Weighted KS Statistics for Inference on Conditional Moment Inequalities

Timothy B. Armstrong*
Yale University

Abstract

This paper proposes set estimators and conservative confidence regions for the identified set in conditional moment inequality models using Kolmogorov-Smirnov statistics with a truncated inverse variance weighting with increasing truncation points. The new weighting differs from those proposed in the literature in two important ways. First, this paper shows that estimators based on KS statistics with the proposed weighting function converge to the identified set at a faster rate than existing procedures based on bounded weight functions in a broad class of models. This provides a theoretical justification for inverse variance weighting in this context, and contrasts with analogous results for conditional moment equalities in which optimal weighting only affects the asymptotic variance. The results on rates of convergence of set estimators are the first such results even for the existing procedures, and involve developing the first general framework for determining consistency and rates of convergence for set estimators and confidence regions in this context. Second, the new weighting changes the asymptotic behavior, including the rate of convergence, of the KS statistic itself, requiring a new asymptotic theory in choosing the critical value. A series of examples illustrates the broad applicability of the results.

First version: March 2011. This version: December 2013.
JEL codes: C01, C14, C34
Keywords: moment inequalities, set inference, adaptive inference, irregular identification

*email: timothy.armstrong@yale.edu. Thanks to Han Hong and Joe Romano for guidance and many useful discussions, and to Liran Einav, Azeem Shaikh, Tim Bresnahan, Guido Imbens, Raj Chetty, Whitney Newey, Victor Chernozhukov, Jerry Hausman, Andres Santos, Elie Tamer, Vicky Zinde-Walsh, Alberto Abadie, Karim Chalak, Xu Cheng, Konrad Menzel, Stefan Hoderlein, Don Andrews, Peter Phillips, Taisuke Otsu, Ed Vytlacil, Xiaohong Chen, Yuichi Kitamura and participants at the 2010 California Econometrics Conference, Stanford third year seminar and MIT econometrics lunch for helpful comments and criticism. All remaining errors are my own. This paper was written with generous support from a fellowship from the endowment in memory of B.F. Haley and E.S. Shaw through the Stanford Institute for Economic Policy Research.

1 Introduction

This paper proposes a class of test statistics for conditional moment inequality models and derives new relative efficiency results for these models that show that set estimates based on these test statistics are more efficient than available methods in a certain precise sense. In doing so, this paper proposes a general set of conditions for deriving rates of convergence of set estimators in these models, and uses them to derive rates for several set estimators proposed in the literature for which rates of convergence have been unknown. While the relative efficiency comparisons are stated for set estimators (which can be interpreted as conservative confidence regions), the results can also be used to make power comparisons for the corresponding testing procedures.

Formally, these models are defined by a restriction of the form $E_P(m(W_i, \theta)|X_i) \geq 0$ almost surely. Here, m is a known parametric function, which may be vector valued (in which case the inequality is interpreted as elementwise). This setup includes many models commonly used in econometrics, including regression models with endogenously censored or missing data, selection models, and certain models of firm and consumer behavior. The problem is to estimate or perform inference on the identified set

$$\Theta_0(P) \equiv \{\theta | E_P(m(W_i, \theta)|X_i) \geq 0 \text{ a.s.}\}$$

given a sample $(X_1, W_1), \ldots, (X_n, W_n)$ from P. This paper proposes sets $\mathcal{C}_n$ that satisfy

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} P(\Theta_0(P) \subseteq \mathcal{C}_n) = 1 \tag{1}$$

for classes of probability distributions $\mathcal{P}$ restricted only by mild regularity conditions. For these sets and several others available in the literature satisfying this requirement, I derive rates of convergence of $\mathcal{C}_n$ to $\Theta_0(P)$. The results give sequences a_n, which depend on the smoothness of $\mathcal{P}$ and the method used to construct $\mathcal{C}_n$, such that

$$\sup_{P\in\mathcal{P}} P(d_H(\Theta_0(P), \mathcal{C}_n) \geq a_n) \to 0, \tag{2}$$

where d_H denotes Hausdorff distance, a notion of distance between sets (see Section 4 for the formal definition). It should be emphasized that, while several methods have been proposed for constructing sets that satisfy the coverage condition (1), the present paper is the first to determine whether or for which rates a_n these sets satisfy the rate of convergence condition

(2) for conditional moment inequalities in the general set identified case. These results show that, in a general class of models, the sets proposed here are the only ones to obtain the best rate a_n in (2) for a variety of classes $\mathcal{P}$ defined by different smoothness conditions without prior knowledge of $\mathcal{P}$. In this sense, the procedures proposed here are adaptive.

The procedures proposed in this paper apply the approach of Chernozhukov, Hong, and Tamer (2007) along with new asymptotic bounds and convergence rate results to a Kolmogorov-Smirnov (KS) statistic weighted by a truncation of the inverse of the sample variance with an increasing sequence of truncation points to obtain set estimates that satisfy (1). The increasing sequence of truncation points I propose changes the asymptotic behavior of the KS statistic relative to the bounded weightings proposed in the literature. In particular, the sequence of random processes maximized by the test statistic is no longer tight, and functional central limit theorems no longer apply. To overcome this, I use maximal inequalities that bound the supremum of a random process by a function of the maximal variance of the process. This approach, while very general, leads only to rates or to extremely conservative upper bounds for the distribution of the test statistic. From a practical standpoint, the results require using either extremely conservative critical values or arbitrary sequences going slowly to infinity. To the extent that this approach is unpalatable, the results in this paper can be taken as theoretical results guiding empirical practice on the optimal choice of test statistic and future research on the theoretical properties of these test statistics (see also Armstrong and Chan (2012) and Chetverikov (2012) for nonconservative critical values for some of these statistics that apply to a weaker notion of coverage).

I derive the rate of convergence to the identified set (sequences a_n that satisfy (2)) for these set estimates under conditions that apply to a broad class of models while still being interpretable. Since general results for rates of convergence to the identified set have not been derived for estimators based on kernel methods or KS statistics with bounded weights, I derive rates of convergence for estimators based on these existing approaches as well. For the class of models I consider, I find that using the inverse variance with increasing trunaction points as the weight function in the KS statistic results in an estimator for the identified set that has a faster rate of convergence to the identified set than the KS statistic based estimators with bounded weights proposed in the literature, and achieves the same rate of convergence as a kernel estimate with the optimal bandwidth. For classes of underlying distributions in which smoothness of two derivatives or less is imposed, these rates correspond with the upper bounds derived by Stone (1982) for estimating conditional means.

To my knowledge, these results provide the first theoretical justification for weighting

moments by their variance in conditional moment inequality problems. If the truncation parameter is allowed to increase fast enough, weighting by the variance in the KS objective function increases the rate of convergence of estimators to the identified set under the conditions I consider. While this result can be thought of as analogous to optimal weighting results for GMM, it is also related to the problem of adaptive choice of smoothing parameters in nonparametric estimation. The results in this paper show that a truncated inverse variance weighting allows the test statistic to automatically optimize a bias variance tradeoff analogous to the one faced in kernel estimation. This allows the set estimate to obtain the optimal rate of convergence without prior knowledge of the smoothness properties of the data generating process.

The results in this paper show that, in certain smoothness classes, estimators based on the methods in this paper achieve the best rate of convergence to the identified set in the Hausdorff metric. While other methods achieve the same rate of convergence if prior information is known about the shape of the conditional mean, these methods will do much worse if incorrect prior information is used to choose a different approach. A succinct way of putting this is that, among the approaches considered here, the approach based on inverse variance weighted KS statistics has the optimal minimax rate for a broad set of smoothness classes. While minimax definitions of relative efficiency are useful, they ignore the possibility that, while the inverse variance weighting approach is better in the worst case in a particular class of distributions, other approaches might do much better under more favorable data generating processes. However, the results in Section 6 show that, even in a very restrictive set of cases that are more favorable for the approach based on bounded weights, the inverse variance weighting proposed in this paper will only lose a $\log n$ term in the rate of convergence to the identified set relative to the rate of convergence using bounded weights (see, in particular, the last part of Theorem B.4). This contrasts with the polynomial differences in rates of convergence in cases where bounded weights or kernel based methods do worse.

The sets considered in this paper can be used as outwardly biased estimates of the identified set. Since they satisfy the coverage requirement (1), they can also be considered conservative confidence sets for the identified set in the sense of Chernozhukov, Hong, and Tamer (2007). With the latter interpretation, these rate results can be thought of as power results for a (conservative) setwise inference procedure. Alternatively, following Imbens and Manski (2004), one may wish to report a confidence region that satisfies only the weaker pointwise coverage requirement that $\inf_{\theta_0 \in \Theta_0(\mathcal{P})} P(\theta_0 \in \mathcal{C}_n)$ increases to one or is above a

prespecified level asymptotically. The rate results in this paper can be used to derive local power results for confidence regions satisfying the weaker Imbens and Manski (2004) coverage requirement, and may be of direct interest for these confidence regions in settings where one wishes to compare the behavior of confidence regions rather than the power of tests.

This paper relates to the recent literature on econometric models defined by moment inequalities and, in particular, conditional moment inequalities where the conditioning variable is continuously distributed. Andrews and Shi (2009), Kim (2008), Menzel (2008, 2010) and Chernozhukov, Lee, and Rosen (2009) treat this problem in different ways. The estimators of the identified set considered in the present paper are similar to those considered by Andrews and Shi (2009) and Kim (2008), but have dramatically different properties, as discussed above. The present paper also contributes to this literature by deriving the first results on rates of convergence to the identified set for these approaches or set estimators based on these approaches that apply in the general set identified case (the rate of convergence results in Kim, 2008, hold only in the point identified case except for in a few very restrictive settings). These estimators and inference procedures build on the idea of transforming conditional moment inequalities to unconditional moment inequalities, which was used by Khan and Tamer (2009) to propose estimates for a point identified model. Their setting differs from most of those considered here in that their model is point identified with a root-n rate of convergence for the point estimate. Galichon and Henry (2009) propose a similar statistic for a class of models under a different setup with possible lack of point identification.

More broadly, this paper relates to the literature on set identified models. Much of this research has been on models defined by finitely many unconditional moment inequalities. Papers that treat this problem include Andrews, Berry, and Jia (2004), Andrews and Jia (2008), Andrews and Guggenberger (2009), Andrews and Soares (2010), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2010), Romano and Shaikh (2008), Bugni (2010), Beresteanu and Molinari (2008), Moon and Schorfheide (2009), Imbens and Manski (2004) and Stoye (2009).

The test statistics in this paper are closely related to the statistics literature on tests for goodness of fit and global hypothesis tests in the gaussian white noise model. Dumbgen and Spokoiny (2001) consider a test related to the test statistic used in the present paper in a one dimensional gaussian setting, while the tests proposed by Andrews and Shi (2009) and Kim (2008) and Chernozhukov, Lee, and Rosen (2009) can be considered generalizations of tests proposed by Bierens (1982) and Bickel and Rosenblatt (1973), respectively. Power

comparisons against various types of alternatives have been considered in this literature. The present paper shows that uniform rates of convergence in the Hausdorff metric in conditional moment inequalities are determined by certain types of alternatives, which are related to those that determine minimax rates in the supremum norm in the gaussian white noise model (see, e.g. Lepski and Tsybakov, 2000). See also Ingster and Suslina (2003), Chapter 14 of Lehmann and Romano (2005) and references therein for more on these nonparametric testing problems. Interpreting these sets as estimators rather than confidence regions, the rates of convergence derived in this paper are related to rates of convergence for nonparametric estimation of the conditional mean (see, e.g. Ibragimov and Hasminskii, 1981; Stone, 1982).

In addition to the existing literature on conditional moment inequalities and set inference, some papers written after or around the same time the present paper first circulated have considered inference using similar test statistics. Armstrong and Chan (2012) and Chetverikov (2012) consider related statistics, but consider pointwise rather than setwise inference (see Imbens and Manski, 2004, for a discussion of the difference between these notions of inference).

