/\partial \bar{g}|_{\bar{g}=g_{0i}}$.

To obtain similar expansions, existing approaches suggest a functional Taylor expansion argument to handle the second term in (4) and express the problem in terms of the natural parameters $\eta := (m, g, f, \gamma)$. That is, setting

$$s(Y_i, X_i, \eta) := \{Y_i - m(v(g(X_{1i}), X_i)|v(g(X_{1i}), X_i))\} \mathbb{I}(f(W_i|W) \geq \tau_n) \psi(X_i, \gamma),$$

the standard approach writes the term $\sqrt{n} E_{\mathcal{Z}_n} [s(Y, X, \widehat{\eta}) - s(Y, X, \eta_0)]$ as

$$\sqrt{n} V_{\eta_0}(\widehat{\eta} - \eta_0) + \sqrt{n} E_{\mathcal{Z}_n} [s(Y, X, \widehat{\eta}) - s(Y, X, \eta_0) - V_{\eta_0}(\widehat{\eta} - \eta_0)],$$

where $\widehat{\eta} := (\widehat{m}, \widehat{g}, \widehat{f}, \widehat{\gamma})$, $\eta_0 := (m_0, g_0, f_0, \gamma_0)$ and $V_{\eta_0}(\widehat{\eta} - \eta_0)$ is a functional derivative in the direction $(\widehat{\eta} - \eta_0)$ at η_0, see, e.g., Newey (1994) or Chen, Linton, and van Keilegom (2003). The standard approach next requires showing that $\sqrt{n} V_{\eta_0}(\widehat{\eta} - \eta_0)$ is asymptotically normal, often by requiring that it has a linear representation and that, for $C \geq 0$,

$$\begin{aligned} \sqrt{n} \left| E_{\mathcal{Z}_n} [s(Y, X, \widehat{\eta}) - s(Y, X, \eta_0)] - V_{\eta_0}(\widehat{\eta} - \eta_0) \right| &\leq \sqrt{n} C \left\| \widehat{\eta} - \eta_0 \right\|^2 \\ &= o_P(1). \end{aligned} \tag{7}$$

In contrast, our expansion in (6) does not require an explicit analysis of pathwise functional derivatives. This is particularly useful here because the map $g \to E[Y|v(g(X_1), X)]$ is only implicitly defined and it

is not clear how the pathwise functional derivative with respect to f (in the trimming) can be computed. Note that $\partial_{\bar{g}} m(W_{0i})$, defined after (6), is a standard (finite-dimensional) derivative of the regression involving the 'true' index W_0. Similarly, our expansion does not require the rates $\| \widehat{\gamma} - \gamma_0 \| = o_P(n^{-1/4})$ or a linear representation for $\widehat{\gamma} - \gamma_0$. This is important, since often in applications $\widehat{\gamma}$ involves derivatives of nonparametric estimates with slow rates of convergence. This feature is made possible by the stochastic equicontinuity of $\hat{\Delta}_n$ in γ. An alternative method of proof would be to treat ϕ_0 itself as a nuisance parameter and exploit the linearity in ϕ_0 to conclude that C in (7) is zero for the term corresponding to $\widehat{\gamma}$. Then, one can apply Lemma 5.1 in Newey (1994), or (4.1.4)' in Theorem 4.1 in Chen (2007) to show that no rates are needed for $\| \widehat{\phi} - \phi_0 \|$. This alternative method of proof still requires an analysis of the pathwise derivative with respect to ϕ_0, although linearity makes this calculation straightforward. Hence, both methods of proof lead to essentially the same conditions on the estimated weights. Our stochastic equicontinuity approach does not rely on the linearity with respect to ϕ_0. When parameters of interest are stochastic equicontinuous with respect to a nuisance parameter, there is no need for rates and linearization on the nuisance estimates, consistency conditions suffice.

Intuitively, our stochastic equicontinuity proof works as follows. Assuming $P(\widehat{\alpha} \in \mathcal{A}) \to 1$, by Theorem 3.1 we can write

$$\hat{\Delta}_n(\widehat{\alpha}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m(\widehat{W}_i | \widehat{W})\} \widehat{\phi}_{\widehat{W}}^{\perp}(X_i) + o_P(1),$$

where $\widehat{\phi}_{\widehat{W}}^{\perp}(X) := \widehat{\phi}(X) - E_{\mathcal{Z}_n}[\widehat{\phi}(X) | \widehat{W}]$. Then, a stochastic equicontinuity argument as in (4) shows that $\hat{\Delta}_n(\widehat{\alpha})$ is asymptotically equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \phi_0^{\perp}(X_i) + \sqrt{n} E_{\mathcal{Z}_n}[\{Y - m(\widehat{W}(X) | \widehat{W})\} \widehat{\phi}_{\widehat{W}}^{\perp}(X) - \varepsilon \phi_0^{\perp}(X)].$$

We then show that the second term in the last expansion can be further written as

$$\begin{aligned} &\sqrt{n} E_{\mathcal{Z}_n}[\{m(W_0(X)|W_0) - m(\widehat{W}(X)|\widehat{W})\} \widehat{\phi}_{\widehat{W}}^{\perp}(X)] - \sqrt{n} E_{\mathcal{Z}_n}[\varepsilon \{\phi_0^{\perp}(X) - \widehat{\phi}_{\widehat{W}}^{\perp}(X)\}] \\ &= \sqrt{n} E_{\mathcal{Z}_n}[\{m(W_0(X)|W_0) - m(\widehat{W}(X)|W_0)\} \widehat{\phi}_{\widehat{W}}^{\perp}(X)] \\ &+ \sqrt{n} E_{\mathcal{Z}_n}[\{m(\widehat{W}(X)|W_0) - m(\widehat{W}(X)|\widehat{W})\} \widehat{\phi}_{\widehat{W}}^{\perp}(X)] + o_P(1) \\ &= \sqrt{n} E_{\mathcal{Z}_n}[\{m(W_0(X)|W_0) - m(\widehat{W}(X)|W_0)\} \widehat{\phi}_{\widehat{W}}^{\perp}(X)] + o_P(1), \end{aligned} \tag{8}$$

where the first equality follows from the orthogonality of ε and functions of X (by the index restriction) and the second equality follows from the orthogonality of $\phi_{\widehat{W}}^{\perp}(X)$ and functions of $W(X)$. The former orthogonality explains the lack of asymptotic impact from estimating weights, while the latter explains the lack of asymptotic impact from the generated regressors through the second argument in $m(W(X_i)|W)$. These asymptotic results allow for arbitrary rates of consistency for $\widehat{\alpha}$. Finally, we provide sufficient conditions that guarantee that the last term in the expansion is asymptotically equivalent to

$$\sqrt{n} E_{\mathcal{Z}_n}[\{g_0(X_1) - \widehat{g}(X_1)\} \partial_{\bar{g}} m(W_0(X)) \widehat{\phi}_{\widehat{W}}^{\perp}(X_i)] = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i E[\partial_{\bar{g}} m(W_{0i}) \phi_0^{\perp}(X_i) | X_{1i}] + o_P(1). \tag{9}$$

These arguments show, without the need to introduce functional derivatives, that there is zero contribution from estimating W_0 in $m(W|W_0)$.

To prove (9) for a generic first step estimator $\widehat{g}$, we define the empirical processes, for $\alpha = (W, \phi)$,

$$R_n(\alpha) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (D_i - g_i) \partial_{\bar{g}} m(W_{0i}) \phi_W^{\perp}(X_i), \text{ and}$$
$$G_n(\alpha) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i \left\{ \partial_{\bar{g}} m(W_{0i}) \phi_W^{\perp}(X_i) - E[\partial_{\bar{g}} m(W_{0i}) \phi_W^{\perp}(X_i) | X_{1i}] \right\}. \tag{10}$$

Define also the rates $p_n := P(f(W_0(X)|\, W_0) \leq 2\tau_n)$, $w_n := \|\widehat{g} - g_0\|_\infty$ and $q_n := \tau_n^{-1} d_n + w_n$. Note that q_n is the rate of convergence of quantities like $\| m(\widehat{W}(\cdot)|\widehat{W}) - m(W_0(\cdot)|W_0) \|_\infty$.

Assumption 9 *The function $m(v(\bar{g}, x)|W_0)$ is twice continuously differentiable in the scalar $\bar{g}$ with uniformly bounded derivatives in x.*

Assumption 10 *The estimator $\widehat{\alpha}$ is such that: (i) $|R_n(\widehat{\alpha}) - G_n(\widehat{\alpha})| = o_P(1)$; (ii) $E[D^2|X] < C$ a.s., $\|\widehat{g} - g_0\|_2 = o_P(n^{-1/4})$, $W_0 \in \mathcal{W}$ and $P(\widehat{W} \in \mathcal{W}) \to 1$; and (iii) $\| \widehat{\phi} - \phi_0 \|_2 = o_P(1)$, $\phi_0 \in \Phi$, and $P(\widehat{\phi} \in \Phi) \to 1$.*

Assumption 11 *τ_n is a sequence of positive numbers satisfying $\tau_n \to 0$, $n\tau_n^{-6} d_n^4 \to 0$ and $n(\tau_n^{-l} q_n^l + p_n)^2 \to 0$, for some $l \geq 2$.*

Assumption 9 is a standard smoothness condition. Assumption 10(i) is a high level assumption. However, a special case of our Theorem 3.1 provides primitive conditions for this assumption to hold when $\widehat{g}$ is a NW kernel estimator, see Assumption C.1(i). Section 4 contains primitive conditions for Assumption 10 in a binary choice model with selection. When $\widehat{g}$ is a series estimator, primitive conditions for Assumption 10(i) can be found in Escanciano and Song (2010). Assumption 11 can be substantially relaxed if $\{\varepsilon_i\}_{i=1}^n$ are conditionally uncorrelated given $\{\widehat{\alpha}_i\}_{i=1}^n$.

Theorem 3.2 *Let Assumptions 1 – 6 and 8 – 11 hold. Assume that $E[Y|X] = E[Y|W_0(X)]$ a.s. Then the expansion in (6) holds uniformly in $a_n \leq \widehat{h}_n \leq b_n$.*

Remark 3.3 *Theorem 3.2 can be easily extended to the case where the index restriction does not hold. In that case, there will be an additional term in the asymptotic expansion coming from the second term in (8). Details are omitted.*

Theorem 3.2 quantifies the estimation effect of $\widehat{\alpha}$ in the empirical process $\hat{\Delta}_n(\widehat{\alpha})$. Only estimating g has an impact on the limiting distribution of $\hat{\Delta}_n(\widehat{\alpha})$, the impact from estimating the weights is asymptotically negligible. Although W depends on g_0, the Theorem shows that the estimation error $\widehat{g} - g_0$ only has an asymptotic impact through the first argument in $m(W(X_i)|W)$. For the process $\hat{\Delta}_n(\widehat{\alpha})$, the contribution from the second argument is asymptotically negligible due to the orthogonality of $\phi_W^{\perp}$ with functions of W.

To illustrate the usefulness of Theorem 3.1 and Theorem 3.2 for estimation and testing, and to show how they would be applied in practice, in the next section we use them to derive asymptotic theory for a new estimator of a binary choice model with selection, and then we apply them to the construction of a new directional specification test.

4 Example: A Binary Choice Model with Selection

Suppose a latent binary variable Y^* satisfies the ordinary threshold crossing binary response model $Y^* = \mathbb{I}\left(X^\top\theta_0 - e \geq 0\right)$ with e independent of X, in short $e \perp X$, and the distribution function of e, F_e, may be unknown. Suppose further that we only observe Y^* for some subset of the population, indexed by a binary variable D, i.e. we only observe $Y = Y^*D$. This is a sample selection model with a binary outcome. The econometrician is assumed to know relatively little about selection D other than that it is binary, so let D be given by the nonparametric threshold crossing model $D = \mathbb{I}\left[g_0\left(X\right) - u \geq 0\right]$ where $u \perp X$ and the function $g_0(X)$ is unknown. Based on Matzkin (1992), we may without loss of generality assume $g_0\left(X\right) = E[D|X]$ and u has a uniform distribution, since then $P(D = 1|X) = P\left[u \leq g_0\left(X\right)\right] = g_0\left(X\right)$.

We then have the model

$$D = \mathbb{I}\left[g_0\left(X\right) - u \geq 0\right] \tag{11}$$

$$Y = \mathbb{I}(X^\top\theta_0 - e \geq 0)D \tag{12}$$

The latent error terms e and u are not independent of each other, so the model does not have selection on observables. When $g_0(\cdot)$ is assumed to be linear in X, e.g. $g_0(X) = X^\top\delta_0$ for some unknown coefficient vector δ_0, and the joint distribution of the errors $(e, u)^\top$ is assumed to be normal, model (11) – (12) is known as the 'Censored Probit Model' or a 'Probit Model with Sample Selection,' see e.g. van de Ven and van Praag (1981) and Meng and Schmidt (1985). In the absence of joint normality, Klein, Shen and Vella (2012) provide semiparametric maximum likelihood estimators of θ_0 and δ_0. Our estimator is similar to theirs, except that they assume the parameterization $g_0(X) = X^\top\delta_0$, while we estimate $g_0(X)$ nonparametrically, and we permit trimming, weighting, and bandwidth choice that are all data dependent. The class of models covered by the special regressor based semiparametric estimator in Lewbel (2007) includes our model equation (11) – (12) when the marginal density of e is parameterized.[4]

Let $(e, u)^\top$ be drawn from an unknown joint distribution function $F\left(e, u\right)$ with $e, u \perp X$. Then, with $g_0\left(X\right) = E[D|X]$ and $m = F$, it holds that

$$E[Y|X] = m[X^\top\theta_0, g_0\left(X\right)],$$

[4]Our methods could also be applied to other related models, e.g., if we replaced (12) with $Y = \left(X^\top\theta_0 - e\right)D$, then this would be a semiparametric generalization of the standard Heckman's (1979) selection model, and if we replaced (12) with $Y = \max\left(X^\top\theta_0 - e, 0\right)D$ then this would be a semiparametric generalization of Cragg's (1971) double hurdle model. See our appendix for more examples.

so an index restriction with $W_0(X) := v(\theta_0, g_0, X) = (X^\top\theta_0, g_0(X))$ holds, and $g_0(X)$ is identified from the selection equation as a conditional expectation.[5] Sufficient conditions for identification of m and θ_0 in this model, which is possible even without an exclusion restriction like that assumed by Klein, Shen and Vella (2012) and others, are provided by Escanciano, Jacho-Chávez and Lewbel (2012), who also propose a semiparametric least squares estimator for this model.[6]

The class of functions

$$\mathcal{W} = \{x \to (x^\top\theta, g(x)) : \theta \in \Theta_0 \subset \mathbb{R}^q,\ g \in \mathcal{G} \subset C^{\eta_g}(\mathcal{X}_X),\ \|g - g_0\|_\infty < \delta\}, \tag{13}$$

for an arbitrarily small $\delta > 0$ and $\eta_g > p$ is used for the remaining part of this section.