The rest of the paper is organized as follows. In Section 2, I describe the estimation problem and estimators of the identified set, and give an informal description of some of the results in the paper and the intuition behind them. In Section 3, I state conditions under which the estimate contains the identified set with probability approaching one. In Section 4, I state conditions for consistency and rates of convergence. In Section 5, I derive rates of convergence of other estimators of the identified set under the conditions in Section 4 and compare them to rates of convergence for the estimators proposed in this paper. In Section 6, I verify the conditions of Section 4 in some examples. Section 7 reports the results of a monte carlo study of the finite sample properties of the estimators. Section 8 concludes, and an appendix contain proofs and additional results referred to in the body of the paper.

I use the following notation throughout the paper. For observations $(X_1, W_1), \ldots, (X_n, W_n)$ and a measurable function h on the sample space, $E_n h(X_i, W_i) \equiv \frac{1}{n}\sum_{i=1}^n h(X_i, W_i)$ denotes the sample mean and $E_P h(X_i, W_i)$ denotes the mean of $h(X_i, W_i)$ under the probability measure P. The support of a random variable X_i under a probability measure P is denoted $\text{supp}_P(X_i)$. I use double subscripts to denote elements of vector observations so that $X_{i,j}$ denotes the jth component of the ith observation X_i. For a vector $x \in \mathbb{R}^k$, use the notation x_{-i} to denote the vector $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)'$. Inequalities on Euclidean space refer to the partial ordering of elementwise inequality. I use $a \wedge b$ to denote the elementwise minimum and $a \vee b$ to denote the elementwise maximum of a and b. For a norm $\|\cdot\|$ on $\mathbb{R}^k$,

$\|t\|_- \equiv \|t \wedge 0\|$. Unless otherwise noted, $\|\cdot\|$ denotes the Euclidean norm.

2 Setup and Informal Description of Results

We observe iid observations $(X_1, W_1), \ldots, (X_n, W_n)$ distributed according to some probability distribution $P \in \mathcal{P}$, and wish to perform inference on the identified set $\Theta_0(P)$ of parameters $\theta \in \Theta \subseteq \mathbb{R}^d$ that satisfy the conditional moment inequalities

$$E_P[m(W_i, \theta)|X_i] \geq 0 \;\; P\text{-a.s.}$$

Here, X_i and W_i are random variables on $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_W}$ respectively, and $m : \mathbb{R}^{d_W} \times \Theta \to \mathbb{R}^{d_Y}$ is a measurable function. See Section 6 for examples of econometric models that fit into this framework. In what follows, $\bar{m}(\theta, x, P)$ will denote a version of $E_P[m(W_i, \theta)|X_i = x]$.

I consider inference on $\Theta_0(P)$ using a standard deviation weighted KS statistic defined as follows. Let $\mathcal{G}$ be a class of functions from $\mathbb{R}^{d_X}$ to $\mathbb{R}_+$. Let $\mu_{P,j}(\theta, g) = E_P m_j(W_i, \theta) g_j(X_i)$ and $\sigma_{P,j}(\theta, g) = \{E_P[m_j(W_i, \theta) g_j(X_i)]^2 - [E_P m_j(W_i, \theta) g_j(X_i)]^2\}^{1/2}$ and define the sample analogues $\hat{\mu}_{n,j}(\theta, g) = E_n m_j(W_i, \theta) g_j(X_i)$ and $\hat{\sigma}_{n,j}(\theta, g) = \{E_n[m_j(W_i, \theta) g_j(X_i)]^2 - [E_n m_j(W_i, \theta) g_j(X_i)]^2\}^{1/2}$. Since the functions in $\mathcal{G}$ are nonnegative, $E_P[m(W_i, \theta)|X_i = x] \geq 0$ for all x implies that $\mu_{P,j}(\theta, g) = E_P m_j(W_i, \theta) g_j(X_i)$ is nonnegative for all g and j. The KS statistics in this paper are designed to be positive and large in magnitude when one of these moments is small (negative and large in magnitude). For a fixed function $S : \mathbb{R}^{d_Y} \to \mathbb{R}_+$ chosen by the researcher, the KS statistic is defined as

$$T_n(\theta) = \sup_{g \in \mathcal{G}} S\left(\frac{\hat{\mu}_{n,1}(\theta, g)}{\hat{\sigma}_{n,1}(\theta, g) \vee \sigma_n}, \ldots, \frac{\hat{\mu}_{n,d_Y}(\theta, g)}{\hat{\sigma}_{n,d_Y}(\theta, g) \vee \sigma_n} \right)$$

where σ_n is a decreasing sequence of truncation points. Here, S is a function that is positive and large in magnitude when one of its arguments is negative and large in magnitude. Possible choices include $t \mapsto \sqrt{\sum_{j=1}^{d_Y} (\min\{t_k, 0\})^2}$ or, more generally, any function that satisfies Assumption 3.3, given in Section 3. If $T_n(\theta)$ is positive and large in magnitude, this is evidence that $\mu_{P,j}(\theta, g)$ is negative for some j and g, so that θ is not in the identified set.

The set estimates in this paper invert this test statistic using critical values that control the probability of false rejection uniformly over Θ, as proposed by Chernozhukov, Hong, and Tamer (2007). For some data dependent value $\hat{c}_n$, the estimator $\mathcal{C}_n(\hat{c}_n)$ for the identified set

is defined as

$$\mathcal{C}_n(\hat{c}_n) \equiv \left\{ \theta \in \Theta \middle| \frac{\sqrt{n}}{\sqrt{\log n}} T_n(\theta) \leq \hat{c}_n \right\}.$$

The results in this paper place various conditions on the user defined sequences $\hat{c}_n$ and σ_n. While weaker conditions suffice for the results in this paper, one can ensure that these conditions hold by choosing, for example, $\hat{c}_n = \sqrt{\log \log n}$ and $\sigma_n = \sqrt{(\log \log n)(\log n)/n}$ (here, the $\log \log n$ sequence is arbitrary and can be replaced by any slowly increasing sequence).

2.1 Intuition for the Results

To describe the intuition behind the results in this paper, consider a special case of the KS statistic based estimators I treat in this paper applied to a particular model. Consider an interval regression model, in which we posit a linear conditional mean for a latent variable W_i^* given an observed variable X_i, $E_P(W_i^*|X_i) = \theta_1 + X_i'\theta_{-1}$, but only observe intervals known to contain W_i^*. Here, X_i is a continuously distributed random variable on $\mathbb{R}^{d_X}$. While surveys that elicit interval responses are an obvious application, this encompasses other forms of incomplete data including selection models and missing data (see Section B.3 for an example). I give a more thorough treatment of this model in Sections 6.1 and 6.2. To keep things simple, suppose that we only observe a one sided interval containing W_i^*. That is, we observe a variable W_i^H known to be greater than or equal to W_i^*. Then the problem can be defined formally as estimating or performing inference on the identified set $\Theta_0(P)$ of values of $\theta = (\theta_1, \theta_{-1})$ that satisfy $E_P(W_i^H|X_i) \geq \theta_1 + X_i'\theta_{-1}$.

To fix ideas, consider using the KS statistic defined above with the class of functions $\mathcal{G}$ given by the set of indicator functions $I(\|X_i - s\| \leq h)$ with s ranging over real numbers and h ranging over nonnegative reals. For some positive weighting function $\omega_n(\theta, s, h)$, define the KS statistic $T_{n,\omega}(\theta) = \sup_{s,h} |\omega_n(\theta, s, h) E_n(W_i^H - \theta_1 - X_i'\theta_{-1}) I(\|X_i - s\| \leq h)|_-$ where $|r|_- \equiv |r \wedge 0|$. This corresponds to the KS statistic defined above with $S(r) = |r|_-$ and with the weight function $\frac{1}{\hat{\sigma}(\theta,s,h) \vee \sigma_n}$ (here $\hat{\sigma}(\theta, s, h) \equiv \{E_n[(W_i^H - \theta_1 - X_i'\theta_{-1}) I(\|X_i - s\| \leq h)]^2 - [E_n(W_i^H - \theta_1 - X_i'\theta_{-1}) I(\|X_i - s\| \leq h)]^2\}^{1/2}$) replaced by an arbitrary weight function $\omega_n(\theta, s, h)$.

Following Andrews and Shi (2009) and Kim (2008), one can show that $T_{n,\omega}(\theta)$ will converge at a $\sqrt{n}$ rate under regularity conditions if $\omega_n(\theta, s, h)$ is bounded uniformly in n. However, since the variance of the moment indexed by (θ, s, h) will be arbitrarily small when h is small (X_i has a continuous distribution), setting $\omega_n(\theta, s, h)$ equal to $\frac{1}{\hat{\sigma}(\theta,s,h) \vee \sigma_n}$ gives a

weight function that increases without bound as σ_n decreases with the sample size. This decreases the rate of convergence from $\sqrt{n}$ to $\sqrt{n/\log n}$. The estimators of the identified set I propose in this paper are based on inverting KS tests with this weighting function, where $\sqrt{n/\log n}T_{n,\omega}(\theta)$ is compared to a critical value $\hat{c}_n$ that is bounded or increases slowly. With a bounded weight function that does not increase with n, $\sqrt{n}T_{n,\omega}(\theta)$ is compared to a bounded or slowly increasing critical value.

While the results in this paper apply to rates of convergence in the Hausdorff metric, much of the intuition for the results in this paper can be exposited in the context of a single sequence of local alternatives. Consider a value of θ such that the regression line $\theta_1 + X_i'\theta_{-1}$ is tangent to the conditional mean $E_P(W_i^H|X_i)$ at a single point x_0, and X_i has a density bounded away from zero and infinity near x_0. This will typically be the case at least for some, if not all, elements on the boundary of the identified set. The results are the same if x_0 is replaced by a finite set, and can be extended to cases of set identification at infinity or at a finite boundary in which x_0 may be infinite and the density of X_i may go to zero or infinity near x_0 by transforming the model (see Section B.3). Suppose that, for some $\alpha > 0$,

$$E_P(W_i^H - \theta_1 - X_i'\theta_{-1}|X_i = x) \text{ behaves like } \|x - x_0\|^\alpha \tag{3}$$

as $\|x - x_0\|$ increases for x close to x_0. If $E_P(W_i^H|X_i = x)$ is twice differentiable and x_0 is on the interior of the support of X_i, this will hold with $\alpha = 2$, and a Lipschitz condition on $E_P(W_i^H|X_i = x)$ leads to $\alpha = 1$. While other values of α appear less natural in this context, they are common in irregularly identified cases such as the selection model considered in Section B.3.

Consider the power of KS tests against local alternatives of the form $\theta_n = (\theta_{1,0}+a_n, \theta_{-1,0})$, where $\theta_0 = (\theta_{1,0}, \theta_{-1,0})$ is on the boundary of the identified set and satisfies the above conditions for some α. Since moments centered at x_0 will have more negative expected values under this sequence of alternatives, the moments with the most power for detecting this sequence of local alternatives will be those indexed by $s = x_0$ and some sequence of values of h. For both classes of weight functions, the order of magnitude of the value of h that indexes the moment with the most power will be determined by a tradeoff between variance and the magnitude of the expectation. The KS objective function evaluated at some (θ, s, h) is the sum of a mean zero term $(E_n - E_P)(W_i^H - \theta_1 - X_i'\theta_{-1})I(\|X_i - s\| \leq h)$ and a drift term $E_P(W_i^H - \theta_1 - X_i'\theta_{-1})I(\|X_i - s\| \leq h)$. Under (θ_n, s, h) with $s = x_0$, the

drift term is

$$E_P(W_i^H - \theta_{1,n} - X_i'\theta_{-1,n})I(\|X_i - x_0\| \leq h) = E_P(W_i^H - \theta_{1,0} - a_n - X_i'\theta_{-1,0})I(\|X_i - x_0\| \leq h)$$
$$= E_P(W_i^H - \theta_{1,0} - X_i'\theta_{-1,0})I(\|X_i - x_0\| \leq h) - a_n E_P I(\|X_i - x_0\| \leq h). \quad (4)$$

Some calculation shows that the first term in the above display is of order $h^{\alpha+d_X}$, while the second term in the above display is of order $-a_n h^{d_X}$.

Which values of h result in the corresponding moment having power depends on the mean zero term and the scaling, which depends on the weight function. First, consider the increasing sequence of weight functions given by $\omega_n(\theta_n, x_0, h) = \frac{1}{\hat{\sigma}(\theta_n, x_0, h) \vee \sigma_n}$. In this case, the $\mathcal{O}(h^{\alpha+d_X} - a_n h^{d_X})$ term in the above display will be divided by $\hat{\sigma}(\theta_n, x_0, h) \vee \sigma_n$, which, for σ_n small enough, will be approximately equal to the standard deviation of the moment indexed by (θ_n, x_0, h), which is of order $h^{d_X/2}$, and compared to a critical value that is of order $(n/\log n)^{-1/2}$ (the mean zero term will be of the same order of magnitude as the normalized critical value, so it will not affect the power calculation). Thus, the local alternative indexed by a_n will be detected if $\mathcal{O}\left(\frac{h^{\alpha+d_X} - a_n h^{d_X}}{h^{d_X/2}}\right) \leq -\mathcal{O}(n/\log n)^{-1/2}$ for some h. The left hand side is minimized when h is equal to a small constant times $a_n^{1/\alpha}$, which leads to the left hand side being of order $-a_n^{(d_X+2\alpha)/(2\alpha)}$. This will be less than the $-\mathcal{O}(n/\log n)^{-1/2}$ critical value if a_n is greater than or equal to a large enough constant times $(n/\log n)^{-\alpha/(d_X+2\alpha)}$.

Now consider using a KS statistic with a bounded weight function. The drift term will still be of order $h^{\alpha+d_X} - a_n h^{d_X}$ before being multiplied by the weight function, but, since the weight function is bounded uniformly in n, weighting will not increase the order of magnitude of the drift term. In this case, the KS statistics will be compared to a critical value of order $n^{-1/2}$, and the mean zero term will be of a smaller order of magnitude, so that the local alternative indexed by a_n will be detected if $\mathcal{O}(h^{\alpha+d_X} - a_n h^{d_X}) \leq -\mathcal{O}(n^{-1/2})$. As before, the left hand side is minimized when h is equal to some small constant times $a_n^{1/\alpha}$. In this case, this leads to the left hand side being of order $a_n^{(d_X+\alpha)/\alpha}$. This will be less than the $-\mathcal{O}(n^{-1/2})$ critical value of a_n is greater than some large constant times $n^{-\alpha/(2d_X+2\alpha)}$. This is a slower rate of convergence than the $(n/\log n)^{-\alpha/(d_X+2\alpha)}$ rate for estimators that use the inverse variance weighting with increasing truncation points.

The increase in power from weighting low variance moments by the inverse of their standard deviations comes from the fact that local alternatives violate the conditional moment inequality on a shrinking subset of the support of the conditioning variable. If we require that the weight be bounded uniformly in n, low variance moments cannot be weighted properly

because the inverse of the standard deviation will be greater than the truncation point.

3 Coverage of the Identified Set

In this section, I state conditions under which the set $\mathcal{C}_n(\hat{c}_n)$ contains the identified set $\Theta_0(P)$ with probability approaching one. Under these conditions, these estimates control the probability of falsely concluding that the data are not consistent with some parameter value. I show that the probability that the estimate contains the identified set converges to one uniformly in any class of probability distributions $\mathcal{P}$ that satisfy a set of assumptions stated below. Since these conditions do not restrict the smoothness of the conditional mean $\bar{m}(\theta, x, P)$ or the distribution of the conditioning variable, this shows that the estimator is robust to many types of data generating processes, at least in the sense of controlling the probability of type I error (of the corresponding testing procedure). In contrast, rates of convergence derived later in the paper depend on additional smoothness conditions on the data generating process.

I make the following assumptions.

Assumption 3.1. *$g_j(X_i) \geq 0$ P-a.s. for j from 1 to d_Y for $g \in \mathcal{G}$ and $P \in \mathcal{P}$.*

Assumption 3.1 states that the conditional moment inequalities are integrated against nonnegative functions, so that going from conditional moment inequalities to unconditional moment inequalities does not change the sign of the moment inequalities.

Assumption 3.2. *For some fixed $\overline{Y} \geq 0$, we have the following.*

1. *For j from 1 to d_Y, define the classes of functions $\mathcal{F}_{j,1} = \{sm_j(W_i, \theta)g_j(X_i) + t | \theta \in \Theta, g \in \mathcal{G}, s, t \in [-(\overline{Y} \vee 1), \overline{Y} \vee 1]\}$ and $\mathcal{F}_{j,2} = \{(sm_j(W_i, \theta)g_j(X_i) + t)^2 | \theta \in \Theta, g \in \mathcal{G}, s, t \in [-(\overline{Y} \vee 1), \overline{Y} \vee 1]\}$. Suppose that, for j from 1 to d_Y and $i = 1, 2$, $\sup_Q N(\varepsilon, \mathcal{F}_{j,i}, L_1(Q)) \leq A\varepsilon^{-V}$ for $0 < \varepsilon < 1$ for some $A, V > 0$, where the supremum over Q is over all probability measures and $N(\varepsilon, \mathcal{F}_{j,i}, L_1(Q))$ is the L_1 covering number defined in Pollard (1984).*

2. *$|m_j(W_i, \theta)g_j(X_i)| \leq \overline{Y}$ P-a.s. for j from 1 to d_Y for all $P \in \mathcal{P}$.*

Part (1) of Assumption 3.2 bounds the complexity of the classes of functions involved so that empirical process methods can be used. This condition will hold if the corresponding bounds hold for $\mathcal{G}$ and $\{w \mapsto m(w, \theta) | \theta \in \Theta\}$ individually. In Section A.5 of the appendix, I

state sufficient conditions for Assumption 3.2, and verify them for some classes of functions $\mathcal{G}$ and the moment functions m from the examples in Section 6. See Pollard (1984) or van der Vaart and Wellner (1996) for definitions and additional sufficient conditions for these covering number bounds.

Part (2) of Assumption 3.2 is natural in many cases, such as models defined by quantile restrictions. In other cases, it restricts some variables to a finite interval. While this is clearly stronger than just bounding some of the moments of $m_j(W_i, \theta)g_j(X_i)$, when combinded with part (1) of this assumption, it leads to rates of convergence that are uniform in θ and g and in the underlying distribution with no additional assumptions on the shape of the conditional mean or variance or the smoothness of the cdfs of the random variables.

I make the following assumption on the function S. These assumptions are satisfied by the function $t \to \|t\|_- \equiv \|t \wedge 0\|$ for any norm $\|\cdot\|$ on Euclidean space.

Assumption 3.3. *$S : \mathbb{R}^{d_Y} \to \mathbb{R}_+$ satisfies (i) $S(t) > 0$ iff. $t_j < 0$ for some j and (ii) for some positive constants $K_{S,1}$ and $K_{S,2}$, we have, for any $c > 0$, $S(t) \geq c \implies t_j \leq -cK_{S,1}$ some j and $S(t) \leq c \implies t_j \geq -cK_{S,2}$ all j.*

Under these conditions with $\hat{c}_n$ chosen large enough and σ_n decreasing slowly enough, the probability of type I error (in the sense of the estimate not containing the identified set) converges to zero uniformly in $P \in \mathcal{P}$.

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold with $\sigma_n\sqrt{n/\log n} \geq K$ and $\hat{c}_n \geq K$ with probability approaching one uniformly in $P \in \mathcal{P}$. If K is larger than some constant that depends only on V and $\overline{Y}$ in Assumption 3.2*

$$\inf_{P \in \mathcal{P}} P(\Theta_0(P) \subseteq \mathcal{C}_n(\hat{c}_n)) \overset{n\to\infty}{\to} 1.$$

While interesting as a theoretical result, Theorem 3.1 does not explicitly state a value for the constant K, making it difficult to use in practice. In fact, the constant can be calculated by carefully following the maximal inequality bounds in the proof. While this gives a feasible critical value that can be used for inference, the resulting bound is extremely conservative. This critical value is given in Theorem A.2 in Section A.2 in the appendix.

While this gives a feasible critical value that can be used for inference on the identified set under very general conditions, the resulting critical value will typically be too conservative to be of use in practice unless the sample size is extremely large. While general methods for obtaining less conservative critical values are not yet available for the problem of inference

on the identified set considered in this paper, a few papers written after or around the same time the first draft of the present paper circulated have made important contributions in proposing less conservative critical values for the related problem of inference on points in the identified set (see Armstrong and Chan, 2012; Chetverikov, 2012). While these results do not apply in our setting, they do give some indication of how conservative the bound is in Theorem A.2. For example, applying the bound in Theorem A.2 to critical values for pointwise inference (θ fixed) with the class $\mathcal{G}$ given by $\{x \mapsto I(s < x < s+t) | s, t \in \mathbb{R}^{d_X}\}$ gives a critical value that is conservative by a factor of just over 128 compared to critical values based on asymptotic distribution results in Armstrong and Chan (2012) (see the discussion in Section A.2).

4 Consistency and Rates of Convergence

To get consistency and rates of convergence, we need additional assumptions that lead to $E_P m(W_i, \theta) g(X_i)$ being large enough for parameters far from the identified set. Consistency and rate of convergence results are stated for the Hausdorff metric on sets. For a metric d on Θ, define the Hausdorff distance $d_H(A, B)$ between any two sets A and B by

$$d_H(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\}.$$

Here, I define d_H to be the Hausdorff distance that arises when d is defined to be the metric associated with the Euclidean norm. Note that under the conditions of Theorem 3.1, $\Theta_0(P) \subseteq \mathcal{C}_n(\hat{c}_n)$ with probability approaching one uniformly in $P \in \mathcal{P}$. When this holds, $\sup_{b \in \Theta_0(P)} \inf_{a \in \mathcal{C}_n(\hat{c}_n)} d(a, b) = 0$ so that we just need to bound $\sup_{a \in \mathcal{C}_n(\hat{c}_n)} \inf_{b \in \Theta_0(P)} d(a, b)$.

4.1 Consistency

The following assumption states that for θ bounded away from the identified set, some moment $E_P m_j(W_i, \theta) g_j(X_i)$ is negative and is bounded away from zero. This assumption is used to obtain consistency, and is in general stronger than what would be needed for power against fixed points in $\Theta \backslash \Theta_0(P)$, since consistency in the sense of convergence under some metric on sets requires that the power against fixed alternatives be uniform in alternatives bounded away from the identified set in this metric.

Assumption 4.1. *For every $\varepsilon > 0$, there exists a $\delta > 0$ such that, for all $P \in \mathcal{P}$,*

$d_H(\theta, \Theta_0(P)) > \varepsilon$ implies that there exists a $g \in \mathcal{G}$ such that $E_P m_j(W_i, \theta) g_j(X_i) < -\delta$ for some j.

Theorem 4.1. *Suppose that Assumption 4.1 and the assumptions of Theorem 3.1 hold, and that $\sup_{P \in \mathcal{P}} P(\hat{c}_n \sqrt{(\log n)/n} > \eta) \to 0$ for all $\eta > 0$. Then, for every $\varepsilon > 0$,*

$$\sup_{P \in \mathcal{P}} P(d_H(\Theta_0(P), \mathcal{C}_n(\hat{c}_n)) > \varepsilon) \overset{n \to \infty}{\to} 0.$$

4.2 Rates of Convergence under High Level Conditions

While the focus of this paper is the interpretable conditions for rates of convergence of the estimate of the identified set given in Section 4.3, I first present a result using a high level condition. The derivations of the rates of convergence in Section 4.3 use this result along with additional arguments relating the variance and expectation of the moments to the conditions in this section. The conditions in this section also encompass the case where local alternatives violate the conditional moment inequality on a non-shrinking set, leading to $\sqrt{n/\log n}$ convergence (such as Assumption B.7 for the application in Section B.3), and it is instructive to compare the verification of the conditions in this section under these two types of set identification.