4.1 Semiparametric Maximum Likelihood Estimation

We here propose a semiparametric maximum likelihood estimator (SMLE) of θ_0 in model (11) – (12). Analogous to Klein, Shen and Vella (2012), we use the Klein and Spady (1993) method of semiparametrically estimating the likelihood function, after plugging in a first stage estimator of g_0, though both our model and our estimator is more general than theirs, as noted earlier. We then apply our uniform convergence theorems to obtain limiting distribution theory for our estimator.

Our model has $Y = D = 0$ if $u > g_0(X)$, $Y = D = 1$ if both $e \leq X^\top\theta_0$ and $u \leq g_0(X)$, and otherwise $Y = 0$ and $D = 1$. Therefore, $P(Y = D = 0|X) = P(D = 0|X) = 1 - E[D|X] = 1 - g_0(X)$, $P(Y = D = 1|X) = P(Y = 1|X) = E[Y|X] = m[X^\top\theta_0, g_0(X)]$ and $P(Y = 0, D = 1|X) = 1 - P(Y = D = 0|X) - P(Y = D = 1|X) = E[D|X] - E[Y|X] = g_0(X) - m[X^\top\theta_0, g_0(X)]$. Based on these probabilities, define the following semiparametric log-likelihood objective function

$$\mathcal{L}_n(\theta, \widehat{g}) := \frac{1}{n}\sum_{i=1}^{n} \{Y_i \log[\widehat{m}_{i\theta}] + (D_i - Y_i)\log[\widehat{g}_i - \widehat{m}_{i\theta}]\}\,\widetilde{t}_{in}\psi_i, \tag{14}$$

where $\widehat{g}_i := \widehat{g}(X_i)$ is the NW estimator of g_0 with possibly data-driven bandwidth $\widehat{h}_{gn}$, $\widehat{m}_{i\theta} := \widehat{m}(W_i(\theta, \widehat{g})|W(\theta, \widehat{g}))$, $W(\theta, g) := (X^\top\theta, g(X))$, $W_i(\theta, g) := (X_i^\top\theta, g(X_i))$ with possible data-driven bandwidth $\widehat{h}_n$, and $\widetilde{t}_{in}$ is a trimming sequence that also accounts for the possibility that $\widehat{m}$ is close to zero or to $\widehat{g}_i$. More specifically, the trimming has the form

$$\widetilde{t}_{in} := \mathbb{I}(\tau_n \leq \widetilde{m}_i \leq \widehat{g}_i - \tau_n) \times \mathbb{I}(\tau_n \leq \widetilde{f}_i),$$

where τ_n is a sequence of positive numbers with $\tau_n \to 0$ as $n \to \infty$, $\widetilde{m}_i := \widehat{m}(\widetilde{W}_i|\widetilde{W})$, $\widetilde{f}_i := \widehat{f}(\widetilde{W}_i|\widetilde{W})$, $\widetilde{W}_i := W_i(\widetilde{\theta}, \widehat{g})$, $\widetilde{W} := W(\widetilde{\theta}, \widehat{g})$, and $\widetilde{\theta}$ is a preliminary consistent estimator for θ_0. For instance, $\widetilde{\theta}$ can be a SMLE with $\psi_i = 1$ and fixed trimming $\widehat{t}_{ni} = \mathbb{I}(X_i \in A)$ for a compact set $A \subset \mathcal{X}_X$ (or

[5] If instead of the assumption $e, u \perp X$ we had the more general assumption $u \perp X$ and $e|u, X \sim e|u, W_0$, then the above model would still hold with $F_{eu}(e, u|W_0)$ denoting the conditional distribution of $e, u|W_0$ and the function $m(r, g)$ now defined as $m(r, g) = F_{eu}(r, g|W_0)$.

[6] Closely related identification and estimation results include Newey (2007), Blundell and Powell (2004) and Ichimura and Lee (1991).

non-data-dependent asymptotic trimming). Note that weighting $\psi_i = 1$ and either fixed or non-data-dependent asymptotic trimming will in general make $\widetilde{\theta}$ inefficient, but that does not violate the required Assumption 14 below for an initial consistent $\widetilde{\theta}$. The estimator for θ_0 we propose is then

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta, \widehat{g}). \tag{15}$$

The estimator $\widehat{\theta}$ extends the related estimator in Klein and Spady (1993) for the single-index binary choice model in two ways. First, the objective function (15) is different and has a nonparametric generated regressor $\widehat{g}$ associated with selection, which complicates the relevant asymptotic theory. Second, and intimately related to the first, is that unlike in Klein and Spady (1993), adaptive weighting is necessary here to improve efficiency due to the presence of the nonparametric generated regressor.

Given identification of θ_0 in this model as discussed in the previous section, it is straightforward to demonstrate consistency of $\widehat{\theta}$. Given consistency, we now apply our convergence theorems to derive limiting distribution theory for $\widehat{\theta}$, allowing for data dependent choice of bandwidth, data dependent asymptotic trimming, and data dependent adaptive weighting for efficiency.

Recall the definitions $\varepsilon_i = Y_i - m_{0i}$ and $u_i = D_i - g_{0i}$. Further, define $\upsilon_i := \varepsilon_i - u_i\partial_{\bar{g}}m(W_{0i})$, $\sigma_{0i}^2 \equiv \sigma_0^2(X_i) := E[\upsilon_i^2|X_i]$, $\psi_i := \psi(W_{0i})$ and $\partial_\theta m(W_{0i}) := \partial m(W_i(\theta, g_0)|W(\theta, g_0))/\partial\theta|_{\theta=\theta_0}$. Also note that

$$\sigma_{0i}^2 = m_{0i}(1-m_{0i}) + (\partial_{\bar{g}}m(W_{0i}))^2 g_{0i}(1-g_{0i}) - 2\partial_{\bar{g}}m(W_{0i})m_{0i}(1-g_{0i}). \tag{16}$$

We shall assume that the following matrix is non-singular and finite (this is little more than a linear index model identification condition),

$$\Gamma_0 := E\left[\frac{g_{0i}\partial_\theta m(W_{0i})\partial_\theta^\top m(W_{0i})}{m_{0i}(g_{0i}-m_{0i})}\psi_i\right]. \tag{17}$$

Define the rates $\tau_{ng} := \inf_{\{x:f(W_0(x)|W_0)<2\tau_n\}} f_X(x)$,

$$d_{mn} := \sqrt{\frac{\log a_n^{-2} \vee \log\log n}{na_n^8}} \qquad \text{and} \qquad d_{gn} := \sqrt{\frac{\log l_n^{-p} \vee \log\log n}{nl_n^{2p+2}}}.$$

Finally, to simplify the notation define $m_\theta := m(W(\theta, g_0)|W(\theta, g_0))$. The rate conditions $\tau_n^{-2}d_{mn} = O(1)$ and $\tau_{ng}^{-2}d_{gn} = O(1)$ will be part of the low-level sufficient conditions for $P(\widehat{m} \in \mathcal{T}^\eta) \to 1$, $P(\partial_\theta\widehat{m} \in \mathcal{T}^\eta) \to 1$, $P(\widehat{g} \in \mathcal{G}) \to 1$ and related high-level conditions.

Assumption 12 *(i) The kernel function satisfies Assumption 4, is twice continuously differentiable and also satisfies* $\left|\partial^{(j)}k(t)/\partial t^j\right| \leq C|t|^{-v}$ *for* $|t| > L_j$, $0 < L_j < \infty$, *for* $j = 1, 2$; *(ii) the sequence* τ_n *is such that* $\tau_n \to 0$, $\tau_n^{-2}d_{mn} = O(1)$; *(iii) The functions* $\sigma_0^2(\cdot)$, $\inf_{\theta\in\Theta_0} m_\theta$ *and* $\inf_{\theta\in\Theta_0}(g_0 - m_\theta)$ *are bounded away from zero and* $\sup_{\theta\in\Theta_0}\partial_\theta m$ *is bounded.*

Assumption 13 *(i) The regression function* $g_0(X) = E[D|X]$ *is estimated by a NW kernel estimator* $\widehat{g}$ *with a kernel function that is* $(p+1)-$*times continuously differentiable, satisfies Assumption 4 with* $r = \rho$ *and a possibly stochastic bandwidth* $\widehat{h}_{gn}$ *satisfying* $P(l_n \leq \widehat{h}_{gn} \leq u_n) \to 1$ *as* $n \to \infty$, *for deterministic sequences of positive numbers* l_n *and* u_n *such that* $u_n \to 0$, $\tau_{ng}^{-2}d_{gn} = O(1)$ *and* $n\tau_{ng}^{-2}u_n^{2\rho} \to 0$; *(ii) the function* g_0 *and the density* $f_X(\cdot)$ *of* X *are* ρ*-times continuously differentiable in* x, *with bounded derivatives. Furthermore,* g_0 *is bounded away from zero,* $g_0 \in \mathcal{G} \subset C^{\eta_g}(\mathcal{X}_X)$ *for some* $\eta_g > p$.

Assumption 14 *The parameter space Θ_0 is a compact subset of $\mathbb{R}^p$ and θ_0 is an element of its interior. The estimators $\widetilde{\theta}$ and $\widehat{\theta}$ are consistent for θ_0. The matrix Γ_0 is non-singular and $E\left[\psi_1^2\right] < \infty$.*

Theorem 4.1 *Let Assumption 1 hold for model (11) – (12), and let Assumptions 3 – 5, 9, 11 – 14 hold. Then $\widehat{\theta}$ is asymptotically normal, i.e.*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \longrightarrow_d N(0, \Gamma_0^{-1}\Sigma_0\Gamma_0^{-1}),$$

where Γ_0 is given in (17) and

$$\Sigma_0 := E\left[\frac{\sigma^2(W_{0i})g_{0i}^2\partial_\theta m(W_{0i})\partial_\theta^\top m(W_{0i})}{m_{0i}^2(g_{0i} - m_{0i})^2}\psi_i^2\right].$$

Remark 4.1 *Consider a bandwidth of the form $\widehat{h}_n = ch_n$, with c a constant to be chosen and h_n a suitable deterministic sequence satisfying the assumptions in Theorem 4.1 above. Then, a natural data-driven choice for the constant c is one that maximizes an estimated semiparametric likelihood criterion, i.e.*

$$\widehat{c}_n = \underset{c\in[\epsilon,\epsilon^{-1}]:\widehat{h}_n = ch_n}{\arg\max} \frac{1}{n}\sum_{i=1}^n \left\{Y_i \log[\widetilde{m}_{i;\widehat{h}_n}] + (D_i - Y_i)\log[\widehat{g}_i - \widetilde{m}_{i;\widehat{h}_n}]\right\}\mathbb{I}(X_i \in A),$$

where ϵ is an arbitrarily small positive number, and we have made explicit the dependence of the leave-one-out version of estimator $\widetilde{m}_i$ on the bandwidth $\widehat{h}_n$. Note that the resulting bandwidth $\widehat{c}_n h_n$ will automatically satisfy our required assumptions by construction. Furthermore, when using this choice in (14) no changes to Theorem 4.1 are needed by virtue of our uniformity-in-bandwidth results.

It can be easily shown that using weights $\psi_i^* := m_{0i}[g_{0i} - m_{0i}]/[\sigma_{0i}^2 g_{0i}]$ leads to a more efficient estimator[7] with asymptotic variance Γ_*^{-1}, where the positive definite matrix Γ_* is given by

$$\Gamma_* := E\left[\frac{\partial_\theta m(W_0(X))\partial_\theta^\top m(W_0(X))}{\sigma^2(W_0(X))}\right].$$

Now let $\widehat{\psi}_i^* = \widetilde{m}_i(\widehat{g}_i - \widetilde{m}_i)/(\widehat{\sigma}_i^2\widehat{g}_i)$, where $\widehat{\sigma}_i^2 = \widetilde{m}_i(1 - \widetilde{m}_i) + (\partial_{\overline{g}}\widetilde{m}_i)^2\widehat{g}_i(1 - \widehat{g}_i) - 2\partial_{\overline{g}}\widetilde{m}_i\widetilde{m}_i(1 - \widehat{g}_i)$ and $\partial_{\overline{g}}\widetilde{m}_i := \partial\widehat{m}(W_i(\widetilde{\theta}, \overline{g})|W(\widetilde{\theta}, \widehat{g}))/\partial\overline{g}|_{\overline{g}=\widehat{g}_i}$. The estimator $\widehat{\sigma}_i^2$ although positive in large samples, may be negative in small samples. Our theory can be straightforwardly extended to other estimators for σ_i^2.

Let $\widehat{\theta}^*$ be the resulting estimator when the optimal weight $\psi_i = \widehat{\psi}_i^*$ is used in (14). Furthermore, let $\widehat{\Gamma}_* = n^{-1}\sum_{i=1}^n \partial_\theta\widehat{m}_{i\widehat{\theta}^*}\partial_\theta^\top\widehat{m}_{i\widehat{\theta}^*}/\widehat{\sigma}_i^2$, where $\partial_\theta\widehat{m}_{i\widehat{\theta}^*} := \partial\widehat{m}(W_i(\theta, \widehat{g})|W(\theta, \widehat{g}))/\partial\theta|_{\theta=\widehat{\theta}^*}$. The following result is the analogous to Theorem 4.1 for the case of optimal estimated weights.

Corollary 4.1 *Let the Assumptions of Theorem 4.1 hold. Then $\sqrt{n}(\widehat{\theta}^* - \theta_0) \longrightarrow_d N(0, \Gamma_*^{-1})$, and $\widehat{\Gamma}_* \to_P \Gamma_*$.*

The proof of Corollary 4.1 is provided in the Appendix A. The proof of Theorem 4.1 itself is almost the same (except simpler, since it does not involve estimated weights), and so is omitted. It is shown that the asymptotic distribution of $\widehat{\theta}^*$ is the same as it would be if the optimal weights ψ_i^* and the regression function m were known instead of estimated.

[7] Whether Γ_*^{-1} coincides with the semiparametric efficiency bound of θ_0 in model (11) – (12) is an open question.

4.2 A Tailor-Made Specification Test

We next illustrate the usefulness of our uniform convergence results for constructing test statistics. A policy parameter of considerable interest in applications of binary choice models is the average structural function (ASF). See, e.g., Stock (1989), Blundell and Powell (2004) and Newey (2007). In our binary choice model with selection described above, the ASF is given by

$$\gamma^* := \int\int m(x^\top \theta_0, g) f_g(g) f_X^*(x) dx dg,$$

where f_X^* is a particular marginal density for X and f_g is the density of g_0. In this context, suppose we are concerned with possible misspecification of the semiparametric binary choice model only to the extent that it leads to inconsistent estimation of the ASF γ^*. Our goal is the construction of both a test and an associated bandwidth choice procedure that concentrates power in this direction.

Consider a directional specification test with these alternatives in mind, testing the correct specification of the model

$$H_0 : E[Y|X] = m[X^\top \theta_0, g_0(X)] \text{ a.s,}$$

against alternatives for which

$$E[\{Y - m(X^\top \theta_0, g_0(X))\}\phi_*(X, W_0(X))] \neq 0, \tag{18}$$

where $\phi_*(X, W_0(X)) := f_g(g_0(X)) f_X^*(X) / f(W_0(X)|W_0)$.

We propose constructing such a test based on

$$T_{n,h_n} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(\widehat{W}_i|\widehat{W}_i)\}\widehat{\phi}_*(X_i, \widehat{W}_i)\widehat{t}_{ni},$$

where $\widehat{W}_i := (X_i^\top \widehat{\theta}, \widehat{g}_i)$, $\widehat{\theta}$ is a consistent estimator for θ_0, such as (15) from the previous section, $\widehat{\phi}_*(X_i, \widehat{W}_i) = \widehat{f}_{ig}(\widehat{g}_i) f_X^*(X_i)/\widehat{f}_i$, $\widehat{f}_{ig}$ is a kernel estimator for the density of g_0 resulting from integrating $\widehat{f}_i \equiv \widehat{f}(\widehat{W}_i|\widehat{W}_i)$, $\widehat{t}_{in} = \mathbb{I}(\widehat{f}_i \geq \tau_n)$ and h_n in T_{n,h_n} denotes the bandwidth used in estimating $\widehat{m}$. Set $\sigma^2 := E[\varepsilon_i^2 \phi_*^{\perp 2}(X_i, W_{0i})]$, and consider the variance estimator

$$\widehat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}(\widehat{W}_i|\widehat{W})\}^2 \widehat{\phi}_*^{\perp 2}(X_i, \widehat{W}_i)\widehat{t}_{ni},$$

where $\widehat{\phi}_*^{\perp}$ is based on a uniformly consistent estimator of $E[f_X^*(X_i)|W_0(X_i)]$. Define the rate

$$d_{Tn} := \sqrt{\frac{\log a_n^{-2} \vee \log\log n}{n a_n^6}}.$$

Next theorem justifies the asymptotic test based on rejecting when $W_{n,h_n} := \widehat{\sigma}^{-2} T_{n,h_n}^2$ is larger than the corresponding critical value from the chi-squared distribution with one degree of freedom.

Theorem 4.2 *Let Assumption 1 hold for model (11) – (12), and let Assumptions 3 – 5, 9, 11 and 13 – 14 hold. Furthermore, assume $\tau_n^{-2} d_{Tn} = O(1)$ and that $\phi_*(\cdot, W_0)$ is bounded. Then, under H_0,*

$$\max_{a_n \leq h_n \leq b_n} W_{n,h_n} \longrightarrow_d \chi_1^2,$$

whereas under the alternative (18), $\max_{a_n \leq h_n \leq b_n} W_{n,h_n} \longrightarrow_P \infty$.

Remark 4.2 *Let $\widehat{h}_n$ denote any solution of $\arg\max_{a_n \leq h_n \leq b_n} W_{n,h_n}$. Our uniform in bandwidth theorems allow us to choose the bandwidth by this optimization, which leads to a test with better finite sample power properties than a test that uses a bandwidth $\widehat{h}_n$ optimized for estimation like that described in remark 4.1. See e.g. Horowitz and Spokoiny (2001) for a related approach in a different context.*

5 Concluding Remarks

We have obtained a new uniform expansion for standardized sample means of weighted regression residuals from nonparametric or semiparametric models, with possibly nonparametric generated regressors. We have shown by examples how these results are useful for deriving limiting distribution theory for estimators and tests. Additional example applications of our uniform expansions are provided in a supplement to this paper available from the authors.

Our asymptotic theory deals with general forms of data dependent bandwidths. An example we employ is where the bandwidth rate is chosen by the practitioner based on theory, and the data dependent constant is chosen by minimizing the same objective function that is used for estimation of model parameters. While this is a natural criterion to use for bandwidth choice, a topic for future research would be the development of data dependent bandwidth selection rules with some formally proven optimality properties. Still less is known about optimal bandwidth rates for testing. In our testing example we choose the bandwidth to maximize the test statistic in a region of admissible bandwidths, and show that this choice is permitted by our asymptotic results.

The limiting distributions in our applications, given in Theorems 4.1 and 4.2 as well as Corollary 4.1, might have been obtainable through more traditional methods, such as Newey and McFadden (1994), Chen, Linton, and van Keilegom (2003), and Ichimura and Lee (2010). For example, Lemma 6 in Rothe (2009, p. 62) required that all nuisance parameters converge at rate faster than $n^{-1/4}$, including the nuisance parameters comprising $\widehat{\phi}$. Moreover, use of these traditional methods generally requires first calculating pathwise derivatives for each object of interest, using the Riesz representation theorem to solve integral equations in each case, see e.g. Newey (1994). Once these solutions were found, one would need to plug them into the influence function, and work out the Hoeffding decomposition of general data-dependent U-statistics. In the final step, one would need to show the stochastic equicontinuity of the resulting influence functions with respect to a smoothing parameter with a 'rule-of-thumb' asymptotic representation, see e.g. Li and Li (2010).

In contrast, our proofs of Corollary 4.1 and Theorem 4.2 in the Appendix A show how repeated applications of our Theorems 3.1 and 3.2 provides an easier, shorter method of proof that sometimes avoids many of the aforementioned steps. The relative ease of deriving the asymptotic properties of semiparametric estimators and test statistics using our proposed tools is further illustrated in a supplement to this paper, where we sketch how Theorems 3.1 and 3.2 could be generically applied to derive the asymptotic properties of well-known semiparametric estimators such as Ichimura (1993), Klein and Spady (1993) and Rothe (2009).

Similarly, U-statistic or U-processes theory, which is the standard method of proof in nonparametric or semiparametric testing analysis, see e.g. Delgado and González Manteiga (2001), would be hardly

applicable to our testing problem. In contrast, our method fits in naturally and leads to the desired expansion in terms of iid terms under mild regularity conditions.

The appealing properties of our estimators and tests regarding data driven bandwidths, possibly nonparametric generated regressors, random trimming and estimated weights are made possible by the use of uniform convergence in these aspects which, when combined with standard stochastic equicontinuity arguments, allows us to establish the desired expansions.

Our results should have applications beyond the types considered here. In particular, expansions of the kind provided by Theorem 3.1 and Theorem 3.2 are the key ingredient in proving the consistency of bootstrap procedures for estimation and testing in semiparametric models.

Appendix A Main Proofs

Before we prove our main results we need some preliminary results from empirical processes theory. Let $N(\varepsilon, \mathcal{G}, \|\cdot\|)$ be the *covering number with respect to* $\|\cdot\|$, i.e. the minimal number of ε-balls with respect to $\|\cdot\|$ needed to cover $\mathcal{G}$.

Lemma A.1 *Let $\mathcal{F}$ and $\mathcal{G}$ be classes of functions with a bounded and a squared integrable envelope F and G, respectively, then*

$$N_{[\cdot]}(\epsilon, \mathcal{F}\cdot\mathcal{G}, \|\cdot\|_2) \leq N_{[\cdot]}(C\epsilon, \mathcal{F}, \|\cdot\|_2) \times N_{[\cdot]}(C\epsilon, \mathcal{G}, \|\cdot\|_2).$$

Proof of Lemma A.1: The proof when the brackets involve positive functions is trivial. For the general case, we proceed as in the proof of Lemma A.3 below. *Q.E.D.*

Lemma A.2 *Let Assumptions 3 and 8 hold. Then, for each W_1 and W_2 in $\mathcal{W}$, and all $\delta > 0$,*

$$\sup_{\phi\in\Phi} \|E[\phi(X)|W_1(X) = W_1(\cdot)] - E[\phi(X)|W_2(X) = W_2(\cdot)]\|_\infty \leq C \|W_1 - W_2\|_\infty .$$

Proof of Lemma A.2: The proof follows from Lemma A2(ii) in Song (2008, p. 1495), noting that Assumption 3 implies his condition (A.35) with $s = 1$. *Q.E.D.*

Let $\mathcal{S}$ be a class of measurable functions of X. Let $\{\xi_i, X_i^\top\}_{i=1}^n$ denote a random sample from the joint distribution of $(\xi, X^\top)$ taking values in $\mathcal{X}_\xi \times \mathcal{X}_X \in \mathbb{R}^{1+p}$, and define the weighted empirical process, indexed by $s \in \mathcal{S}$,

$$\Psi_n(s) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\xi_i s(X_i) - E[\xi_i s(X_i)]\} .$$

We say that Ψ_n is asymptotically uniformly ρ-equicontinuous at $s_0 \in \mathcal{S}$, for a pseudo-metric ρ on $\mathcal{S}$, if for all $\varepsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} P\left[\sup_{s_1\in\mathcal{S}:\rho(s_1,s_0)<\delta} |\Psi_n(s_1) - \Psi_n(s_0)| > \varepsilon\right] \leq \eta.$$

The following result gives sufficient conditions for uniform $\|\cdot\|_2$-equicontinuity of Ψ_n. One important implication of the uniform equicontinuity is that $\Psi_n(\widehat{s}) = \Psi_n(s_0) + o_P(1)$, provided $\|\widehat{s} - s_0\|_2 = o_P(1)$.

Lemma A.3 *Assume $E[\xi_i^2|X_i] < L$ a.s., and let $\mathcal{S}$ be a class of uniformly bounded functions such that $\log N_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2) \leq C\varepsilon^{-v_s}$ for some $v_s < 2$. Then, Ψ_n is asymptotically uniformly $\|\cdot\|_2$-equicontinuous at $s_0 \in \mathcal{S}$, for all s_0.*

Proof of Lemma A.3: Define the class of functions $\mathcal{G} := \{(\xi, x) \to \xi s(x) : s \in \mathcal{S}\}$. Let $a^+ := \max\{a, 0\}$ and $a^- := \max\{-a, 0\}$ denote the positive and negative parts of a, respectively. Let $\{[s_{lj}, s_{uj}] : j = 1, ..., N_\varepsilon \equiv N_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2)\}$ be a family of ε-brackets (with respect to $\|\cdot\|_2$) covering $\mathcal{S}$. Then, it holds that $\{[\xi^+ s_{lj} - \xi^- s_{uj}, \xi^+ s_{uj} - \xi^- s_{lj}] : j = 1, ..., N_\varepsilon\}$ is also a family of $L^{1/2}\varepsilon$-brackets covering $\mathcal{G}$. Then, by our assumptions, $\mathcal{G}$ has finite bracketing entropy, and hence, Ψ_n is $\|\cdot\|_2$-equicontinuous at all points in $\mathcal{S}$. *Q.E.D.*

Lemma A.4 *Under the Assumptions of Theorem 3.1, $\sup_{\alpha\in\mathcal{A}} |R_n(\alpha)| = o_P(1)$, where*

$$R_n(\alpha) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - \widehat{m}(W(X_i)|W)\}\Delta t_{ni}\phi(X_i).$$

Proof of Lemma A.4: Write

$$\begin{aligned} R_n(\alpha) := & \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - m(W(X_i)|W)\}\Delta t_{ni}\phi(X_i) \\ & + \frac{1}{\sqrt{n}} \sum_{i=1}^n \{m(W(X_i)|W) - \widehat{m}(W(X_i)|W)\}\Delta t_{ni}\phi(X_i) := R_{1n}(\alpha) + R_{2n}(\alpha). \end{aligned}$$

By Markov's inequality, we obtain

$$\begin{aligned} \sup_{\alpha\in\mathcal{A}} |R_{1n}(\alpha)| &= O_P\left(\sqrt{n} E\left[|\Delta t_{ni}|\right]\right) \\ &= o_P(1). \end{aligned}$$

The proof that $\sup_{a_n \leq \widehat{h}_n \leq b_n} \sup_{\alpha\in\mathcal{A}} |R_{2n}(\alpha)| = o_P(1)$ follows as for R_{1n}, hence, it is omitted. *Q.E.D.*

Proof of Theorem 3.1: We write,

$$\begin{aligned} \hat{\Delta}_n(\alpha) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - \widehat{m}(W(X_i)|W)\}t_{ni}\phi(X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - \widehat{m}(W(X_i)|W)\}\Delta t_{ni}\phi(X_i) \\ &=: S_n(\alpha) + R_n(\alpha), \end{aligned}$$

By Lemma A.4, $R_n(\alpha) = o_P(1)$, uniformly in $\alpha \in \mathcal{A}$ and $a_n \leq \widehat{h}_n \leq b_n$.