The next assumption is a high level assumption that incorporates both the variance and expectation of the moments defined by each $g \in \mathcal{G}$. The assumption is similar to the polynomial minorant condition in Chernozhukov, Hong, and Tamer (2007).

Assumption 4.2. *For some positive constants C, ψ, γ, and δ with $\psi < 1$, we have, (i) for all $P \in \mathcal{P}$ and $\theta \in \Theta$ with $d_H(\theta, \Theta_0(P)) \leq \delta$,*

$$\inf_{g,j} \frac{\mu_{P,j}(\theta, g)}{\sigma_{P,j}(\theta, g) \vee d_H(\theta, \Theta_0(P))^{\psi/\gamma}} \leq -C d_H(\theta, \Theta_0(P))^{1/\gamma}$$

where the infimum is taken over $g \in \mathcal{G}$ and $j \in \{1, \ldots, d_Y\}$ and (ii) $\sigma_n (n/\log n)^{\psi/2}$ is bounded uniformly in P.

Part (ii) of this assumption states that the cutoff σ_n must go to zero fast enough that the moments with the most identifying power relative to their variance are scaled by their standard deviation. This condition can be made to hold for all $\psi < 1$ by choosing, e.g., $\sigma_n = \sqrt{(\log \log n)(\log n)/n}$, as discussed in Section 2.

The following theorem gives rates of convergence to the identified set under this assumption.

Theorem 4.2. *Suppose that Assumption 4.1 and 4.2 hold, and that Assumptions 3.1, 3.2 and 3.3 hold with σ_n and $\hat{c}_n$ chosen to satisfy the requirements of Theorems 3.1 and 4.1. Then, for some large B that does not depend on P,*

$$\sup_{P \in \mathcal{P}} P\left(\left(\frac{n}{\hat{c}_n^2 \log n}\right)^{\gamma/2} d_H(\mathcal{C}_n(\hat{c}_n), \Theta_0(P)) > B\right) \overset{n\to\infty}{\to} 0.$$

The results in the next section use Theorem 4.2 along with additional arguments to formalize the intuition described in Section 2.1. The balancing of the mean and variance described in Section 2.1 plays out through the ratio of the mean $\mu_{P,j}(\theta, g)$ and the standard deviation $\sigma_{P,j}(\theta, g)$ in Assumption 4.2. This determines the best attainable value of γ in Assumption 4.2. If a sequence of g functions can be found such that, as the distance of θ to the identified set decreases, the magnitude of $\mu_{P,j}(\theta, g)$ decreases much more slowly than $\sigma_{P,j}(\theta, g)$, the left hand side of the display in Assumption 4.2 will be large in magnitude, so that the condition will hold with a larger value of γ. It is useful to contrast this with the case where local alternatives violate one of the conditional moment inequalities on a non-shrinking set. In this case, g can be chosen to be some fixed function that is positive only on this set. This leads to $\sigma_{P,j}(\theta, g)$ being fixed while $\mu_{P,j}(\theta, g)$ typically goes to zero at a rate proportional to $d_H(\theta, \Theta_0(P))$, so that Assumption 4.2 holds with $\gamma = 1$, and Theorem 4.2 gives a $\sqrt{n/\log n}$ rate of convergence for the set estimator (see the proof of the part of Theorem B.4 that applies under Assumption B.7 for more details). In cases like those described in Section 2.1, the best attainable ratio of $\mu_{P,j}(\theta, g)$ to $\sigma_{P,j}(\theta, g)$ depends on smoothness properties of the data generating process and leads to a smaller γ and a slower rate of convergence. The results in the next section cover this case.

4.3 Interpretable Conditions for Rates of Convergence

Assumption 4.2 is a high level condition that incorporates both the expectation and variance of each g function. The next assumptions place restrictions on the shape of the conditional mean $\bar{m}(\theta, x, P) = E_P(m(X_i, \theta)|X_i = x)$ as a function of x and θ that can be used to verify Assumption 4.2. These conditions shed light on how the shape of the data generating process and $\bar{m}(\theta, x, P)$ as a function of θ and x determine the rate of convergence, and are easier to verify in many applications. Once consistency is established, these assumptions only need to hold for $d_H(\theta, \Theta_0(P)) < \varepsilon$ for some $\varepsilon > 0$.

Assumption 4.3. *$\bar{m}(\theta, x, P)$ is differentiable in θ with derivative $\bar{m}_\theta(\theta, x, P)$ that is continuous as a function of θ uniformly in (θ, x, P)*

Assumption 4.4. *For some $\eta > 0$ and $C > 0$, we have, for all $\theta \in \Theta \backslash \Theta_0(P)$, there exists a $j_0(\theta, P)$, $\theta_0(\theta, P)$ and $x_0(\theta, P)$ such that*

$$\bar{m}_{\theta, j_0(\theta,P)}(\theta_0(\theta, P), x_0(\theta, P), P)(\theta - \theta_0(\theta, P)) \leq -\eta \|\theta - \theta_0(\theta, P)\|,$$

$\bar{m}_{j_0(\theta,P)}(\theta_0(\theta, P), x_0(\theta, P), P) = 0$*, and, for* $\|x - x_0\| < \eta$*,*

$$|\bar{m}_{j_0(\theta,P)}(\theta_0(\theta, P), x, P) - \bar{m}_{j_0(\theta,P)}(\theta_0(\theta, P), x_0(\theta, P), P)| \leq C\|x - x_0(\theta, P)\|^\alpha.$$

The first part of Assumption 4.4 states that, for θ close to the identified set, there is some element in the identified set such that that moving from this element to θ corresponds to moving some index of the conditional mean downward. This assumption restricts the angle between the path from θ to some point on the identified set and the directional derivative of the conditional mean for θ along this path. To see that the first part of Assumption 4.4 comes from a condition on the magnitude of the derivative of the conditional mean with respect to θ and the angle of between the derivative and the difference between θ and some point on the identified set, note that, letting ϕ be the angle between $\bar{m}_{\theta, j_0(\theta,P)}(\theta_0(\theta, P), x_0(\theta, P), P)$ and $\theta - \theta_0(\theta, P)$,

$$\begin{aligned}
&\bar{m}_{\theta, j_0(\theta,P)}(\theta_0(\theta, P), x_0(\theta, P), P)(\theta - \theta_0(\theta, P)) \\
&= \|\bar{m}_{\theta, j_0(\theta,P)}(\theta_0(\theta, P), x_0(\theta, P), P)\| \|\theta - \theta_0(\theta, P)\| \cos\phi.
\end{aligned}$$

Thus, the first part of Assumption 4.4 will be satisfied if $\|\bar{m}_{\theta, j_0(\theta,P)}(\theta_0(\theta, P), x_0(\theta, P), P)\|$ is bounded away from zero and $\cos\phi$ is negative and bounded away from zero. As shown in a counterexample in Section A.4, the rate of convergence can be slower when such a condition is not placed on the angle ϕ.

The second part of Assumption 4.4 is a restriction on the shape of the conditional mean as a function of x for θ on the boundary of the identified set. Combining this with the first part of the assumption determines which functions in $\mathcal{G}$ have power under local alternatives. As verified for several models in Section 6, this typically follows from Hölder conditions or conditions on the first two derivatives of conditional means or quantiles of variables in the data, leading to some value of α between zero and 2, or from conditions on densities and conditional means near the boundary of the support of the conditioning variable, which can

lead to larger values of α after a transformation of the data.

The next assumption states that, for any $P \in \mathcal{P}$, all points must either be outside of the support of X_i under P, or have sufficient probability mass nearby. While this assumption rules out X_i having infinite support or having a density that goes to zero near the boundary of its support, these cases can typically be handled by transforming the data to make this assumption hold. I do this for one application in Section B.3.

Assumption 4.5. *For some $\eta > 0$, we have, for all $P \in \mathcal{P}$ and all $\varepsilon > 0$, $P(\|X_i - x\| \leq \varepsilon)/\varepsilon^{d_X} \geq \eta$ for all x on the support of X_i.*

The next assumption ensures that the set of functions $\mathcal{G}$ is rich enough to contain functions that behave like indicators of small sets. This assumption holds for any class that contains indicator sets of the open balls for any norm on $\mathbb{R}^{d_X}$, or, for any nonnegative bounded kernel function $k : \mathbb{R}^{d_X} \to \mathbb{R}_+$ with finite support and $k(x)$ bounded away from zero near $x = 0$, the class $\{x \mapsto k((x-t)/h) | t \in \mathbb{R}, h \geq 0\}$ that contains all dilations and translations of the kernel function k (in the following assumption, the upper bound is taken to be one, $C_{\mathcal{G},1}$ can be taken to be the positive lower lower bound on the kernel function in a neighborhood of zero, and $C_{\mathcal{G},2}$ can be taken to be such that $\{x | \|x\| \leq C_{\mathcal{G},2}\}$ is contained in this neighborhood of zero).

Assumption 4.6. *The functions in $\mathcal{G}$ are uniformly bounded and for some constants $0 < C_{\mathcal{G},1} < 1$ and $0 < C_{\mathcal{G},2} < 1$, we have that, for all $s \in \mathbb{R}^{d_X}$ and $t \geq 0$, $\mathcal{G}$ contains a function g such that $C_{\mathcal{G},1} I(\|X_i - s\| < C_{\mathcal{G},2} t) \leq g(X_i) \leq I(\|X_i - s\| < t)$.*

The next theorem gives rates of convergence under these assumptions.

Theorem 4.3. *Suppose that Assumptions 4.3, 4.4, 4.5 and 4.6 hold. Then part (i) of Assumption 4.2 holds with $\gamma = 2\alpha/(d_X + 2\alpha)$ and $\psi = d_X/(d_X + 2\alpha)$.*

Applying Theorem 4.2, this gives a $(n/\log n)^{\alpha/(d_X+2\alpha)}$ rate of convergence as long as the cutoff point σ_n for the standard deviation weighting decreases at least as quickly as $((\log n)/n)^{\psi/2} = (n/\log n)^{d_X/(2d_X+4\alpha)}$, but slightly more slowly than $((\log n)/n)^{1/2}$, so that σ_n satisfies the conditions of Theorem 3.1. As discussed in Section 2, one can ensure that both of these conditions hold by choosing $\sigma_n = \sqrt{(\log\log n)(\log n)/n}$.

5 Rates of Convergence for Other Estimators

In order to compare the estimators based on KS statistics with increasing variance weights proposed in this paper to estimation procedures based on kernels or KS statistics with

bounded weights, we need rates of convergence for these estimators as well. Since these results are not available in the literature, I derive these results in this section. The results show that, in contrast to the truncated inverse weighting approach which achieves a $(n/\log n)^{\alpha/(d_X+2\alpha)}$ rate, an approach with bounded weights leads to a slower rate of $n^{\alpha/(2d_X+2\alpha)}$, and a kernel approach leads to the same rate as the truncated inverse weighting approach only if the rate for the optimal bandwidth is known, and otherwise leads to set estimates that converge to the identified set more slowly.