To handle S_n we shall apply Theorem 2.11.9 in van der Vaart and Wellner (1996, p. 211) to the array $Z_{ni}(\lambda) = n^{-1/2}(Y_i - m(X_i))t_{ni}\phi(X_i)$, where $\lambda = (m, \phi) \in \Lambda$, and $\Lambda := \mathcal{T}^\eta \times \Phi$. By Triangle inequality and definition of $\mathcal{T}^\eta$, it follows that

$$\sum_{i=1}^n E[\sup |Z_{ni}(\lambda_2) - Z_{ni}(\lambda_1)|^2] \leq C\delta^2,$$

where the sup is taken over $\lambda_2 = (m_2, \phi_2) \in \Lambda$ such that $\|m_2 - m_1\|_\infty < \delta$ and $\|\phi_2 - \phi_1\|_2 < \delta$, for a fixed $\lambda_1 = (m_1, \phi_1) \in \Lambda$. Then, using the notation of Theorem 2.11.9 in van der Vaart and Wellner (1996, p. 211), for any $\varepsilon > 0$, by Lemma A.1

$$N_{[\cdot]}(\varepsilon, \Lambda, L_2^n) \leq N_{[\cdot]}\left(\varepsilon C, \mathcal{T}^\eta, \|\cdot\|_\infty\right) \times N_{[\cdot]}\left(\varepsilon C, \Phi, \|\cdot\|_2\right).$$

Hence, by Lemma B.2 in Ichimura and Lee (2010, p. 262) and Assumption 8, Λ satisfies $\int_0^1 \sqrt{N_{[\cdot]}(\varepsilon, \Lambda, L_2^n)} < \infty$. On the other hand, for any $\delta > 0$, by Chebyshev's inequality

$$\begin{aligned}\sum_{i=1}^n E[\sup_\lambda |Z_{ni}(\lambda)| \, \mathbb{I}(\sup_\lambda |Z_{ni}(\lambda)| > \delta)] &\leq C n^{1/2} E[|Y| \, \mathbb{I}(|Y| > C n^{1/2}\delta)] \\ &\leq \frac{C E[|Y|^2]}{n^{1/2}\delta^2} \to 0.\end{aligned}$$

Hence, the conditions of Theorem 2.11.9 in van der Vaart and Wellner (1996, p. 211) are satisfied and $\sum_{i=1}^n Z_{ni}(\lambda) - E[Z_{ni}(\lambda)]$ is asymptotically stochastic equicontinuous with respect to the pseudo-metric $\rho(\lambda_1, \lambda_2) := \max\{\|m_2 - m_1\|_\infty, \|\phi_2 - \phi_1\|_2\}$. The stochastic equicontinuity, Assumption 6 and our results in Appendix B imply that, uniformly in $\alpha \in \mathcal{A}$,

$$\begin{aligned}S_n(\alpha) = &\frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - m(W(X_i)|W)\} t_{ni} \phi(X_i) \\ &- \sqrt{n} E_{\mathcal{Z}_n}[\{\widehat{m}(W(X_i)|W) - m(W(X_i)|W)\} t_{ni} \phi(X_i)] + o_P(1) =: \Delta_{0n}(\alpha) - \Delta_{1n}(\alpha) + o_P(1).\end{aligned}$$

We shall prove that

$$\sup_{\alpha \in \mathcal{A}} \left| \Delta_{0n}(\alpha) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - m(W(X_i)|W)\} \phi(X_i) \right| = o_P(1) \tag{A-19}$$

and

$$\sup_{a_n \leq \widehat{h}_n \leq b_n} \sup_{\alpha \in \mathcal{A}} \left| \Delta_{1n}(\alpha) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - m(W(X_i)|W)\} E[\phi(X_i)|W(X_i)] \right| = o_P(1). \tag{A-20}$$

The equality in (A-19) follows from an application of Theorem 2.11.9 in van der Vaart and Wellner (1996, p. 211) to the array $Z_{ni}(\alpha) = n^{-1/2}\{Y_i - m(W(X_i)|W)\}\phi(X_i)(t_{ni} - 1)$ and the pointwise convergence to zero of $\sum_{i=1}^n Z_{ni}(\lambda) - E[Z_{ni}(\lambda)]$ for each λ. We prove now (A-20). To simplify notation denote $\iota_{n\alpha}(w) := E[\phi(X_i) t_{ni} | W(X_i) = w]$, $\alpha \in \mathcal{A}$, and note that $\Delta_{1n}(\alpha) = \sqrt{n} E_{\mathcal{Z}_n}[\iota_{n\alpha}(W(X_i)) (\widehat{m}(W(X_i)|W) - m(W(X_i)|W))]$. We write

$$\widehat{m}(w|W) - m(w|W) = a_n(w|W) + r_n(w|W),$$

where

$$a_n(w|W) := f^{-1}(w|W) \left(\widehat{T}(w|W) - T(w|W) - m(w|W) \left(\widehat{f}(w|W) - f(w|W) \right) \right),$$

$T(w|W) := m(w|W) f(w|W)$ and

$$r_n(w|W) := -\frac{\widehat{f}(w|W) - f(w|W)}{\widehat{f}(w|W)} a_n(w|W).$$

Note that uniformly in w and W, $\mathbb{I}(X_i \in \mathcal{X}_n) \leq t_n(w|W) := \mathbb{I}(f(w|W) \geq \tau_n)$. This inequality and the results in Appendix B imply that $\sup |r_n(w|W)| t_n(w|W) = o_P(n^{-1/2})$ under our assumptions on the bandwidth. It then follows that $\Delta_{1n}(\alpha)$ is uniformly bounded by

$$\int \iota_{n\alpha}(w) [\widehat{T}(w|W) - T(w|W)] dw \tag{A-21}$$

$$- \int \iota_{n\alpha}(w) m(w|W) [\widehat{f}(w|W) - f(w|W)] dw \tag{A-22}$$

$$+ o_P(n^{-1/2}).$$

We now look at terms (A-21)-(A-22). Firstly, it follows from our results in Appendix B that the difference between $T(w|W)$ and $E[\widehat{T}(w|W)]$ is $o_P(n^{-1/2})$. Hence, uniformly in $\alpha \in \mathcal{A}$,

$$\begin{aligned}
&\int \iota_{n\alpha}(w) [\widehat{T}(w|W) - T(w|W)] dw = \int \iota_{n\alpha}(w) [\widehat{T}(w|W) - E(\widehat{T}(w|W))] dw + o_P(n^{-1/2}) \\
&= \frac{1}{n} \sum_{j=1}^{n} Y_j \int \iota_{n\alpha}(w) K_h(W_j - w) dw - \int \iota_{n\alpha}(w) E(Y_j K_h(W_j - w)) dw + o_P(n^{-1/2}), \\
&= \frac{1}{n} \sum_{j=1}^{n} \iota_{n\alpha}(W_j) Y_j - E[\iota_{n\alpha}(W_j) m(W_j|W)] + o_P(n^{-1/2}),
\end{aligned}$$

where the last equality follows from the change of variables $u = h^{-1}(W_j - w)$, Assumptions 3, 5 and the fact that, uniformly in $\alpha \in \mathcal{A}$ and $a_n \leq \widehat{h}_n \leq b_n$,

$$\begin{aligned}
\int \iota_{n\alpha}(w) K_h(W_j - w) dw &= \int_{\mathcal{X}_n} \phi(x) \left(\int f_X(x|w, W) K_h(W_j - w) dw \right) dx \\
&= \int_{\mathcal{X}_n} \phi(x) f_X(x|W_j, W) dx + O(b_n^r).
\end{aligned}$$

Likewise, the term (A-22) becomes $n^{-1/2} \sum_{j=1}^{n} \iota_{n\alpha}(W_j) m(W_j|W) - E[\iota_{n\alpha}(W_j) m(W_j|W)] + o_P(n^{-1/2})$. In conclusion, we have uniformly in $\alpha \in \mathcal{A}$, that $\Delta_{1n}(\alpha) = n^{-1/2} \sum_{j=1}^{n} \iota_{n\alpha}(W_j) [Y_j - m(W_j|W)] + o_P(n^{-1/2})$. Applying again Theorem 2.11.9 in van der Vaart and Wellner (1996, p. 211) to the array $n^{-1/2}[Y_j - m(W_j|W)][\iota_{n\alpha}(W_j) - \iota_\alpha(W_j)]$, where $\iota_\alpha(w) := E[\phi(X_i)|W(X_i) = w]$, $\alpha \in \mathcal{A}$, proves (A-20) and hence the result of the Theorem. *Q.E.D.*

Proof of Theorem 3.2: We write, using $\widehat{m}_i := \widehat{m}(\widehat{W}_i|\widehat{W})$ and $t_{ni} := \mathbb{I}(f(W_0(X_i)|W_0) > \tau_n)$,

$$\begin{aligned}
\hat{\Delta}_n(\widehat{\alpha}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}_i\} t_{ni} \widehat{\phi}(X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}_i\} \Delta t_{ni} \widehat{\phi}(X_i) \\
&=: \widehat{S}_n(\widehat{\alpha}) + \widehat{R}_n(\widehat{\alpha}).
\end{aligned}$$

We write $\widehat{R}_n(\widehat{\alpha}) = n^{-1/2}\sum_{i=1}^n \varepsilon_i \Delta t_{ni}\widehat{\phi}(X_i) - n^{-1/2}\sum_{i=1}^n \{\widehat{m}_i - m_{0i}\}\Delta t_{ni}\widehat{\phi}(X_i) =: \widehat{R}_{1n} - \widehat{R}_{2n}$.

Using the simple inequalities, with $f_i \equiv f\left(W_0(X_i)\middle| W_0\right)$ and $\widehat{f}_i \equiv \widehat{f}\left(\widehat{W}_i\middle|\widehat{W}\right)$,

$$|\Delta t_{ni}| \le \mathbb{I}(\tau_n \le f_i \le 2\tau_n) + \mathbb{I}(|\widehat{f}_i - f_i| > \tau_n),$$

$$|\Delta t_{ni}| \le \mathbb{I}(f_i \ge \tau_n)\mathbb{I}(\widehat{f}_i < 2\tau_n) + \mathbb{I}(\widehat{f}_i \ge 2\tau_n)\mathbb{I}(|\widehat{f}_i - f_i| > \tau_n), \tag{A-23}$$

$$\mathbb{I}(\widehat{f}_i < \tau_n) \le \mathbb{I}(f_i \le 2\tau_n) + \mathbb{I}(|\widehat{f}_i - f_i| > \tau_n), \tag{A-24}$$

the term $\widehat{R}_{1n}$ is shown to satisfy $\widehat{R}_{1n}(\widehat{\alpha}) = O_P\left(\sqrt{n}(\tau_n^{-l}q_n^l + p_n)\right) = o_P(1)$, by the arguments of the proof of Lemma A.4 and the uniform rates for $\| \widehat{f} - f \|^2_{\mathcal{W},\infty}$.

While the second term $\widehat{R}_{2n}$ can be further decomposed as

$$\begin{aligned}\widehat{R}_{2n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\widehat{m}_i - m(\widehat{W}_i|\widehat{W})\}\Delta t_{ni}\widehat{\phi}(X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \{m(\widehat{W}_i|\widehat{W}) - m_{0i}\}\Delta t_{ni}\widehat{\phi}(X_i)\\ &:= \widehat{R}_{2n;a} + \widehat{R}_{2n;b}.\end{aligned}$$

We can further write, using (A-23),

$$\begin{aligned}E|\widehat{R}_{2n;a}| \le{}& \| (\widehat{m} - m)\mathbb{I}(f \ge \tau_n) \|_{\mathcal{W},\infty} \sqrt{n}P(\widehat{f}(\widehat{W}_i|\widehat{W}) < 2\tau_n)\\ &+ \| (\widehat{m} - m)\mathbb{I}(\widehat{f} \ge 2\tau_n) \|_{\mathcal{W},\infty} \frac{\sqrt{n}\| \widehat{f} - f \|^l_\infty}{\tau_n^l}.\end{aligned} \tag{A-25}$$

The results in Appendix B and (A-24) yield

$$\sqrt{n}P(\widehat{f}(\widehat{W}_i|\widehat{W}) < \tau_n) = O_P(n^{1/2}(p_n + \tau_n^{-l}q_n^l)),$$

$\| (\widehat{m} - m)\mathbb{I}(f \ge \tau_n) \|_{\mathcal{W},\infty} = O_P(\tau_n^{-1}d_n)$ and $\| (\widehat{m} - m)\mathbb{I}(\widehat{f} \ge 2\tau_n) \|_{\mathcal{W},\infty} = O_P(\tau_n^{-1}d_n)$. Thus, from (A-25) and the previous rates

$$\begin{aligned}\widehat{R}_{2n;a} &= O_P(\tau_n^{-1}d_n)O_P(n^{1/2}(p_n + \tau_n^{-l}q_n^l))\\ &= o_P(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\widehat{R}_{2n;b} &= O_P(w_n)O_P(n^{1/2}(p_n + \tau_n^{-l}q_n^l))\\ &= o_P(1).\end{aligned}$$

Recall $\widehat{\phi}^{\perp}_{\widehat{W}}(X) := \widehat{\phi}(X) - E[\widehat{\phi}(X)|\widehat{W}]$. Then, using $\widehat{R}_n(\widehat{\alpha}) = o_P(1)$ and Theorem 3.1, we can write, provided $P(\widehat{\alpha} \in \mathcal{A}) \to 1$, uniformly in $a_n \le \widehat{h}_n \le b_n$,

$$\begin{aligned}\hat{\Delta}_n(\widehat{\alpha}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - m(\widehat{W}_i|\widehat{W})\}\widehat{\phi}^{\perp}_{\widehat{W}}(X_i) + o_P(1)\\ &= \widetilde{\Delta}_n(\widehat{\alpha}) + o_P(1).\end{aligned}$$

We now apply Lemma A.1 to the classes $\mathcal{F} \equiv \mathcal{S} = \{s(x) = \phi(x) - E[\phi(X)|W = W(x)] : \alpha \in \mathcal{A}\}$ and $\mathcal{G} \equiv \mathcal{D} = \{y - m(W(x)|W) : W \in \mathcal{W}\}$ to conclude that the product class $\mathcal{M} = \mathcal{D} \times \mathcal{S}$ is Donsker. Hence, by Lemma A.3, it holds that

$$\hat{\Delta}_n(\widehat{\alpha}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \phi_0^{\perp}(X_i) + \sqrt{n} E_{\mathcal{Z}_n}[\{Y - m(\widehat{W}(X)|\widehat{W})\}\widehat{\phi}_{\widehat{W}}^{\perp}(X) - \varepsilon \phi_0^{\perp}(X)] + o_P(1). \tag{A-26}$$

From the arguments in (8), the second term in the last expansion is asymptotically equivalent to

$$\sqrt{n} E_{\mathcal{Z}_n}[\{m(W_0(X)|W_0) - m(\widehat{W}(X)|W_0)\}\widehat{\phi}_{\widehat{W}}^{\perp}(X)].$$

By the rates on the L_2-norm, and Assumptions 9 and 10, this term is asymptotically equivalent to

$$\sqrt{n} E_{\mathcal{Z}_n}[\{g_0(X_1) - \widehat{g}(X_1)\}\partial_{\bar{g}} m(W_0(X))\widehat{\phi}_{\widehat{W}}^{\perp}(X_i)]. \tag{A-27}$$

To handle this term, note that the class $\{\{g_0(x_1) - g(x_1)\}\partial_{\bar{g}} m(W_0(x))s(x) : \alpha \in \mathcal{A}, s \in \mathcal{S}\}$ is Donsker, by an application of Lemma A.1, where $\mathcal{S}$ is defined above. Hence, by stochastic equicontinuity and the consistency of $\widehat{g}$,

$$\begin{aligned}
&\sqrt{n} E_{\mathcal{Z}_n}[\{g_0(X_1) - \widehat{g}(X_1)\}\partial_{\bar{g}} m(W_0(X))\widehat{\phi}_{\widehat{W}}^{\perp}(X_i)] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g_0(X_1) - \widehat{g}(X_1)\}\partial_{\bar{g}} m(W_0(X))\widehat{\phi}_{\widehat{W}}^{\perp}(X_i) + o_P(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{D_i - \widehat{g}(X_1)\}\partial_{\bar{g}} m(W_0(X))\widehat{\phi}_{\widehat{W}}^{\perp}(X_i) \\
&- \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i \partial_{\bar{g}} m(W_0(X))\widehat{\phi}_{\widehat{W}}^{\perp}(X_i) + o_P(1).
\end{aligned}$$

On the other hand, by our Assumption 10(i) this expansion can be further simplified as

$$\frac{-1}{\sqrt{n}} \sum_{i=1}^{n} u_i E_{\mathcal{Z}_n}[\partial_{\overline{g}} m(W_{0i})\widehat{\phi}_{\widehat{W}}^{\perp}(X)|X_{1i}].$$

Finally, an application of Lemma A.3 with $\mathcal{S} = \{s(x) = E[\partial_{\overline{g}} m(W_{0i})\phi_W^{\perp}(X)|X_1 = x_1] : \alpha \in \mathcal{A}\}$, yields

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i E_{\mathcal{Z}_n}[\partial_{\overline{g}} m(W_{0i})\widehat{\phi}_{\widehat{W}}^{\perp}(X)|X_{1i}] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i E[\partial_{\overline{g}} m(W_{0i})\phi_W^{\perp}(X)|X_{1i}] + o_P(1). \tag{A-28}$$

These results along with the equality in (A-26) yield the desired result. *Q.E.D.*

Proof of Corollary 4.1: Our estimator satisfies the first order condition, for large n,

$$0 = \frac{1}{n} \sum_{i=1}^{n} [Y_i \widehat{g}_i - D_i \widehat{m}_{i\widehat{\theta}^*}] \frac{\partial_\theta \widehat{m}_{i\widehat{\theta}^*}}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})} \widehat{\psi}_i^* \widetilde{t}_{in}.$$

Simple algebra and a standard Taylor series expansion around θ_0 yield

$$Y_i \widehat{g}_i - D_i \widehat{m}_{i\widehat{\theta}^*} = [Y_i - \widehat{m}_{i\theta_0}]\widehat{g}_i - [D_i - \widehat{g}_i]\, \widehat{m}_{i\theta_0} - D_i \partial_\theta^\top \widehat{m}_{i\bar{\theta}}(\widehat{\theta}^* - \theta_0),$$

where $\bar{\theta}$ is such that $|\bar{\theta} - \theta_0| \leq |\widehat{\theta}^* - \theta_0|$ a.s. It then follows that, for a sufficiently large n,

$$\sqrt{n}(\widehat{\theta}^* - \theta_0) = \Gamma_n^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y_i - \widehat{m}_{i\theta_0}] \frac{\widehat{g}_i \partial_\theta \widehat{m}_{i\widehat{\theta}^*}}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})} \widehat{\psi}_i^* \widetilde{t}_{in} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [D_i - \widehat{g}_i] \frac{\partial_\theta \widehat{m}_{i\widehat{\theta}^*}}{(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})} \widehat{\psi}_i^* \widetilde{t}_{in} \right)$$
$$\equiv \Gamma_n^{-1}(A_{1n} - A_{2n}), \tag{A-29}$$

where

$$\Gamma_n := \frac{1}{n} \sum_{i=1}^{n} \frac{D_i \partial_\theta \widehat{m}_{i\widehat{\theta}^*} \partial_\theta^\top \widehat{m}_{i\bar{\theta}}}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})} \widehat{\psi}_i^* \widetilde{t}_{in}.$$

We first show that all the nonparametric estimates in these expressions converge in the sup–norm (and hence in the L_2-norm) to their corresponding population analogs. We focus on the derivative term $\partial_\theta \widehat{m}_{i\widehat{\theta}^*}$, since other terms are simpler. Note that

$$\widehat{f}(w|\theta, g)\, \partial_\theta \widehat{m}(w|\theta, g) = \partial_\theta \widehat{T}(w|\theta, g) - \widehat{m}(w|\theta, g)\, \partial_\theta \widehat{f}(w|\theta, g). \tag{A-30}$$

To show the convergence of the right hand side (r.h.s), using a simplified notation, we write

$$\sup |\partial_\theta \widehat{f}(w|\theta, g) - \partial_\theta f(w|\theta, g)| \leq \sup |\partial_\theta \widehat{f}(w|\theta, g) - E\partial_\theta \widehat{f}(w|\theta, g)| + \sup |E\partial_\theta \widehat{f}(w|\theta, g) - \partial_\theta f(w|\theta, g)|$$
$$\equiv I_{1n} + I_{2n},$$

where the sup is over the set $a_n \leq \widehat{h}_n \leq b_n$, $\theta \in \Theta_0$, $w \in \mathcal{Q}_\mathcal{W}$ and $g \in \mathcal{G}$. From Lemma B.8, it follows that

$$I_{1n} = O_P\left(\sqrt{\frac{\log a_n^{-2} \vee \log\log n}{n a_n^4}} \right).$$

By the classical change of variables and integration by parts, for any $a_n \leq h \leq b_n$,

$$E\left[\partial_\theta \widehat{f}(w|\theta, g) - \partial_\theta f(w|\theta, g)\right] = \frac{1}{h^3} E\left[X \partial_{w_1} K\left(\frac{w - W(\theta, g)}{h} \right) - \partial_\theta f(w|\theta, g) \right]$$
$$= \int \partial_{w_1} m(w - uh|\theta, g) K(u)\, du - \partial_\theta f(w|\theta, g),$$

where $m(w|\theta, g) = r(w|\theta, g) f(w|\theta, g)$ and $r(w|\theta, g) := E[X|W(\theta, g) = w]$. By a Taylor series expansion,

$$I_{2n} = O\left(b_n^{r-1} \frac{1}{r!} \left\| \partial_w^{r-1} \partial_{w_1} m \right\|_{\Theta \times \mathcal{G}, \infty} \right) = O(b_n^{r-1}).$$

The proof for $\partial_\theta \widehat{T}$ follows the same arguments as for $\partial_\theta \widehat{f}$, and hence is omitted. Therefore by simple but somewhat tedious algebra one can show that

$$\left\| \widehat{f} \partial_\theta \widehat{m}_{\widehat{\theta}^*}(\cdot|\widehat{\theta}^*, \widehat{g}) - f \partial_\theta m_{\theta_0}(\cdot|W_0) \right\|_\infty = o_P(1).$$

Using (A-30) and that $\partial_\theta \widehat{m}_{\widehat{\theta}^*}$ is bounded for large n, write A_{1n} as

$$A_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y_i - \widehat{m}_{i\theta_0}] \frac{\partial_\theta \widehat{T}_{i\widehat{\theta}^*}}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})} \frac{\widetilde{m}_i(\widehat{g}_i - \widetilde{m}_i)}{\widehat{\sigma}_i^2 \widehat{f}_{i\widehat{\theta}^*}} \widetilde{t}_{in}$$
$$- \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y_i - \widehat{m}_{i\theta_0}] \frac{\partial_\theta \widehat{f}_{i\widehat{\theta}^*}}{(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})} \frac{\widetilde{m}_i(\widehat{g}_i - \widetilde{m}_i)}{\widehat{\sigma}_i^2 \widehat{f}_{i\widehat{\theta}^*}} \widetilde{t}_{in}$$
$$=: A_{11n} - A_{12n}.$$

To analyze A_{11n}, notice that the function $\Upsilon(x_1, x_2, x_3, x_4) = (x_3/x_1)\cdot(x_4/x_2)$ is Lipschitz on $[\varepsilon, \varepsilon^{-1}]^2 \times [-C, C] \times [0, 1]$, for any $\varepsilon > 0$. We shall apply Theorem 2.10.6 in van der Vaart and Wellner (1996, p. 192) to the transformation Υ. To that end, we consider the classes of functions for any $\varepsilon > 0$,

$$\mathcal{F}_{1\varepsilon} := \{l = m(g - m) : m \in T^\eta, g \in \mathcal{G}, l > \varepsilon\},$$

$$\mathcal{F}_{2\varepsilon} := \left\{l = \sigma^2 f : \sigma^2 \in \Sigma, f \in T^\eta, l > \varepsilon\right\},$$

where

$$\Sigma := \left\{ \begin{array}{c} l = m(1-m) + m_g g(1-g) - 2m_g m(1-g) \\ : m, m_g \in T^\eta, g \in \mathcal{G}, l > \varepsilon \end{array} \right\}.$$

By sucesive applications of Theorem 2.10.6, each of the coordinate classes is Donsker, and hence, the class $\Upsilon(\mathcal{F}_{1\varepsilon}, \mathcal{F}_{2\varepsilon}, T^\eta, \mathcal{F}_{2\varepsilon})$ is Donsker and satisfies the entropy condition needed for Φ in Theorem 3.2. Moreover, by the results in Appendix C and the conditions $\tau_n^{-2} d_{mn} = O(1)$ and $\tau_{ng}^{-2} d_{gn} = O(1)$, it holds that $P(\widehat{m} \in \mathcal{T}^\eta) \to 1$, $P(\widetilde{m}_i \in \mathcal{T}^\eta) \to 1$, $P(\partial_\theta \widehat{T}_{i\widehat{\theta}^*} \in \mathcal{T}^\eta) \to 1$, $P(\widehat{g} \in \mathcal{G}) \to 1$ and $P(\widehat{\sigma}^2 \in \Sigma) \to 1$. Hence, we can apply Theorem 3.2 to conclude

$$A_{11n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \phi_{10}^{\perp}(X_i) + o_P(1),$$

where $\phi_{10}(X_i) := \partial_\theta T_{i\theta_0} / (f_{i\theta_0} \sigma_{0i}^2)$. Applying the same arguments to A_{12n}, we get

$$A_{12n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \phi_{20}^{\perp}(X_i) + o_P(1),$$

where $\phi_{20}(X_i) := \partial_\theta f_{i\theta_0} m_{0i} / (f_{i\theta_0} \sigma_{0i}^2)$. Then, using $E[\partial_\theta m(W_0(X_i))|W_0] = 0$ a.s., which can be shown as in Lemma 5.6 in Ichimura (1993, p. 95), we conclude

$$A_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \frac{\partial_\theta m_{i\theta_0}}{\sigma_{0i}^2} + o_P(1). \tag{A-31}$$

The same arguments applied to A_{1n} are also applied to A_{2n}, to prove that $A_{2n} = o_P(1)$. Thus, we conclude that

$$\sqrt{n}(\widehat{\theta}^* - \theta_0) = \Gamma_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \frac{\partial_\theta m_{i\theta_0}}{\sigma_{0i}^2} + o_P(1),$$

and the result of the corollary follows from the Lindeberg-Lévy Central Limit Theorem and Slutsky's Lemma, after proving that $\Gamma_n^{-1} = \Gamma^{-1} + o_P(1)$. The latter equality, and that of $\widehat{\Gamma}_* = \Gamma_* + o_P(1)$, is easy to prove, and details are therefore omitted. *Q.E.D.*

Proof of Theorem 4.2: With $\Upsilon(x, y) = x/y$, consider the class of functions $\Upsilon(\mathcal{F}_1, \mathcal{F}_{2\varepsilon})$, where

$$\mathcal{F}_1 = \left\{h(x) = f_X^*(x) \cdot f(g(x)) : f \in C^1([0,1]), g \in \mathcal{G}\right\},$$

$$\mathcal{F}_{2\varepsilon} = \{l \in T^\eta, l > \varepsilon\}.$$

By Theorem 2.10.6 in van der Vaart and Wellner (1996, p. 192) the class $\Upsilon(\mathcal{F}_1, \mathcal{F}_{2\varepsilon})$ is Donsker and satisfies the entropy condition needed for Φ in Theorem 3.2. Moreover, by the results in Appendix C, the condition $\tau_n^{-2} d_{Tn} = O(1)$ is sufficient for $P(\widehat{f}_i \in \mathcal{F}_{2\varepsilon}) \to 1$ and $P(f_X^*(x) \cdot \widehat{f}_{ig} \in \mathcal{F}_1) \to 1$. Using our uniform rates for kernel estimators one can show that $\widehat{\phi}_*(X_i, \widehat{W}_i)$ converges uniformly to $\phi_*(X, W_0(X))$. Hence, by Theorem 3.2, uniformly in $a_n \leq \widehat{h}_n \leq b_n$,

$$T_{n,h} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \phi_*^{\perp}(X_i, W_{0i}) + o_P(1).$$

On the other hand, it is straightforward to prove that $\widehat{\sigma}^2 = \sigma^2 + o_P(1)$. The limiting null distribution then follows from the Lindeberg-Lévy Central Limit Theorem and Slutsky's Lemma. Under the alternative,

$$\frac{1}{\sqrt{n}} T_{n,h} := \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}(\widehat{W}_i | \widehat{W}_i)\} \widehat{\phi}_*(X_i, \widehat{W}_i) \widehat{t}_{ni},$$

converges to $E\left(\varepsilon_i \phi_*(X_i, W_{0i})\right) \neq 0$, and hence the consistency follows. *Q.E.D.*

Appendix B Uniform Consistency Results for Kernel Estimators

This section establishes rates for uniform consistency of kernel estimators used in the paper. These auxiliary results complement related ones in Andrews (1995) and Sperlich (2009), among others, but we impose different conditions on the kernel functions and provide alternative methods of proof. Unlike Mammen, Rothe and Schienle (2012), we consider uniform in bandwidth consistency and rates, though we do not provide uniform limiting distributions. Einmahl and Mason (2005) also study uniform in bandwidth consistency of kernel estimators, but they did not consider the extension to kernel estimators of (possibly nonparametrically) generated observations, as we do. The results of this section, which should be potentially useful in other settings, are more general than required for the proofs of the main the results in the text.