5.1 Bounded Weight Functions

Consider a set estimate based on a KS statistic similar to the ones considered so far, but with the weight function $1/(\hat{\sigma}(\theta,g) \vee \sigma_n)$ replaced by some bounded weight function $\omega_n(\theta,g) = (\omega_{n,1}(\theta,g),\ldots,\omega_{n,d_Y}(\theta,g))$. Here, $\omega_n(\theta,g)$ is unrestricted, except for the requirement that, for some $\overline{\omega}$ we have $\|\omega_n(\theta,g)\| \leq \overline{\omega}$ for all n, θ, and g. Define

$$T_{n,\omega}(\theta) \equiv \sup_{g\in\mathcal{G}} S\left(\omega_{n,1}(\theta,g)\hat{\mu}_{n,1}(\theta,g),\ldots,\omega_{n,d_Y}(\theta,g)\hat{\mu}_{n,d_Y}(\theta,g)\right).$$

Following Andrews and Shi (2009) (with additional conditions to control the complexity of $m_j(W_i,\theta)g_j(X_i)$ over θ as well as g), $T_{n,\omega}(\theta)$ will converge at a $\sqrt{n}$ rate, so define the estimate of the identified set for critical value $\hat{c}_n$ to be

$$\mathcal{C}_{n,\omega}(\hat{c}_n) \equiv \left\{\theta\in\Theta \middle| \sqrt{n}T_{n,\omega}(\theta) \leq \hat{c}_n\right\}.$$

Under upper bounds on the smoothness of the conditional mean that correspond to the lower bounds given in Section 4, upper bounds on the rate of convergence of set estimates based on KS statistics with bounded weights can be derived. These conditions are stated in the following assumption.

Assumption 5.1. *For some $\theta_0 \in \delta\Theta_0(P)$ such that θ_0 is in the interior of Θ, the following holds for some neighborhood $B(\theta_0)$ of θ_0. (i) $\bar{m}(\theta,x,P)$ is differentiable in θ with derivative $\bar{m}_\theta(\theta,x,P)$ bounded over $\theta\in B(\theta_0)$. (ii) For some $\eta>0$, we have, for all $\theta_0' \in (\delta\Theta_0(P)) \cap B(\theta_0)$, the set $\mathcal{X}_0(\theta_0')$ of points x_0 such that $\min_k \bar{m}_k(\theta_0',x_0,P)=0$ satisfies*

$$|\bar{m}_j(\theta_0',x,P) - \bar{m}_j(\theta_0',x_0,P)| \geq \eta\left(\|x-x_0\|^\alpha \wedge \eta\right),$$

for all j, and the number of elements in $\mathcal{X}_0(\theta_0')$ is bounded uniformly over θ_0'. (iii) X_i has

finite support and a bounded density on its support. (iv) There exists a path $t \mapsto \theta_t$ *such that* $\theta_t \to \theta_0$ *as* $t \to 0$ *and* $t \to d(\theta_t, \theta_0)$ *is continuous for* t *in a neighborhood of* 0.

Assumption 5.1 gives an upper bound on the smoothness of the conditional mean similar to the lower bound of Assumption 4.4. It states that α is the best (greatest) possible value of α for which Assumption 4.4 can hold. Without this assumption, rates of convergence derived using Assumption 4.4 and some value of α could be conservative, since the same assumption could also hold with a larger value of α. The next theorem uses this condition to get an upper bound on the rate of convergence of the set estimator $\mathcal{C}_{n,\omega}(\hat{c}_n)$ when the sequence of weight functions is uniformly bounded.

Theorem 5.1. *Under Assumptions 3.1, 3.2, 3.3 and 5.1, if* $\hat{c}_n$ *is bounded away from zero and* $g(X_i)$ *and* $m(W_i, \theta)$ *are uniformly bounded, then, for some* $\varepsilon > 0$,

$$P\left(n^{\alpha/(2d_X+2\alpha)} d_H\left(\mathcal{C}_{n,\omega}(\hat{c}_n), \Theta_0(P)\right) \geq \varepsilon\right) \overset{n\to\infty}{\to} 1.$$

Under the smoothness conditions of Section 4, this slower rate of convergence can be achieved (up to an arbitrarily slow rate of growth of the critical value) using bounded weights with an estimated set that contains $\Theta_0(P)$ with probability approaching one.

Theorem 5.2. *Suppose that Assumptions 3.1, 3.2, 3.3, 4.1, 4.3, 4.4, 4.5 and 4.6 hold. Let the weight function* $\omega_n(\theta, g)$ *satisfy* $\underline{\omega} \leq \omega_n(\theta, g) \leq \overline{\omega}$ *for some* $0 < \underline{\omega} \leq \overline{\omega} < \infty$, *and suppose that* $\hat{c}_n \to \infty$ *with* $\hat{c}_n/\sqrt{n} \to 0$. *Then*

$$\inf_{P\in\mathcal{P}} P(\Theta_0(P) \subseteq \mathcal{C}_{n,\omega}(\hat{c}_n)) \overset{n\to\infty}{\to} 1$$

and, for B *large enough,*

$$\sup_{P\in\mathcal{P}} P\left(\left(n/\hat{c}_n^2\right)^{\alpha/(2d_X+2\alpha)} d_H(\mathcal{C}_n(\hat{c}_n), \Theta_0(P)) > B\right) \overset{n\to\infty}{\to} 0.$$

The $n^{\alpha/(2d_X+2\alpha)}$ rate of convergence for the estimator using bounded weights is slower than the $(n/\log n)^{\alpha/(d_X+2\alpha)}$ rate of convergence derived in Section 4 for the estimator using the truncated variance weights. The rate of convergence is slower because sequences of local alternatives violate a shrinking set of moment inequalities. This leads to sequences of functions in $\mathcal{G}$ with the most power having a shrinking sequence of variances, so that a bounded weighting function cannot give them enough weight. While the examples in Section 6 show that this case is likely to be common in practice, bounded weight functions will have

advantages in other important cases (for example, when two inequalities form an equality conditional on X_i in a positive probability set, leading to point identification, bounded weights will typically perform better). Under conditions such as Assumption B.7 for the selection model in Section B.3, sequences of local alternatives lead to a single function in $\mathcal{G}$ with positive variance having power. In this case, using a bounded sequence of weight functions does not cause such a problem, and the increasing sequence of truncation points does worse by a power of $\log n$ because of the larger critical value needed for the KS statistic.

5.2 Kernel Methods

Suppose that we estimate the conditional mean $E_P(m_j(W_i, \theta)|X_i = x) = \bar{m}_j(\theta, x, P)$ using the kernel estimate

$$\hat{\bar{m}}_j(\theta, x) \equiv \frac{E_n m_j(W_i, \theta) k((X_i - x)/h_n)}{E_n k((X_i - x)/h_n)}$$

for some sequence $h_n \to 0$. Chernozhukov, Lee, and Rosen (2009) and Ponomareva (2010) propose methods for inference on conditional moment inequalities based on this estimate of the conditional mean. Following Chernozhukov, Lee, and Rosen (2009) this estimate of the conditional mean will converge at a $\sqrt{nh^{d_X}/\log n}$ rate uniformly over x. Using the results in this paper, this rate can be shown to be uniform over θ as well, so that the statistic

$$T^{\text{kern}}_{n,k,h_n}(\theta) \equiv \sup_{x \in \text{supp}_P(X_i)} S(\hat{\bar{m}}(\theta, x))$$

can be used to form an estimate

$$\mathcal{C}^{\text{kern}}_n(\hat{c}_n) \equiv \left\{ \theta \in \Theta \middle| \frac{\sqrt{nh^{d_X}}}{\sqrt{\log n}} T^{\text{kern}}_{n,k,h_n}(\theta) \leq \hat{c}_n \right\}$$

that will contain the identified set with probability approaching one for $\hat{c}_n$ large enough.

I place the following conditions on the choice of kernel function k. All of these conditions are fairly mild regularity conditions, except for the requirement that k be positive, which rules out higher order kernels. Ruling out higher order kernels is important. Since the class of KS statistics used in this paper integrate the conditional moment inequality against positive functions, these statistics cannot take advantage of smoothness conditions of more than two derivatives, while higher order kernels with a properly chosen bandwidth can.

Assumption 5.2. *(i) k is nonnegative (ii) k integrates to one, is bounded and square integrable over $\mathbb{R}^{d_X}$ and $k(t)$ is bounded away from zero for t in some neighborhood of 0 (iii) Assumption 3.2 holds with $\mathcal{G}$ replaced by the class of functions $t \mapsto k((t-x)/h)$ where x and h vary.*

As with set estimators based on KS statistics with bounded weights, the upper bounds on the smoothness of the conditional mean in Assumption 5.1 lead to upper bounds on the rate of convergence of estimates of the identified set based on kernel estimates. For the first order kernel estimates described above, estimates of the identified set will converge no faster than estimates based on variance weighted KS statistics, and will only achieve the same rate if the tuning parameter h_n is chosen to go to zero at the proper rate. Although this means that properly weighted KS statistics will generally do at least as well as first order kernel estimates and sometimes better in terms of rates of convergence, kernel estimates with a properly chosen sequence h_n may do better against alternatives that approach the identified set at a given rate.

The upper bound on rates of convergence for kernel based estimators is stated in the following theorem. In this theorem, the requirements that the critical value $\hat{c}_n$ be large and that the bandwidth h_n not shrink too quickly ensure that the procedure controls the probability of false rejection. If these conditions do not hold, we may have $\Theta_0(P) \not\subseteq \mathcal{C}_n^{\text{kern}}(\hat{c}_n)$ with high probability asymptotically.

Theorem 5.3. *Suppose that Assumptions 4.5, 5.1 and 5.2 hold. If $\hat{c}_n$ is chosen large enough, and if $h_n^{d_X} n / \log n \geq a$ for a large enough, then, for some $\varepsilon > 0$,*

$$P\left(\left(\frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}} \wedge h_n^{-\alpha}\right) d_H(\mathcal{C}_n^{kern}(\hat{c}_n), \Theta_0(P)) \geq \varepsilon\right) \overset{n\to\infty}{\to} 1.$$

The upper bound on the rate of convergence in Theorem 5.3 is the slower of $\frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}}$, which comes from a variance term, and $h_n^{-\alpha}$, which comes from a bias term. The optimal rate of convergence for estimates based on first order kernels will be achieved only when these terms are of the same order of magnitude, which corresponds to $h_n^{-\alpha} = \mathcal{O}\left(\frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}}\right)$ or $h_n = \mathcal{O}\left(\frac{\log n}{n}\right)^{1/(d_X+2\alpha)}$. Thus, choosing the optimal h_n requires knowing or estimating the Hölder constant α. While kernel based estimates may give more power when h_n is chosen optimally, variance weighted KS statistics give the same rate of convergence as kernel based estimates with the optimally chosen h_n without knowing α. If h_n is chosen to go to zero

at a different rate from the optimal rate for a given data generating process, kernel based estimates of the identified set will converge more slowly than estimates based on variance weighted KS statistics. If the choice of h_n is far enough off from the optimal choice (i.e. if the researcher is wrong enough about the smoothness of the data generating process), even the rate of convergence for unweighted KS statistics in Theorem 5.2 will be better than the rate of convergence of the kernel based estimate.

6 Applications

In this section, I verify the conditions for rates of convergence stated above for some applications under primitive conditions. Section B verifies the conditions for rates of convergence in some additional applications.

6.1 One Sided Regression

We posit a linear regression model $E_P(W_i^*|X_i) = X_i'\beta$ for a latent variable W_i^*, but we only observe (X_i, W_i^H), where W_i^H is known to be greater than or equal to W_i^*. This leads to the conditional moment inequality $E_P(W_i^H|X_i) \geq X_i'\beta$, which fits into the framework of this paper with $d_Y = 1$, $W_i = (X_i, W_i^H)$ and $m(W_i, \theta) = W_i^H - \theta_1 - X_i'\theta_{-1}$. Here, $\bar{m}(\theta, x) = E_P(W_i^H|X_i = x) - \theta_1 - x'\theta_{-1}$. I verify the conditions used above to derive rates of convergence (Assumptions 4.3 and 4.4) under the following assumptions.