We first state some well-known results from the empirical process literature. Define the generic class of measurable functions $\mathcal{G} := \{z \to m(z, \theta, h) : \theta \in \Theta, h \in \mathcal{H}\}$, where Θ and $\mathcal{H}$ are endowed with the pseudo-norms $|\cdot|_\Theta$ and $|\cdot|_\mathcal{H}$, respectively. The following result is Theorem 2.14.2 in van der Vaart and Wellner (1996, p. 240).

Lemma B.1 *Let $\mathcal{G}$ be a class of measurable functions with a measurable envelope G. Then, there exists a constant C such that*

$$E[\|\mathbb{G}_n\|_\mathcal{G}] \leq C \|G\|_2 \int_0^1 \sqrt{1 + \log N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|_2)} d\varepsilon.$$

The following result is the celebrated Talagrand's inequality (see Talagrand, 1994). Rademacher variables are iid variables $\{\varepsilon_i\}_{i=1}^n$ such that $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$.

Lemma B.2 *Let $\mathcal{G}$ be a class of measurable functions satisfying $\|g\|_\infty \leq M < \infty$ for all $g \in \mathcal{G}$. Then it holds for all $t > 0$ and some universal positive constants A_1 and A_2 that*

$$P\left(\max_{1\leq m\leq n}\|\mathbb{G}_m\|_{\mathcal{G}} \geq A_1\left(E\left\|\sum_{i=1}^n \varepsilon_i g(Z_i)\right\|_{\mathcal{G}} + t\right)\right) \leq 2\left\{\exp\left(-\frac{A_2 t^2}{n\sigma_{\mathcal{G}}^2}\right) + \exp\left(-\frac{A_2 t}{M}\right)\right\},$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of iid Rademacher variables, independent of the sample $\{Z_i\}_{i=1}^n$ and $\sigma_{\mathcal{G}}^2 := \sup_{g\in\mathcal{G}} var(g(Z))$.

We now proceed with the main results of this section. Let Υ be a class of measurable real-valued functions of Z and let $\mathcal{W}$ be a class of measurable functions of X with values in $\mathbb{R}^d$. Define $\mathcal{Q}_{\mathcal{W}} := \{W(x) \in \mathbb{R}^d : W \in \mathcal{W} \text{ and } x \in \mathcal{X}_X\}$. We denote by $\psi := (\varphi, w, W)$ a generic element of the set $\Psi := \Upsilon \times \mathcal{Q}_{\mathcal{W}} \times \mathcal{W}$. Let $\Psi_I := \Upsilon \times I \times \mathcal{W}$, for a compact set $I \subset \mathcal{Q}_{\mathcal{W}}$. Let $f(w|W)$ denote the density of $W(X)$ evaluated at w. Define the regression function $c(\psi) := E[\varphi(Z)|W(X) = w]$. Henceforth, we use the convention that a function evaluated outside its support is zero. Then, an estimator for $m(\psi) := c(\psi)f(w|W)$ is given by

$$\widehat{m}_h(\psi) = \frac{1}{nh^d}\sum_{i=1}^n \varphi(Z_i) K\left(\frac{w - W(X_i)}{h}\right),$$

where $K(w) = \prod_{l=1}^d k(w_l)$, $k(\cdot)$ is a kernel function, $h := h_n > 0$ is a bandwidth and $w = (w_1, \ldots, w_d)^\top$. We consider the following regularity conditions on the data generating process, kernel, bandwidth and classes of functions.

Assumption B.1 *The sample observations $\{Z_i := (Y_i^\top, X_i^\top)^\top\}_{i=1}^n$ are a sequence of independent and identically distributed (iid) variables, distributed as $Z \equiv (Y^\top, X^\top)^\top$.*

Assumption B.2 *Assumption 2 in the main text holds.*

Assumption B.3 *The density $f(w|W)$ is uniformly bounded, i.e. $\|f\|_{\mathcal{W},\infty} < C$.*

Assumption B.4 *Assumption 4 in the main text holds.*

Assumption B.5 *The possibly data-dependent bandwidth h satisfies $P(a_n \leq h \leq b_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers a_n and b_n such that $b_n \to 0$ and $a_n^d n/\log n \to \infty$.*

Given the class $\mathcal{W}$ and the compact set $I \subset \mathcal{Q}_{\mathcal{W}}$, we define the class of functions

$$\mathcal{K}_0 := \left\{x \to K\left(\frac{w - W(x)}{h}\right) : w \in I, W \in \mathcal{W}, h \in (0, 1]\right\}.$$

Our first result bounds the complexity of the class $\mathcal{K}_0$, which is crucial for the subsequent analysis.

Lemma B.3 *Under Assumption B.4, for a positive constant C_1,*

$$N_{[\cdot]}(C_1\varepsilon, \mathcal{K}_0, \|\cdot\|_2) \leq C\varepsilon^{-v} N(\varepsilon^2, \mathcal{W}, \|\cdot\|_\infty), \text{ for some } v \geq 1. \tag{B-32}$$

Proof of Lemma B.3: Let $W_1, ..., W_{N_{1\varepsilon}}$ be the centers of an ε^2-cover of $\mathcal{W}$ with respect to $\|\cdot\|_\infty$, where $N_{1\varepsilon} = N(\varepsilon^2, \mathcal{W}, \|\cdot\|_\infty)$. Fix j, $j = 1, ..., N_{1\varepsilon}$, and consider the marginal class

$$\mathcal{K}_{0,j} := \left\{ x \to K\left(\frac{w - W_j(x)}{h}\right) : w \in I, h \in (0,1] \right\}.$$

We will show that under our assumptions, $\mathcal{K}_{0,j}$ is a VC class for each j, hence $N(\varepsilon, \mathcal{K}_{0,j}) \leq C\varepsilon^{-v}$ for some $v \geq 1$. Notice that $\mathcal{K}_{0,j} = \prod_{l=1}^{d} \mathcal{K}_{0,j,l}$ where

$$\mathcal{K}_{0,j,l} := \left\{ x \to k\left(\frac{w_l - W_{jl}(x)}{h}\right) : w_l \in I_l, h \in (0,1] \right\},$$

where $I_l := \{w_l : w \in I\}$ and $W_j(x) = (W_{j1}(x), ..., W_{jd}(x))^\top$. Hence, by Lemma 2.6.18 in van der Vaart and Wellner (1996, p. 147) it suffices to prove that $\mathcal{K}_{0,j,l}$ is a VC subgraph class. Moreover, by the same lemma, without loss of generality (as k is of bounded variation), we can assume that k is non-decreasing on $\mathbb{R}$. Recall that $\mathcal{K}_{0,j,l}$ is a VC subgraph class if and only if its class of subgraphs is a VC class of sets, which holds if the class

$$\mathcal{S}_{\mathcal{K}} = \left\{ \left\{ (x,t) : k\left(\frac{w_l - W_{jl}(x)}{h}\right) < t \right\} : w_l \in I_l, h \in (0,1] \right\},$$

is a VC subgraph class. But this follows from an application of Lemma 2.6.18 in van der Vaart and Wellner (1996, p. 147), after noticing that

$$\mathcal{S}_{\mathcal{K}} = \left\{ \left\{ (x,t) : hk^{-1}(t) - w_l + W_{jl}(x) > 0 \right\} : w_l \in I_l, h \in (0,1] \right\},$$

where $k^{-1}(t) = \inf\{u : k(u) \geq t\}$. Then, Lemma 2.6.15 and Lemma 2.6.18(iii) in van der Vaart and Wellner (1996, pp. 146-147) imply that the class $\mathcal{K}_{0,j}$ is a VC class for each $j = 1, ..., N_{1\varepsilon}$. Set, for each $\varepsilon > 0$, $N_{2\varepsilon j} := N(\varepsilon, \mathcal{K}_{0,j})$.

On the other hand, our assumptions on the kernel imply that

$$|K(x) - K(y)| \leq |x - y| K^*(y), \tag{B-33}$$

where $K^*(y)$ is bounded and integrable, see Hansen (2008, p. 741). Hence, for any $K(w - W(\cdot)/h) \in \mathcal{K}_0$, there exist $W_j \in \mathcal{W}$, $w_{jk} \in I$ and $h_{jk} \in (0,1]$, $j = 1, ..., N_{1\varepsilon}$ and $k = 1, ..., N_{2\varepsilon j}$, such that

$$\begin{aligned}
E\left[\left|K\left(\frac{w - W(X)}{h}\right) - K\left(\frac{w_{jk} - W_j(X)}{h_{jk}}\right)\right|^2\right] &\leq 2E\left[\left|K\left(\frac{w - W(X)}{h}\right) - K\left(\frac{w - W_j(X)}{h}\right)\right|^2\right] \\
&\quad + 2E\left[\left|K\left(\frac{w - W_j(X)}{h}\right) - K\left(\frac{w_{jk} - W_j(X)}{h_{jk}}\right)\right|^2\right] \\
&\leq C\varepsilon^2 h^{-1} E\left[K^*\left(\frac{w - W_j(X)}{h}\right)\right] + 2\varepsilon^2 \\
&\leq C_1^2 \varepsilon^2,
\end{aligned}$$

where recall W_j is such that $\|W - W_j\|_\infty \leq \varepsilon^2$, and the second inequality uses that K is bounded to conclude $|K((w - W(X))/h) - K((w - W_j(X))/h)| \leq C$. Hence, (B-32) follows. *Q.E.D.*

The following lemma extends some results in Einmahl and Mason (2005) to kernel estimators with nonparametric generated regressors.

Lemma B.4 *Let $J = I^\varepsilon = \{w \in \mathcal{Q}_\mathcal{W} : |w - v| \leq \varepsilon,\ v \in I\}$, for I a compact set of $\mathcal{Q}_\mathcal{W} \subset \mathbb{R}^d$ for some $0 < \varepsilon < 1$. Also assume that Assumptions B.1 – B.5 hold. Further, assume that Υ is a VC class, with envelope function G satisfying either*

$$\exists M > 0 :\ G(Z)\,\mathbb{I}\{W(X) \in J\} \leq M,\ a.s. \tag{B-34}$$

or for some $s > 2$

$$\sup_{(W,w)\in\mathcal{W}\times J} E[G^s(Z)\,|W(X) = w] < \infty. \tag{B-35}$$

Then we have for any $c > 0$ and $b_n \downarrow 0$, with probability 1,

$$\limsup_{n\to\infty} \sup_{c_n^\gamma \leq h \leq b_n} \frac{\sup_{\psi\in\Psi_I} \sqrt{nh^d}\,|\widehat{m}_h(\psi) - E\widehat{m}_h(\psi)|}{\sqrt{((\log(1/h^d)) \vee \log\log n)}} =: Q(c) < \infty,$$

where $c_n := c(\log n/n)$, $\gamma := 1/d$ in the bounded case (B-34) and $\gamma := (1/d - 2/ds)$ under assumption (B-35).

Proof of Lemma B.4: We only prove this lemma for the unbounded case, the proof for the bounded case follows similar steps and therefore is omitted. For any $k = 1, 2, ...,$ and $\varphi \in \Upsilon$, set $n_k := 2^k$, and

$$\varphi_k(Z_i) := \varphi(Z_i)\,\mathbb{I}\left\{G(Z_i) < c_{n_k}^{-1/s}\right\},$$

where s is as in (B-35).
For fixed h_0, $0 < h_0 < 1$, and for $n_{k-1} \leq n \leq n_k$, $w \in I$, $c_{n_k}^\gamma \leq h \leq h_0$ and $\varphi \in \Upsilon$, let

$$\widehat{m}_h^{(k)}(\psi) = \frac{1}{nh^d}\sum_{i=1}^n \varphi_k(Z_i)\,K\left(\frac{w - W(X_i)}{h}\right).$$

First, we shall prove that under our assumptions there exists a constant $Q_1(c) < \infty$, such that with probability 1,

$$\limsup_{k\to\infty} \Delta_k = Q_1(c), \tag{B-36}$$

where

$$\Delta_k := \max_{n_{k-1}\leq n\leq n_k} \sup_{c_{n_k}^\gamma \leq h \leq h_0} \frac{\sup_{\psi\in\Psi_I} \sqrt{nh^d}\left|\widehat{m}_h^{(k)}(\psi) - E\widehat{m}_h^{(k)}(\psi)\right|}{\sqrt{((\log(1/h^d)) \vee \log\log n)}}.$$

To that end, for $\psi \in \Psi_I$ and $c_{n_k}^\gamma \leq h \leq h_0$, let

$$v_h(Z_i, \psi) := \varphi(Z_i)\,K\left(\frac{w - W(X_i)}{h}\right) \text{ and } v_h^{(k)}(Z_i, \psi) := \varphi_k(Z_i)\,K\left(\frac{w - W(X_i)}{h}\right).$$

Define the class $\mathcal{V}_k(h) := \{v_h^{(k)}(\cdot,\psi) : \psi \in \Psi_I\}$ and note that for each $v_h^{(k)} \in \mathcal{V}_k(h)$,

$$\sup_{z\in\mathcal{Z}} \| v_h^{(k)}(z,\cdot) \|_{\Psi_I} := \sup_{z\in\mathcal{Z}} \sup_{\psi\in\Psi_I} |v_h^{(k)}(z,\psi)| \leq \|K\|_\infty c_{n_k}^{-1/s}.$$

Also, observe that

$$E[|v_h^{(k)}(Z,\psi)|^2] \leq E[|v_h(Z,\psi)|^2] \leq E\left[\left|\varphi(Z_i) K\left(\frac{w - W(X_i)}{h}\right)\right|^2\right].$$

Using a conditioning argument, we infer that the last term is

$$\begin{aligned}
&\leq \int E[G^2(Z)|W(X) = w']K^2\left(\frac{w-w'}{h}\right) f(w'|W)\, dw' \\
&\leq C\int h^d K^2(u) f(w - uh|W)\, du \\
&\leq C\|K\|_{2,\lambda}^2 \|f\|_{\mathcal{W},\infty} h^d =: C_1 h^d.
\end{aligned}$$

Thus,

$$\sup_{v\in\mathcal{V}_k(h)} E[|v(Z)|^2] \leq C_1 h^d. \tag{B-37}$$

Set for $j,k \geq 0$, $h_{j,k} := 2^j c_{n_k}^\gamma$ and define $\mathcal{V}_{j,k} := \{v_h^{(k)}(\cdot,\psi) : \psi \in \Psi_I \text{ and } h_{j,k} \leq h \leq h_{j+1,k}\}$. Clearly by (B-37),

$$\sup_{v\in\mathcal{V}_{j,k}} E[|v(Z)|^2] \leq C h_{j,k}^d =: \sigma_{j,k}^2. \tag{B-38}$$

Define the product class of functions $\mathcal{G}_0 := \mathcal{K}_0 \cdot \mathcal{C} \cdot \Upsilon$, where

$$\mathcal{K}_0 = \left\{x \to K\left(\frac{w - W(x)}{h}\right) : w \in I, W \in \mathcal{W}, h \in (0,1]\right\}$$

and $\mathcal{C} = \{z \to f(z) = \mathbb{I}\{G(z) < c\} : c > 0\}$. It is straightforward to prove that, for some positive constant C,

$$N_{[\cdot]}(\varepsilon, \mathcal{G}_0, \|\cdot\|_2) \leq N(C\varepsilon, \mathcal{K}_0, \|\cdot\|_2) \times N(C\varepsilon, \mathcal{C}, \|\cdot\|_2) \times N(C\varepsilon, \Upsilon, \|\cdot\|_2). \tag{B-39}$$

Hence, by Lemma B.3 and our assumptions on the class Υ, we obtain that $\log N_{[\cdot]}(\varepsilon, \mathcal{G}_0, \|\cdot\|_2) \leq C\varepsilon^{-v_0}$, for some $v_0 < 2$. Note that $\mathcal{V}_{j,k} \subset \mathcal{G}_0$, so $\log N_{[\cdot]}(\varepsilon, \mathcal{V}_{j,k}, \|\cdot\|_2) \leq C\varepsilon^{-v_0}$ also holds.