Assumption 6.1. *For some $C > 0$ and $0 < \alpha \leq 1$, $\|E_P(W_i^H|X_i = x) - E_P(W_i^H|X_i = x')\| \leq C\|x - x'\|^\alpha$ for x and x' on the support of X_i for all $P \in \mathcal{P}$.*

Assumption 6.1 places a Hölder condition on the conditional mean of the upper bound of the outcome given X_i. This is a smoothness condition on the data generating process. For $\alpha = 1$, Assumption 6.1 states that this conditional mean must be Lipschitz continuous. For smaller α, the conditional mean must still be continuous, but can be less smooth.

For $\alpha > 1$, a condition like Assumption 6.1 would restrict $E_P(W_i^H|X_i = x)$ to be constant, since its slope would have to converge to zero at every point. However, as described in Section 2.1, this condition factors into the rate of convergence only in restricting $E_P(W_i^H - \theta_1 - X_i'\theta_{-1}|X_i = x)$ to increase no faster than a multiple of $\|x - x_0\|^\alpha$ near some tangency point x_0 for $\theta = (\theta_1, \theta_{-1})$ on the boundary of the identified set. The same argument will still go through as long as this restriction on the difference between $E_P(W_i^H|X_i = x)$ and a tangent line holds for some α, even if $\alpha > 1$. While placing this condition directly on

$E_P(W_i^H - \theta_1 - X_i'\theta_{-1}|X_i = x)$ near tangency points is a bit awkward in general, this condition has a natural interpretation when $\alpha = 2$. In this case, it requires that the difference between the conditional mean $E_P(W_i^H|X_i = x)$ and any tangent line behave quadratically near the tangent point, which is implied by a bound on the second derivative. This is the content of the next assumption.

Assumption 6.2. *(i) $E_P(W_i^H|X_i = x)$ has a second derivative that is bounded uniformly in P and x and (ii) for any $P \in \mathcal{P}$, $\theta_0 \in \Theta_0(P)$, $E_P(W_i^H|X_i = x)$ is bounded away from $\theta_{0,1} + x'\theta_{0,-1}$ on the boundary of the support of X_i*

Part (ii) of Assumption 6.2 restricts the local alternatives that determine Hausdorff distance of the set estimator to be in the interior of the support of X_i. Without such a condition, the estimators considered here suffer from the familiar problems of kernel estimators at the boundary of the support of a conditioning variable, and the slower rate with $\alpha = 1$ is obtained.

The next assumption ensures that the condition on the tangent angle in Assumption 4.4 holds. Under this assumption, rates of convergence to the identified set depend on sequences of parameters in which only the intercept parameter varies. This condition ensures that varying the intercept parameter a small amount near the boundary of the identified set gives an element that is still in the parameter space Θ.

Assumption 6.3. *The subvector θ_{-1} of θ is bounded over $\theta \in \Theta$ and, for any $\theta \in \Theta$, $(\theta_1', \theta_{-1}) \in \Theta$ for all θ_1' such that $\inf_{P\in\mathcal{P}} \inf_x E_P(W_i^H|X_i = x) - x'\theta_{-1} \leq \theta_1' \leq \theta_1$.*

Theorem 6.1. *Suppose that Assumptions 6.3 holds in the one sided linear regression model and X_i has compact support for all $P \in \mathcal{P}$. Then, if Assumption 6.1 holds, Assumptions 4.3 and 4.4 will hold for α specified in Assumption 6.1. If Assumption 6.2 holds, Assumptions 4.3 and 4.4 will hold for $\alpha = 2$.*

If the parameter space Θ is restricted so that all sequences of local alternatives corresponded to rotating the regression line around a tangent point, Assumption 6.3 will fail and the rate of convergence will be slower. The verification of the assumptions of Theorem 4.3 will not go through in this case because the first part of Assumption 4.4 will fail. As an example, suppose $E_P(W_i^H|X_i = x) = x^2$. If the parameter space Θ does not restrict the intercept parameter, the proof of Theorem 6.1 will go through. However, if $\Theta = \{(0, \theta_1)|\theta_1 \in \mathbb{R}\}$ (that is, we restrict the intercept to be 0), the rate of convergence will be determined by local alternatives of the form $(0, a_n)$. I show in Section A.4 of the appendix that the estimate of the

identified set converges no faster than at a $((\log n)/n)^{1/5}$ rate, rather than the $((\log n)/n)^{2/5}$ rate for the case where the parameter space is unrestricted, and that the estimate based on bounded weights has an even slower rate of convergence.

These issues also make it more difficult to state primitive conditions that lead to Assumption 4.4 in the case of two sided interval regression, in which we add the conditional moment inequality $m_2(W_i, \theta) = \theta_1 + X_i'\theta_{-1} - W_i^L$. As with restricting the parameter space, adding the second conditional moment inequality can lead to the rate of convergence being determined by sequences of local alternatives that correspond to rotating the regression line around a tangent point. One example that leads to this is when $E_P(W_i^H|X_i = x) = x^2$ and $E_P(W_i^L|X_i = x) = -x^2$. Adding the moment inequality on W_i^L has the same effect as restricting the intercept to be zero in the example above. The rate of convergence to the identified set is determined by local alternatives of the form $(0, a_n)$, which leads to a slower rate of convergence. The argument in Section A.4 applies here as well, leading to a slower $((\log n)/n)^{1/5}$ rate of convergence.

For the case where X_i is a scalar, these cases can be ruled out in the interval regression model by placing conditions on certain tangency points. I go through this argument in the next section. However, higher dimensions appear to require further conditions.

6.2 Interval Regression with a Scalar Regressor

In what follows, I consider an interval regression model in which, in addition to W_i^H defined as in Section 6.1, we observe W_i^L that is known to satisfy $W_i^L \leq W_i^*$, so that $E_P(W_i^L|X_i) \leq \theta_1 + X_i'\theta_{-1}$. This fits into the framework of this paper with $m(W_i, \theta) = (W_i^H - \theta_1 - X_i'\theta_{-1}, \theta_1 + X_i'\theta_{-1} - W_i^L)$. I restrict attention to the case where $d_X = 1$, so that $\theta_{-1} = \theta_2$ is a scalar.

In addition to the assumptions used in Section 6.1, I impose the following assumption, which rules out cases like the one described above in which local alternatives correspond to rotating the regression line around a tangent point. For the following condition, let $(\theta_1^u(P), \overline{\theta}_2(P)) \in \Theta_0(P)$ and $(\theta_1^\ell(P), \underline{\theta}_2(P)) \in \Theta_0(P)$ be such that $\overline{\theta}_2(P) = \sup_{\theta \in \Theta_0(P)} \theta_2$ and $\underline{\theta}_2(P) = \inf_{\theta \in \Theta_0(P)} \theta_2$. Define $x_{0,1}^u(P) = \max\{x | E_P(W_i^H|X_i = x) = \theta_1^u(P) + \overline{\theta}_2(P)x\}$, $x_{0,2}^u(P) = \min\{x | E_P(W_i^L|X_i = x) = \theta_1^u(P) + \overline{\theta}_2(P)x\}$, $x_{0,1}^\ell(P) = \max\{x | E_P(W_i^L|X_i = x) = \theta_1^\ell(P) + \underline{\theta}_2(P)x\}$ and $x_{0,2}^\ell(P) = \min\{x | E_P(W_i^H|X_i = x) = \theta_1^\ell(P) + \underline{\theta}_2(P)x\}$.

Assumption 6.4. *(i) The support of X_i is bounded uniformly in $P \in \mathcal{P}$. (ii) The absolute value of the slope parameter θ_2 is bounded uniformly on the identified sets $\Theta_0(P)$ of $P \in \mathcal{P}$. (iii) $x_{0,1}^u(P) - x_{0,2}^u(P)$ and $x_{0,1}^\ell(P) - x_{0,2}^\ell(P)$ are bounded from below away from zero uniformly over $P \in \mathcal{P}$.*

Theorem 6.2. *In the interval regression model with $d_X = 1$, suppose that Assumption 6.4 holds. Then, if Assumption 6.1 holds as stated and with W_i^H replaced by W_i^L, Assumptions 4.3 and 4.4 will hold for α specified in Assumption 6.1 (and $d_X = 1$). If Assumption 6.2 holds as stated and with W_i^H replaced with W_i^L, Assumptions 4.3 and 4.4 will hold for $\alpha = 2$ (and $d_X = 1$).*

Part (iii) of Assumption 6.4 ensures that the tangent points for x for the largest and smallest values of the slope parameter in the identified set are bounded away from each other. If this does not hold, the rate of convergence can be slower, as shown in Section A.4. A sufficient condition for part (iii) of Assumption 6.4 is that the slope parameter is bounded uniformly over the parameter space and that $E_P(W_i^H|X_i = x) - E_P(W_i^L|X_i = x)$ is bounded away from zero uniformly over x and $P \in \mathcal{P}$. However, Assumption 6.4 also allows for point identified cases or "nearly" point identified cases in which $E_P(W_i^H|X_i = x) - E_P(W_i^L|X_i = x)$ is small (or zero), so long as the two conditional means are not only close at a single point. In particular, Assumption 6.4 allows $E_P(W_i^H|X_i = x)$ and $E_P(W_i^L|X_i = x)$ to be equal on a positive probability set (leading to point identification). In this case, the results in Theorem 6.2 still apply, but are conservative, since local alternatives will violate the conditional moment inequalities on a set with nonvanishing probability. Since Theorem 6.2 gives the minimax rate of convergence over a class of underlying distributions that includes both regular point identified and set identified cases, the resulting rate corresponds to the slower of the two cases, which turns out to be the set identified case.

7 Monte Carlo

To examine the finite sample properties of the set estimates proposed in this paper, and to illustrate their implementation, I perform a monte carlo study. I compare the weighted KS statistic based set estimators to the other estimators in Section 5 in a quantile regression model with missing data on the outcome variable, where no additional assumptions are imposed on the process generating the missing values. Letting W_i^* be the true value of the outcome variable, I simulate from a model where the median of W_i^* given $X_i = x$ is given by $\theta_1 + \theta_2 x$, but W_i^* is not always observed. This falls into the framework of the interval quantile regression model described in Section B.2, with $W_i^H = W_i^L = W_i^*$ when the outcome variable is observed, and $W_i^H = \infty$ and $W_i^L = -\infty$ when the outcome variable is unobserved. The identified set contains all values of (θ_1, θ_2) that are consistent with the median regression model and some, possibly endogenous, censoring mechanism generating the missing values.

I generate data as follows. For X_i and U_i^* generated as independent variables with $X_i \sim \text{unif}(-3,3)$ and $U_i^* \sim \text{unif}(-1,1)$ and $(\theta_{1,*}, \theta_{2,*}) = (1/4, 1/2)$, I set $W_i^* = \theta_{1,*} + \theta_{2,*}X_i + U_i^*$. Then, I set W_i^* to be missing (that is, $(W_i^L, W_i^H) = (-\infty, \infty)$) with probability $1/5 - X_i^2/20 + X_i^4/200$, and observed ($W_i^L = W_i^H = W_i^*$) with the remaining probability $1 - (1/5 - X_i^2/20 + X_i^4/200)$. Note that, while the data are generated by taking a particular point $(\theta_{1,*}, \theta_{2,*})$ in the identified set and using a censoring process that satisfies the missing at random assumption (that the event of W_i^* not being observed is independent of U_i^* conditional on X_i^*), the identified set for this model is larger than a single point, and contains all values of (θ_1, θ_2) that are consistent with median regression and any form of censoring, including those where the probability of not observing W_i^* depends on the outcome W_i^* itself. I generate monte carlo data sets with the data generating process described above and n equal to 200, 500, and 1000 observations. I use 1000 replications for each monte carlo design.