Define $l_k := \max\{j : h_{j,k} \leq 2h_0\}$ if this set is non-empty, which is obviously the case for large enough k. Also, define

$$a_{j,k} := \sqrt{n_k h_{j,k}^d \left(\left|\log(1/h_{j,k}^d)\right| \vee \log\log n_k\right)}.$$

Then, by Lemma B.1 and (B-38), for some positive constant C_3, for all k sufficiently large and all $0 \leq j \leq l_k - 1$,

$$\begin{aligned}
E\left[\sup_{v\in\mathcal{V}_{j,k}} \left|\sum_{i=1}^{n_k} \varepsilon_i v(Z_i)\right|\right] &\leq C_3\sqrt{n_k h_{j,k}^d} \\
&\leq C_3 a_{j,k}
\end{aligned} \tag{B-40}$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of iid Rademacher variables, independent of the sample $\{Z_i\}_{i=1}^n$.

By definition, $2h_{l_k,k} = h_{l_k+1,k} \geq 2h_0$, which implies that for $n_{k-1} \leq n \leq n_k$, $[c_n^\gamma, h_0] \subset [c_{n_k}^\gamma, h_{l_k,k}]$. Thus, for large enough k and for any $\rho > 1$,

$$A_k(\rho) := \{\Delta_k \geq 2A_1(C_3+\rho)\} \subset \bigcup_{j=0}^{l_k-1} \left\{ \max_{n_{k-1}\leq n\leq n_k} \left\|\sqrt{n}\mathbb{G}_n\right\|_{\mathcal{V}_{j,k}} \geq A_1(C_3+\rho)a_{j,k} \right\},$$

where C_3 is the constant in (B-40) and A_1 is the universal constant in Lemma B.2.

Set for any $\rho > 1$, $j \geq 0$ and $k \geq 1$,

$$p_{j,k}(\rho) := P\left(\max_{n_{k-1}\leq n\leq n_k} \left\|\sqrt{n}\mathbb{G}_n\right\|_{\mathcal{V}_{j,k}} \geq A_1(C_3+\rho)a_{j,k} \right).$$

Note that $\sqrt{n_k h_{j,k}^d} c_{n_k}^{1/s} = 2^{jd/2}\sqrt{n_k c_{n_k}}$, $n_k c_{n_k} \log\log n_k \geq c\,(\log\log n_k)^2$ and that $a_{j,k}^2/n_k h_{j,k}^d \geq \log\log n_k$, for all k sufficiently large. Hence, applying Talagrand's inequality, see Lemma B.2, with $\sigma_{\mathcal{G}}^2 = \sigma_{j,k}^2$, $M = \|K\|_\infty c_{n_k}^{-1/s}$ and $t = \rho a_{j,k}$, we obtain

$$\begin{aligned}
p_{j,k}(\rho) &\leq 2\left[\exp\left(-\frac{A_2\rho^2 a_{j,k}^2}{n_k C h_{j,k}^d}\right) + \exp\left(-\frac{A_2\rho a_{j,k} c_{n_k}^{1/s}}{\|K\|_\infty}\right)\right] \\
&\leq 2\left[\exp\left(-\frac{A_2\rho^2}{C}\log\log n_k\right) + \exp\left(-\frac{2^{jd/2}A_2\rho}{\|K\|_\infty}\sqrt{n_k c_{n_k}\log\log n_k}\right)\right] \\
&\leq 2\,(\log n_k)^{-\frac{A_2\rho^2}{C}} + 2\,(\log n_k)^{-\frac{A_2\rho 2^{jd/2}c^{1/2}}{\|K\|_\infty}} \\
&\leq 4\,(\log n_k)^{-\rho A_3},
\end{aligned}$$

where $A_3 := A_2\left(1/C \wedge c^{1/2}/\|K\|_\infty\right)$. Since $l_k \leq 2\log n_k$ for large enough k,

$$P(A_k(\rho)) \leq P_k(\rho) := \sum_{j=0}^{l_k-1} p_{j,k}(\rho) \leq 8(\log n_k)^{1-\rho A_3}.$$

Then, (B-36) follows from Borel-Cantelli by taking ρ sufficiently large, e.g. $\rho \geq 3/A_3$.

Next, for $n_{k-1} \leq n \leq n_k$, $w \in I$, $c_{n_k}^\gamma \leq h \leq h_0$ and $\varphi \in \Upsilon$, let

$$\overline{m}_h^{(k)}(\psi) = \frac{1}{nh^d}\sum_{i=1}^n \overline{\varphi}_k(Z_i)\,K\left(\frac{w - W(X_i)}{h}\right),$$

where $\overline{\varphi}_k(Z_i) = \varphi(Z_i)\,\mathbb{I}\left\{G(Z) \geq c_{n_k}^{-1/s}\right\}$. Then, following the same steps as in Lemma 4 in Einmahl and Mason (2005, p. 1400), we obtain, with probability 1,

$$\lim_{k\to\infty} \max_{n_{k-1}\leq n\leq n_k} \sup_{c_n^\gamma \leq h\leq h_0} \frac{\sup_{\psi\in\Psi_I} \sqrt{nh^d}\left|\overline{m}_h^{(k)}(\psi) - E\overline{m}_h^{(k)}(\psi)\right|}{\sqrt{((\log(1/h^d)) \vee \log\log n)}} = 0. \tag{B-41}$$

Finally, (B-36) and (B-41) together prove the result. *Q.E.D.*

Our next results involve uniform convergence rates for kernel estimators. For a_n and b_n as in Assumption B.5 and r as in Assumption B.4 , define

$$d_n := \sqrt{\frac{\log a_n^{-d} \vee \log\log n}{n a_n^d}} + b_n^r.$$

The following are classical smoothness conditions that are needed to control bias.

Assumption B.6 *Assumption 3 in the text holds.*

Define as in the main text $m(w|W) := E[Y|W=w]$, $w \in \mathcal{X}_W \subset \mathbb{R}^d$, and its nonparametric NW estimator is $\widehat{m}(w|W) := \widehat{T}(w|W)/\widehat{f}(w|W)$, where $\widehat{T}(w|W) := n^{-1}h^{-d}\sum_{i=1}^n Y_i K((w-W(X_i))/h)$ and $\widehat{f}(w|W) := n^{-1}h^{-d}\sum_{i=1}^n K((w-W(X_i))/h)$.

Lemma B.5 *Let Assumptions B.1 – B.6(i) hold. Then, we have,*

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{Q}_\mathcal{W}; W \in \mathcal{W}} |\widehat{f}(w|W) - f(w|W)| = O_P(d_n).$$

Proof of Lemma B.5: Write

$$\begin{aligned}\sup|\widehat{f}(w|W) - f(w|W)| &\le \sup|\widehat{f}(w|W) - E\widehat{f}(w|W)| + \sup|E\widehat{f}(w|W) - f(w|W)| \\ &\equiv I_{1n} + I_{2n},\end{aligned}$$

where henceforth the sup is over the set $a_n \le h \le b_n$, $w \in \mathcal{Q}_\mathcal{W}$ and $W \in \mathcal{W}$. An inspection of the proof of Lemma B.4 with $\varphi(\cdot) \equiv 1$ shows that we can take $I = \mathcal{Q}_\mathcal{W}$, so we obtain

$$I_{1n} = O_P\left(\sqrt{\frac{\log a_n^{-d} \vee \log\log n}{n a_n^d}}\right).$$

By the classical change of variables, Taylor expansion and Assumptions B.4 and B.6, $I_{2n} = O_P(b_n^r)$. *Q.E.D.*

The following results establish rates of convergence for kernel estimates of $m(w|W)$ and $T(w|W)$.

Lemma B.6 *Let Assumptions B.1 – B.6 hold. Then, we have*

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{Q}_\mathcal{W}; W \in \mathcal{W}} |\widehat{T}(w|W) - T(w|W)| = O_P(d_n).$$

Proof of Lemma B.6: The proof for $\widehat{T}$ follows the same arguments as for $\widehat{f}$, and hence, it is omitted. *Q.E.D.*

Define $t_n(w|W) := \mathbb{I}(f(w|W) \ge \tau_n)$ and $\widehat{t}_n(w|W) := \mathbb{I}(\widehat{f}(w|W) \ge \tau_n)$.

Lemma B.7 *Let Assumptions B.1 – B.6 hold. Then, we have*

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{Q}_W; W \in \mathcal{W}} |\widehat{m}(w|W) - m(w|W)| t_n(w|W) = O_P(\tau_n^{-1} d_n) + O_P(\tau_n^{-2} d_n^2)$$

and

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{Q}_W; W \in \mathcal{W}} |\widehat{m}(w|W) - m(w|W)| \widehat{t}_n(w|W) = O_P(\tau_n^{-1} d_n).$$

Proof of Lemma B.7: We write

$$\widehat{m}(w|W) - m(w|W) = a_n(w|W) + r_n(w|W),$$

where

$$a_n(w|W) := f^{-1}(w|W)\left[\widehat{T}(w|W) - T(w|W) - m(w|W)(\widehat{f}(w|W) - f(w|W))\right],$$

$T(w|W) := m(w|W) f(w|W)$ and

$$r_n(w|W) := -\frac{\widehat{f}(w|W) - f(w|W)}{\widehat{f}(w|W)} a_n(w|W).$$

Note that, on $f(w|W) \ge \tau_n$,

$$\begin{aligned}\frac{\widehat{f}(w|W) - f(w|W)}{\widehat{f}(w|W)} &= \frac{\widehat{f}(w|W) - f(w|W)}{f(w|W)} \frac{1}{\left(\widehat{f}(w|W)/f(w|W) - 1\right) + 1} \\ &= O_P\left(\frac{\widehat{f}(w|W) - f(w|W)}{f(w|W)}\right),\end{aligned}$$

since

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{Q}_W; W \in \mathcal{W}} \left|\frac{\widehat{f}(w|W)}{f(w|W)} - 1\right| t_n(w|W) = O_P(\tau_n^{-1} d_n).$$

Since $f^{-1}(w|W) t_n(w|W)$ is bounded by τ_n^{-1}, we obtain from previous results that

$$\sup |r_n(w|W)| t_n(w|W) = O_P(\tau_n^{-2} d_n^2).$$

For the second equality, note that

$$\widehat{m}(w|W) - m(w|W) = \left\{\frac{\widehat{T}(w|W) - T(w|W)}{\widehat{f}(w|W)} - m(w|W)\frac{\widehat{f}(w|W) - f(w|W)}{\widehat{f}(w|W)}\right\}.$$

Hence, since m is uniformly bounded, we obtain the uniform bound

$$\begin{aligned}|\widehat{m}(w|W) - m(w|W)| \widehat{t}_n(w|W) &\le \tau_n^{-1} |\widehat{T}(w|W) - T(w|W)| \\ &\quad + \tau_n^{-1} |\widehat{f}(w|W) - f(w|W)| \, |m(w|W)| \\ &= O_P(\tau_n^{-1} d_n).\end{aligned}$$

Q.E.D.

We now consider stronger versions of Assumptions B.5 and B.4 that are applicable to derivatives of kernel estimates such as

$$\dot{m}_h(\psi) = \frac{1}{nh^{d+1}} \sum_{i=1}^{n} \varphi(Z_i) \dot{K}\left(\frac{w - W(X_i)}{h}\right),$$

where $\dot{K}(w/h) = \dot{k}(w_1/h)\prod_{l=2}^{d} k(w_l/h)$, where $\dot{k}(u) = \partial k(u)/\partial u$ is the derivative of the kernel function k.

Assumption B.7 *The possibly data-dependent bandwidth h satisfies $P(a_n \leq h \leq b_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers a_n and b_n such that $b_n \to 0$ and $a_n^{d+2} n / \log n \to \infty$.*

Assumption B.8 *The kernel function $k(t) : \mathbb{R} \to \mathbb{R}$ is bounded, symmetric, twice continuously differentiable and satisfies the following conditions: $\int k(t)\, dt = 1$, $\int t^l k(t)\, dt = 0$ for $0 < l < r$, and $\int |t^r k(t)|\, dt < \infty$, for some $r \geq 2$, $\left|\partial^{(j)} k(t)/\partial t^j\right| \leq C$ and for some $v > 1$, $\left|\partial^{(j)} k(t)/\partial t^j\right| \leq C|t|^{-v}$ for $|t| > L_j$, $0 < L_j < \infty$, for $j = 1, 2$.*

Lemma B.8 *Under the conditions of Lemma B.4 but with B.7 and B.8 replacing B.5 and B.4, respectively, we have for any $c > 0$ and $b_n \downarrow 0$, with probability 1,*

$$\limsup_{n\to\infty} \sup_{c_n^{\gamma} \leq h \leq b_n} \frac{\sup_{\psi \in \Psi_I} \sqrt{nh^{d+2}} |\dot{m}_h(\psi) - E\dot{m}_h(\psi)|}{\sqrt{((\log(1/h^d)) \vee \log\log n)}} =: Q(c) < \infty.$$

Proof of Lemma B.8: The proof follows the same steps as that of Lemma B.4, and hence it is omitted. *Q.E.D.*

Appendix C Some Primitive Conditions

This section provides primitive conditions for some of the high level assumptions in the main text of the paper. These high level conditions can be classified into into three classes: 1. Assumptions on the smoothness and boundedness conditions regarding densities and regression functions; 2. Asymptotic inclusion assumptions for nonparametric estimators; and 3. Other high-level assumptions regarding properties of these estimates. Assumptions 3, 6(i) and 9 in the main text belong to class 1. Assumptions 6(ii) and 10(ii-iii) in the paper are of the type 2, while Assumptions 10(i) in the main text is an example of type 3. Primitive conditions for assumptions in the class 1 are generally model specific, see e.g. Klein and Spady (1993) for parametric generated regressors. Here we focus on primitive conditions for classes 2 and 3.