For each monte carlo replication, I compute set estimates based on the weighted KS statistics in this paper using various choices of the user defined parameters σ_n and $\hat{c}_n$. I also compute set estimates based on a constant weighting, and kernel estimates, as described in Section 5. For the weighted and unweighted KS statistics, the class of functions $\mathcal{G}$ is taken to be the set of indicator functions for intervals $\{x \mapsto I(s < x < s+t) | t \geq 0\}$. For the kernel based estimators, the uniform kernel $x \mapsto I(-h/2 < x < h/2)$ is used, and the supremum is taken over x such that the kernel function is positive only on the support of X_i. For each of these estimators, I form $\hat{c}_n$ has follows. Using monte carlo simulation with 1000 replications, I compute the .95 quantile of the distribution of $\sqrt{n/\log n}$ times the test statistic where $(m_1(W_i, \theta), m_2(W_i, \theta)) = (D_i^* - 1/2, 1/2 - D_i^*)$, where $\{D_i^*\}_{i=1}^n$ are iid bernoulli(1/2) variables independent of $\{X_i\}_{i=1}^n$ and X_i has the same distribution as in data generating process for the monte carlos. This corresponds to a distribution where both moment inequalities are binding for all x since, with probability one, no observations are missing. Letting $c_{.95,n}$ be this value, I take the critical value to be $c_{.95,n}$ times a slowly increasing sequence, which I take to be $\sqrt{\log \log n}$ in most cases, except for the variance weighted KS statistic, for which I also compute sets with the sequence taken to be $\sqrt{\log n}$ and 1 to assess the sensitivity of the estimator to $\hat{c}_n$ (by results in Armstrong and Chan, 2012, $c_{.95,n}$ converges to a positive constant, so that, except for in the case where the increasing sequence is taken to be 1 for all n, these critical values satisfy the conditions of Theorem 3.1; the case where the increasing sequence taken to be 1 for all n is included to examine the performance of the estimator when the sequence of critical values is smaller than what is allowed by the conditions of this paper). For σ_n for the variance weighted KS statistics, I report results for $\sigma_n = (1/2)n^{-1/6}$,

$\sigma_n = (1/2)n^{-1/10}$ and $\sigma_n = (1/2)\sqrt{(\log n)(\log\log n)/n}$, which correspond to the slowest possible rate for $\alpha = 1$ and 2, and (a slowly increasing sequence multiplied by) the fastest possible rate respectively (the 1/2 factor corresponds to the conditional standard deviation of the moment function in the median regression model). For the kernel based estimators, I report results with the bandwidth given by $(\overline{x} - \underline{x})n^{-1/3}$ and $(\overline{x} - \underline{x})n^{-1/5}$, where $\overline{x} = 3$ and $\underline{x} = -3$ are the upper and lower endpoints of the support of X_i. These correspond to the optimal rate for the bandwidth for $\alpha = 1$ and $\alpha = 2$ respectively.

Tables 1 and 2 report median Hausdorff distances for each estimator for the projection of the set estimator onto each parameter, and for the parameter vector itself (comparisons for other quantiles of the Hausdorff distance are similar). The monte carlo coverage probability is at least 99.9% in all cases, except for the case where the increasing sequence that multiplies $c_{n,.95}$ is taken to be 1 (first column of Table 2), for which the monte carlo coverage probabilities are 99.5%, 98.2% and 99.3% for sample sizes 200, 500 and 1000 respectively.

For these data generating processes, the variance weighted statistics generally perform better than the unweighted statistics for the slope parameter β_1 and worse for the intercept parameter β_2. For the parameterization used here, this translates to the Hausdorff distance for both parameters being larger for the weighted statistics, although it should be noted that this depends on the parameterization and the units in which the covariate is measured (e.g., if the outcome is yearly income in dollars and the covariate is years of education, β_1 is measured in dollars, while β_2 is measured in dollars per year of education; if one instead measures education in months, β_2 will be divided by 12, while β_1 will remain the same; in these monte carlos, reparameterizing so that β_2 is multiplied by, say, 10, leads to the Hausdorff distance being smaller for the weighted statistic).

The data generating process used here satisfies the conditions of this paper with $\alpha = 2$ and $d_X = 1$. The values of σ_n are all chosen so that the variance weighted KS statistic based estimator achieves the same rate as the kernel based estimator with the optimal bandwidth (which is proportional to $n^{-1/5}$). Thus, according to the asymptotic results one should expect that the kernel estimator with the optimal bandwidth should perform slightly better than the variance weighted KS statistic based estimators, but that the variance weighted KS statistic based estimators should not be too far behind, and should perform better than the kernel estimator with the bandwidth chosen proportional to $n^{-1/3}$. This holds for the slope parameter β_2, although the weighted KS statistic actually performs better for the intercept parameter and for the Hausdorff distance of both parameters together.

Thus, the asymptotic power results of Theorem B.2 appear to provide a good description

of the behavior of the estimates of the slope parameter for this case, while the asymptotic power results for the intercept parameter appear to be better described by an asymptotic approximation where the conditional moment inequality is violated on a set with nonshrinking probability (as with Assumption B.7 or the local alternatives considered by Andrews and Shi, 2009). A likely explanation for this is that, for this data generating process, the conditional moment inequality is violated for all values of x on the support of X_i when the intercept parameter is fixed at values in the identified set and the intercept parameter is moved in the range of the median distances seen in these monte carlos. In any case, as both of these asymptotic approximations predict, the weighted KS statistic based estimator does close to as well as the best estimator in all cases, while each of the other estimators performs worse in some case.

In the monte carlo results described above, the data generating process is designed so that, according to asymptotic theory, the weighted statistic should perform better than the unweighted statistic. To examine a setting where asymptotic theory predicts that the unweighted statistic performs better, I perform a monte carlo analysis of a version of the missing data model with a constant probability of missing outcomes given X_i, and with the slope variable constrained so that all local alternatives violate the conditional moment inequality on a nonshrinking set. The data generating process for X_i and U_i^* is the same as above, but I set $(\theta_{1,*}, \theta_{2,*}) = (0, 0)$, and set the W_i^* to be missing with probability .1 independently of (X_i, U_i^*). For this model, I estimate the upper endpoint of the identified set for θ_1 with θ_2 set to zero. This can be considered a quantile version of the selection model with an exclusion restriction (the exclusion restriction being that $\theta_2 = 0$) in Section B.3, with the data generating process satisfying the quantile version of Assumption B.7. Thus, according to asymptotic theory, the weighted statistic should perform slightly worse than the unweighted statistic. To assess how well this describes these data generating processes, I compare the Hausdorff distance of estimators based on weighted and unweighted statistics.

The results of the monte carlos for this data generating process are reported in Table 3. This table reports the median distance between the upper endpoint of the estimator of the identified set and the upper endpoint of the identified set. The critical value $\hat{c}_n$ is the same as for the other data generating processes. As predicted, the unweighted statistic performs better in this setting.

8 Conclusion

This paper proposes estimates of the identified set in conditional moment inequality models based on variance weighted KS statistics. I derive rates of convergence of these and other set estimators to the identified set under conditions that apply to many models of practical interest. In many settings, the rate of convergence of the set estimator I propose is the fastest among those available, and, in settings where other estimators are better, the improvement in rate of convergence is no more than a factor of $\log n$. While, in most cases, there is some other estimator that does slightly better, choosing the correct one requires knowledge of smoothness and shape conditions on the data generating process, and guessing incorrectly about these conditions can lead the researcher to use an estimator with a much slower rate of convergence. The advantage of the estimator proposed in this paper is that it performs well under a variety of conditions without prior knowledge of which of these conditions hold.

In settings where local alternatives violate the conditional moment inequalities on a shrinking set, the weights I propose for KS statistics give the statistics more power against local alternatives than bounded weights. The examples in Section 6 show that this situation is common in practice. When sequences of local alternatives violate the conditional moment inequalities on a fixed, positive probability set, the larger critical values required by the increasing sequence of weight functions lead to a loss in power, but only by a factor of $(\log n)^{1/2}$. This provides a theoretical justification for variance weighting in this context. Under certain conditions, weighting the KS statistic objective function by a truncated inverse of the estimated variance increases the rate of convergence of the corresponding estimator of the identified set.

A Proofs and Auxiliary Results

This appendix collects several results not stated in the body of the paper. In Section A.1, I state and prove uniform convergence results for classes of functions weighted by truncated standard deviations. These results are used later in the appendix in proving some of the results stated in the body of the paper. Section A.2 provides (conservative) bounds for the critical values used in the paper that do not require arbitrary increasing sequences. In Section A.3, I provide sufficient conditions for the rate of convergence to be strictly faster than $\sqrt{n}$. In Section A.4, I provide an example of a data generating process for an interval regression where low power against local alternatives when the slope parameter varies leads to a slower rate of convergence to the identified set. In Section A.5, I state conditions under which

Assumption 3.2 holds and verify them for the applications described in Section 6. Section A.6 contains proofs of the theorems stated in the body of the paper and in Appendix B.

A.1 Uniform Convergence Lemma

The following lemma is useful in deriving some of these results. Applied to mean zero functions, the lemma says that any sequence of classes of functions that is not too complex converges uniformly at a $\sqrt{n/\log n}$ rate when scaled by the standard deviation if the minimum standard deviation does not go to zero too fast.

Lemma A.1. *Let $Z_1, \ldots, Z_n$ be iid observations and let $\mathcal{P}$ be a set of probability distributions and $\mathcal{F}_{n,P}$ a set of classes of functions indexed by $n \in \mathbb{N}$ and $P \in \mathcal{P}$ such that, for some $\overline{f}$, $f(Z_i) \leq \overline{f}$ with P-probability one for $P \in \mathcal{P}$ and $f \in \mathcal{F}_{n,P}$ for each n. Let $\mu_{2,P}(f) = (E_P f(Z_i)^2)^{1/2}$ and let $\mu_{2,n}$ be a sequence such that $\mu_{2,n}\sqrt{n/\log n}$ is bounded away from zero. Let $\mathcal{G}_{n,P} = \{f\mu_{2,n}/(\mu_{2,P}(f) \vee \mu_{2,n}) | f \in \mathcal{F}_{n,P}\}$ and suppose that*

$$\sup_{P\in\mathcal{P}} \sup_{n\in\mathbb{N}} \sup_{Q} N(\varepsilon, \mathcal{G}_{n,P}, L_1(Q)) \leq A\varepsilon^{-W}$$

for $0 < \varepsilon < 1$ where the supremum over Q is over all probability measures. Then for some B that does not depend on N,

$$\sup_{P\in\mathcal{P}} P\left(\frac{\sqrt{n}}{\sqrt{\log n}} \sup_{f\in\mathcal{F}_{n,P}} \left|(E_n - E_P)\frac{f(Z_i)}{\mu_{2,P}(f) \vee \mu_{2,n}}\right| \geq B \text{ some } n \geq N\right) \overset{N\to\infty}{\to} 0.$$

Proof. The result follows by applying the following theorem to the classes of functions $\mathcal{G}_{n,P}$. For $g = f\mu_{2,n}/(\mu_{2,P}(f) \vee \mu_{2,n}) \in \mathcal{G}_{n,P}$, $E_P g(Z_i)^2 = E_P f(Z_i)^2 \mu_{2,n}^2/(\mu_{2,P}(f)^2 \vee \mu_{2,n}^2) = \mu_{2,P}(f)^2\mu_{2,n}^2/(\mu_{2,P}(f)^2 \vee \mu_{2,n}^2) \leq \mu_{2,n}^2$, so the theorem applies with the same $\mu_{2,n}$. □

Specialized to a class $\mathcal{P}$ of probability distributions with a single element P, this says that the sequence in the probability statement in the last display of the lemma is bounded by B with P-probability one. The conclusion of the lemma implies that this scaled sequence is $\mathcal{O}_P(1)$ uniformly in $P \in \mathcal{P}$, but is slightly stronger.

The proof of the lemma uses the following theorem, which is a slightly stronger version of Theorem 37 in Pollard (1984), with the conditions stated in a slightly different way. The following theorem basically follows the arguments of the proof of Theorem 37 in Pollard (1984), but changes a few things to get a slightly stronger result. Note that the notation $\mu_{2,P}^2$ is used for the raw second moment of functions rather than their variance, although the

distinction is often not important since applications typically involve the raw second moment going to zero at the same rate as the variance.