Assumption 6(ii) in the main text requires that $P(\widehat{m} \in \mathcal{T}^{\eta}) \to 1$, for some $\eta > \max(1, d/2)$. That is, it requires one to prove that

$$\sup_{x \in \mathcal{X}_X, W_1, W_2 \in \mathcal{W}} \frac{|\widehat{m}(W_1(x)|W_1) - \widehat{m}(W_2(x)|W_2)|}{\|W_1 - W_2\|_{\infty}} = O_P(1) \tag{C-42}$$

and

$$P(\widehat{m}\,(\cdot|\,W) \in C^{\eta}(\mathcal{X}_W)) \to 1 \text{ for all } W \in \mathcal{W}. \tag{C-43}$$

We now provide primitive conditions for (C-42) and (C-43) when densities are bounded away from zero. Similar arguments can be used to find primitive conditions with vanishing densities when random trimming is used, simply multiplying the rates d_{1n} and d_{2n} below by τ_n^{-1}.

To verify (C-42), we write

$$\begin{aligned}\widehat{m}\,(W_1(x)|\,W_1) - \widehat{m}\,(W_2(x)|\,W_2) = &\frac{\widehat{T}\,(W_1(x)|\,W_1) - \widehat{T}\,(W_2(x)|\,W_2)}{\widehat{f}\,(W_1(x)|\,W_1)} \\ &+ \frac{\left[\widehat{f}\,(W_2(x)|\,W_2) - \widehat{f}\,(W_1(x)|\,W_1)\right]\widehat{m}\,(W_2(x)|\,W_2)}{\widehat{f}\,(W_1(x)|\,W_1)}.\end{aligned} \tag{C-44}$$

Define $\psi := (x, W_1, W_2) \in \Psi := \mathcal{X}_X \times \mathcal{W} \times \mathcal{W}$,

$$v_{h,1}(Z_i, \psi) := Y_i \frac{h}{\|W_1 - W_2\|_\infty}\left\{K\left(\frac{W_1(x) - W_1(X_i)}{h}\right) - K\left(\frac{W_2(x) - W_2(X_i)}{h}\right)\right\}$$

and

$$\widehat{m}_{h,1}(x, \psi) := \frac{1}{nh^{d+1}}\sum_{i=1}^{n} v_{h,1}(Z_i, \psi).$$

Using (B-33) and the arguments afterwards, it is straightforward to prove that, for each $\psi \in \Psi$,

$$E[|v_{h,1}(Z_i, \psi)|^2] \leq Ch^d.$$

Then, arguing as in Lemma B.4 one can show that

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} |\widehat{m}_{h,1}(\psi) - E\widehat{m}_{h,1}(\psi)| = O_P(d_{1n}),$$

where

$$d_{1n} := \sqrt{\frac{\log a_n^{-d} \vee \log\log n}{na_n^{d+2}}}.$$

On the other hand, the typical arguments used to handle the bias yield

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} |E\widehat{m}_{h,1}(\psi)| = O_P(1).$$

A similar conclusion can be obtained for the other terms in (C-44). Then, a primitive condition for (C-42) is that $d_{1n} = O(1)$.

To give a primitive condition for (C-43) we consider the case $d = 2$, which arises in models such as the binary choice model with selection discussed in the main text. In this case, we can take $\eta = 1 + \eta_q$, with $0 < \eta_q < 1$. Then, (C-43) reduces to showing that, for $j = 1, 2$, for each $W \in \mathcal{W}$,

$$\sup_{w \neq w'} \frac{|\partial\widehat{m}\,(w|\,W)/\partial w_j - \partial\widehat{m}\,(w'|\,W)/\partial w_j|}{|w - w'|^{\eta_q}} = O_P(1). \tag{C-45}$$

To that end, define for $\psi := (w, w') \in \Psi := \mathcal{X}_W \times \mathcal{X}_W$,

$$\begin{aligned}\widehat{m}_{h,2}(\psi) &:= \frac{1}{nh^4}\sum_{i=1}^{n} Y_i \frac{h}{|w-w'|^{\eta_q}}\left\{\dot{K}_j\left(\frac{w-W(X_i)}{h}\right) - \dot{K}_j\left(\frac{w'-W(X_i)}{h}\right)\right\}, \\ &=: \frac{1}{nh^4}\sum_{i=1}^{n} v_{h,2}(Z_i, \psi),\end{aligned}$$

where $\dot{K}_1(w) = \partial k(w_1)/\partial w_1 k(w_2)$ and $\dot{K}_2(w) = \partial k(w_2)/\partial w_2 k(w_1)$, $w = (w_1, w_2)$. Then, using similar arguments as for $\widehat{m}_{h,1}$ and Lemma B.4 one can show that

$$\sup_{a_n \le h \le b_n} \sup_{\psi \in \Psi} |\widehat{m}_{h,2}(\psi) - E\widehat{m}_{h,2}(\psi)| = O_P(d_{2n}),$$

and

$$\sup_{a_n \le h \le b_n} \sup_{\psi \in \Psi} |E\widehat{m}_{h,2}(\psi)| = O_P(1),$$

where

$$d_{2n} := \sqrt{\frac{\log a_n^{-2} \vee \log\log n}{n a_n^6}}.$$

Hence, for $d = 2$ a primitive condition for Assumption 6(ii) is that $d_{2n} = O(1)$ and Assumption B.8 above hold. For a general d a similar approach can be used, provided that the corresponding kernel derivatives are of bounded variation. Similar conditions such as Assumptions 10(iii) and 13(ii) can be verified analogously.

Next, we consider primitive conditions regarding the first step estimator in Assumption 10 when $\widehat{g}_i$ is the NW estimator. Define the rate

$$d_{gn} := \sqrt{\frac{\log l_n^{-p} \vee \log\log n}{n l_n^{2p+2}}}.$$

Consider the following assumption:

Assumption C.1 *(i) The regression function g_0 is estimated by a NW kernel estimator $\widehat{g}$ with a kernel function that is $(p+1)-$times continuously differentiable, satisfies Assumption 4 with $r = \rho$ and a possibly stochastic bandwidth $\widehat{h}_{gn}$ satisfying $P(l_n \le \widehat{h}_{gn} \le u_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers l_n and u_n such that $d_{gn} = O(1)$ and $nu_n^{2\rho} \to 0$; (ii) the function g_0 and the density $f_X(\cdot)$ of X are ρ-times continuously differentiable in x, with bounded derivatives. The density $f_X(\cdot)$ is bounded away from zero. Furthermore $g_0 \in \mathcal{G} \subset C^{\eta_g}(\mathcal{X}_X)$, for some $\eta_g > p$.*

Examples of random bandwidths that satisfy our assumptions are plug-in bandwidths of the form $\widehat{h}_{gn} = \widehat{c} h_{gn}$ with $\widehat{c}$ is bounded in probability and h_n a suitable deterministic sequence. If $h_{gn} = cn^{-\delta}$, for some constant $c > 0$, then Assumption C.1(i) requires that $1/2\rho < \delta < 1/(2p+2)$, so ρ needs to be greater than $p+1$.

We now prove that under the primitive condition C.1 above, the high level Assumption 10(i) in the main text and $\|\widehat{g} - g_0\|_2 = o_P(n^{-1/4})$ hold. First, the condition $P(\widehat{g} \in C^{\eta_g}(\mathcal{X}_X)) \to 1$ can be verified as

in (C-43) above. Then, using our Theorem 3.1, without trimming and with the class $W(x) = \{x_1\}$, we obtain $|R_n(\widehat{\alpha}) - G_n(\widehat{\alpha})| = o_P(1)$ as desired. On the other hand, an application of Lemma B.7 implies

$$\|\widehat{g} - g_0\|_\infty = \sqrt{\frac{\log n}{n l_n^p}} + u_n^\rho = o_P(n^{-1/4}),$$

under the conditions on the bandwidth.

References

Ahn, H. (1997): "Semiparametric Estimation of a Single-Index Model with Nonparametrically Generated Regressors," *Econometric Theory*, 13(1), 3–31.

Ahn, H. and C. F. Manski (1993): "Distribution Theory for the Analysis of Binary Choice under Uncertainty with Nonparametric Estimation of Expectations," *Journal of Econometrics*, 56(3), 291–321.

Ahn, H. and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Problem," *Journal of Econometrics*, 58(1-2), 3–29.

Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Akritas, M. G.and I. van Keilegom (2001): "Non-parametric Estimation of the Residual Distribution," *Scandinavian Journal of Statistics*, 28(3), 549–567.

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), 43–72.

Andrews, D. W. K. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11(3), 560–596.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag, New York, 1 edn.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71(7), 655–679.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," in Handbook of Econometrics (J. J. Heckman and E. E. Leamer, eds.) volume 6, 5549–5632. Elsevier, Amsterdam.

Chen, X., O. B. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Cragg, J. G. (1971): "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica*, 39(5), 829–44.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 70(1), 33–58.

Delgado, M. A. and W. González Manteiga (2001): "Significance Testing in Nonparametric Regression Based on the Bootstrap," *Annals of Statistics*, 29(5), 1469–1507.

Einmahl, J. H. J., and D. M. Mason (2005): "Uniform in Bandwidth Consistency of Kernel-Type Function Estimators," *Annals of Statistics*, 33(3), 1380–1403.

Escanciano, J. C., D. T. Jacho-Chávez and A. Lewbel (2012): "Identification and Estimation of Semiparametric Two Step Models," Unpublished manuscript.

Escanciano, J. C., and K. Song (2010): "Testing Single-Index Restrictions with a Focus on Average Derivatives," *Journal of Econometrics*, 156(2), 377–391.

Hahn, J., and G. Ridder (2013): "The Asymptotic Variance of Semiparametric Estimators with Generated Regressors," *Econometrica*, 81(1), 315–340.

Hansen, B. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory,* 24(3), 726–748.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261–294.

Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Horowitz, J. L. and V. G. Spokoiny (2001): "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model against a Nonparametric Alternative," *Econometrica*, 69(3), 599–631.

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58(1-2), 71–120.

Ichimura, H., and L. Lee (1991): "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation"," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 3–49. Cambridge University Press.

Ichimura, H., and S. Lee (2010): "Characterization of the Asymptotic Distribution of Semiparametric M-Estimators," *Journal of Econometrics*, 159(2), 252–266.

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Klein, R., C. Shen, and F. Vella (2012) "Semiparametric Selection Models with Binary Outcomes," Unpublished manuscript.

Klein, R. and R. Spady (1993) "An efficient Semiparametric Estimator for Discrete Choice Models", *Econometrica*, 61(2), 387-421.

Lewbel, A. (2007): "Endogenous Selection or Treatment Model Estimation," *Journal of Econometrics*, 141, 777-8067.

Lewbel, A., and O. B. Linton (2007): "Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions," *Econometrica*, 75(4), 1209–1227.

Li, D. and Q. Li (2010): "Nonparametric/semiparametric Estimation and Testing of Econometric Models with Data Dependent Smoothing Parameters," *Journal of Econometrics*, 157(1), 179–190.

Li, Q. and J. M. Wooldridge (2002): "Semiparametric Estimation Of Partially Linear Models For Dependent Data With Generated Regressors," *Econometric Theory*, 18(3), 625–645.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics*, 40(2), 1132–1170.

Mammen, E., C. Rothe, and M. Schienle (2013): "Semiparametric Estimation with Generated Covariates," Unpublished manuscript.

Matzkin, R. L. (1992): "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, 60(2), 239–270.

Meng, C.-L., and P. Schmidt (1985): "On the Cost of Partial Observability in the Bivariate Probit Model," *International Economic Review*, 26(1), 71–85.

Neumeyer, N. (2004): "A Central Limit Theorem for Two-Sample U-Processes,". *Statistics & Probability Letters*, 67, 73-85.

Neumeyer, N., and Van Keilegom, I. (2010): "Estimating the Error Distribution in Nonparametric Multiple Regression with Applications to Model Testing," *Journal of Multivariate Analysis*, 101(5), 1067–1078.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

Newey, W. K. (2007): "Nonparametric Continuous/Discrete Choice Models," *International Economic Review*, 48(4), 1429–1439.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by D. McFadden, and R. F. Engle, vol. IV, pp. 2111–2245. Elsevier, North-Holland, Amsterdam.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67(3), 565–603.

Nickl, R. and B. M. Pötscher (2007). "Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type," *Journal of Theoretical Probability*, 20(2), 177–199.

Olley, S. and A. Pakes (1996). "The Dynamics Of Productivity In The Telecommunications Equipment Industry". *Econometrica*, 64(6), 1263–1297.

Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25(1), 221–247.

Pinkse, J. (2001): "Nonparametric Regression Estimation using Weak Separability," Unpublished manuscript.

Rothe, C. (2009): "Semiparametric Estimation of Binary Response Models with Endogenous Regressors," *Journal of Econometrics*, 153(1), 51–64.

Song, K. (2008): "Uniform Convergence of Series Estimators over Function Spaces," *Econometric Theory*, 24(6), 1463–1499.

Sperlich, S. (2009): "A Note on Non-parametric Estimation with Predicted Variables," *The Econometrics Journal*, 12(2), 382–395.

Stock, J. H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84(406), 567–575.

Talagrand, M. (1994): "Sharper bounds for Gaussian and empirical processes," *Annals of Probability*, 22(1), 28–76.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer Series in Statistics. Springer-Verlag, New York, 1 edn.

van de Ven, W. and B. van Praag (1981). "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," *Journal of Econometrics*, 17(2), 229–252.