Theorem A.1. *Let $Z_1, \ldots, Z_n$ be iid observations and let $\mathcal{P}$ be a set of probability measures and $\mathcal{F}_{n,P}$ a set of classes of functions indexed by $n \in \mathbb{N}$ and $P \in \mathcal{P}$ such that, for some $\overline{f}$, $f(Z_i) \leq \overline{f}$ P-a.s. for $f \in \mathcal{F}_{n,P}$ for $P \in \mathcal{P}$ for each n and, for some positive constants A and W,*

$$\sup_{P \in \mathcal{P}} \sup_{n \in \mathbb{N}} \sup_{Q} N(\varepsilon, \mathcal{F}_{n,P}, L_1(Q)) \leq A\varepsilon^{-W}$$

for $0 < \varepsilon < 1$ where the supremum over Q is over all probability measures. Suppose that, for some sequence $\mu_{2,n}$, $E_P f(Z_i)^2 \leq \mu_{2,n}^2$ for all $f \in \mathcal{F}_{n,P}$ for all $P \in \mathcal{P}$ for all n. Then, if $\mu_{2,n}\sqrt{n/\log n}$ is bounded away from zero we will have, for some B that does not depend on N,

$$\sup_{P \in \mathcal{P}} P\left(\frac{\sqrt{n}}{\mu_{2,n}\sqrt{\log n}} \sup_{f \in \mathcal{F}_{n,P}} |(E_n - E_P) f(Z_i)| \geq B \text{ some } n \geq N \right) \overset{N \to \infty}{\to} 0.$$

Proof. The proof is a slight modification of the proof of Theorem 37 in Pollard (1984). The sequence $\mu_{2,n}$ corresponds to δ_n in that theorem, and, in contrast to the theorem from Pollard (1984) which defines a sequence α_n that must satisfy certain conditions, this theorem corresponds to using the best α_n sequence possible, and noting that α_n need not be nonincreasing as long as it is bounded.

Without loss of of generality, assume that $\overline{f} = 1$. Fix B (conditions on how large B has to be will be stated throughout the theorem) and set $\varepsilon_n = \frac{B\mu_{2,n}\sqrt{\log n}}{8\sqrt{n}}$. Since $var_P((E_n - E_P)f(Z_i))/(4\varepsilon_n^2) \leq (\mu_{2,n}^2/n)/(4B^2\mu_{2,n}^2(\log n)/(64n)) = 16/(B^2 \log n) \leq 1/2$ for n greater than some number that does not depend on P, the inequality (30) in Pollard (1984) will eventually imply

$$P\left(\frac{\sqrt{n}}{\mu_{2,n}\sqrt{\log n}} \sup_{f \in \mathcal{F}_{n,P}} |(E_n - E) f(Z_i)| \geq B \right) = P\left(\sup_{f \in \mathcal{F}_{n,P}} |(E_n - E) f(Z_i)| \geq 8\varepsilon_n \right)$$
$$\leq 4(P \times \nu)\left(\sup_{f \in \mathcal{F}_{n,P}} |\mathbb{P}_n^\circ f(Z_i)| \geq 2\varepsilon_n \right)$$

for all $P \in \mathcal{P}$ where $\mathbb{P}_n^\circ f(Z_i) = \frac{1}{n}\sum_{i=1}^n f(Z_i) \cdot s_i$ and $s_1, \ldots, s_n$ are iid random variables that take on values ± 1 each with probability one half drawn independent of $Z_1, \ldots, Z_n$ and ν

denotes the probability measure of $s_1, \ldots, s_n$. Conditional on the data, this is bounded by

$$(P \times \nu)\left(\sup_{f \in \mathcal{F}_n} |\mathbb{P}_n^\circ f(Z_i)| \geq 2\varepsilon_n \middle| Z_1, \ldots, Z_n\right) \leq 2N(\varepsilon_n, \mathcal{F}_{n,P}, L_1(\mathbb{P}_n)) \exp\left[-\frac{1}{2}\frac{n\varepsilon_n^2}{\left(\sup_{f \in \mathcal{F}_{n,P}} E_n f(Z_i)^2\right)}\right].$$

For any constant $a > 0$, on the event that

$$\sup_{f \in \mathcal{F}_{n,P}} E_n f(Z_i)^2 \leq a^2 \mu_{2,n}^2, \tag{5}$$

the previous display will be bounded by

$$\begin{aligned}
&2N(\varepsilon_n, \mathcal{F}_{n,P}, L_1(\mathbb{P}_n)) \exp\left(-\frac{1}{2}\frac{n\varepsilon_n^2}{a^2\mu_{2,n}^2}\right) \leq 2A\varepsilon_n^{-W} \exp\left(-\frac{1}{2}\frac{n\varepsilon_n^2}{a^2\mu_{2,n}^2}\right) \\
&= 2A \exp\left[-\frac{1}{2} \cdot n \cdot \frac{B^2\mu_{2,n}^2 \log n}{64n} \cdot \frac{1}{a^2\mu_{2,n}^2} - W \log \frac{B\mu_{2,n}\sqrt{\log n}}{8\sqrt{n}}\right] \\
&= 2A \exp\left[-\frac{B^2 \log n}{128a^2} - W \log \frac{B}{8} - W \log \frac{\mu_{2,n}\sqrt{\log n}}{\sqrt{n}}\right]
\end{aligned} \tag{6}$$

The condition that $\mu_{2,n}\sqrt{n/\log n}$ is bounded away from zero is more than enough to guarantee that the term in the last logarithm is bounded from below by a fixed power of n. Thus, the expression in the last display can be made to go to zero at any polynomial rate for any a by choosing B to be large enough (in a way that depends on a but not n or P).

For any $P \in \mathcal{P}$, the P-probability of (5) failing to hold can be bounded using Lemma 33 in Pollard (1984) with $\delta_n = a\mu_{2,n}/8$ (the lemma holds for $a \geq 8$):

$$\begin{aligned}
&P\left(\sup_{f \in \mathcal{F}_{n,P}} E_n f(Z_i)^2 > a^2\mu_{2,n}^2\right) = P\left(\sup_{f \in \mathcal{F}_{n,P}} E_n f(Z_i)^2 > 64\delta_n^2\right) \leq 4E_P[N(\delta_n, \mathcal{F}_{n,P}, L_2(\mathbb{P}_n))] \exp(-n\delta_n^2) \\
&\leq 4A(\delta_n/2)^{-W} \exp(-n\delta_n^2) = 4 \cdot 2^W A \exp(-n\delta_n^2 - W \log \delta_n) \\
&= 4 \cdot 2^W A \exp\left[-na^2\mu_{2,n}^2/64 - W \log \frac{a}{8} - W \log \mu_{2,n}\right] \\
&\leq 4 \cdot 2^W A \exp\left[-\frac{na^2}{64}\frac{c \log n}{n} - W \log \frac{a}{8} - \frac{W}{2} \log \frac{c \log n}{n}\right]
\end{aligned} \tag{7}$$

where $\sqrt{c}$ is a lower bound for $\mu_{2,n}\sqrt{n/\log n}$. This can be made to go to zero at any polynomial rate by choosing a large.

Thus, if we choose a and B large enough, $\sup_{P \in \mathcal{P}} P\left(\frac{\sqrt{n}}{\mu_{2,n}\sqrt{\log n}} \sup_{f \in \mathcal{F}_{n,P}} |(E_n - E_P)f(Z_i)| \geq B\right)$

will be summable over n, so that

$$\sup_{P\in\mathcal{P}} P\left(\frac{\sqrt{n}}{\mu_{2,n}\sqrt{\log n}}\sup_{f\in\mathcal{F}_{n,P}}|(E_n-E_P)f(Z_i)|\geq B \text{ some } n\geq N\right)$$
$$\leq\sum_{n\geq N}\sup_{P\in\mathcal{P}} P\left(\frac{\sqrt{n}}{\mu_{2,n}\sqrt{\log n}}\sup_{f\in\mathcal{F}_{n,P}}|(E_n-E_P)f(Z_i)|\geq B\right)\overset{N\to\infty}{\to}0.$$

□

With this lemma in hand, we can get rates of convergence for classes of functions weighted by their standard deviation under additional conditions that allow the standard deviation to be consistently estimated. In order to get results for functions weighted by the standard deviation rather than the raw second moment, I apply the previous results to classes of functions of the form $f-E_Pf(Z_i)$. Letting $\hat{\sigma}(f)^2=E_n(f(Z_i))^2-(E_nf(Z_i))^2$ and $\sigma_P(f)^2=E_P(f(Z_i))^2-(E_Pf(Z_i))^2$, rates of convergence for

$$\sup_{f\in\mathcal{F}_n}\left|(E_n-E_P)\frac{f(Z_i)}{\hat{\sigma}(f)\vee\sigma_n}\right|$$

will follow by applying the above results to the classes of functions $f-E_Pf(Z_i)$ once we can bound $\frac{\sigma_P(f)\vee\sigma_n}{\hat{\sigma}(f)\vee\sigma_n}$, and for this it is sufficient to show that $\hat{\sigma}(f)/\sigma_P(f)$ converges to one uniformly over $\sigma_P(f)\geq\sigma_n$. The following lemma gives sufficient conditions for this.

Lemma A.2. *Let $Z_1,\ldots,Z_n$ be iid observations and let $\mathcal{F}_n$ be a sequence of classes of functions and $\mathcal{P}$ a set of probability distributions such that, for some $\overline{f}$, $f(Z_i)\leq\overline{f}$ with P-probability one for $P\in\mathcal{P}$ and $f\in\mathcal{F}_n$ for each n. Let $\sigma_P(f)=(E_Pf(Z_i)^2-(E_Pf(Z_i))^2)^{1/2}$ and let σ_n be a sequence such that $\sigma_n\sqrt{n/\log n}$ is bounded away from zero. Define $\mathcal{G}^1_{n,P}=\{(f-E_Pf(Z_i))\sigma_n/(\sigma_P(f)\vee\sigma_n)\}$ and $\mathcal{G}^2_{n,P}=\{(f-E_Pf(Z_i))^2\sigma_n/(\mu_{2,P}([f-E_Pf(Z_i)]^2)\vee\sigma_n)\}$, and suppose that, for some positive constants A and W,*

$$\sup_{P\in\mathcal{P}}\sup_{n\in\mathbb{N}}\sup_{Q}N(\varepsilon,\mathcal{G}^i_{n,P},L_1(Q))\leq A\varepsilon^{-W}$$

for $0<\varepsilon<1$ and $i=1,2$, where the supremum over Q is over all probability measures.

Then, for every $\varepsilon>0$, there exists a c such that, if $\sigma_n\sqrt{n/\log n}\geq c$ for all n,

$$\sup_{P\in\mathcal{P}} P\left(\sup_{f\in\mathcal{F}_n,\sigma_P(f)\geq\sigma_n}\left|\frac{\hat{\sigma}(f)}{\sigma_P(f)}-1\right|\geq\varepsilon \text{ some } n\geq N\right)\overset{N\to\infty}{\to}0.$$

Proof. We have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n,\sigma_P(f)\geq\sigma_n}\left|\frac{\hat{\sigma}^2(f)-\sigma_P^2(f)}{\sigma_P^2(f)}\right| \\
&= \sup_{f\in\mathcal{F}_n,\sigma_P(f)\geq\sigma_n}\left|\frac{(E_n-E_P)(f(Z_i)-E_Pf(Z_i))^2-(E_nf(Z_i)-E_Pf(Z_i))^2}{\sigma_P^2(f)}\right| \\
&\leq \sup_{f\in\mathcal{F}_n,\sigma_P(f)\geq\sigma_n}\left|\frac{(E_n-E_P)(f(Z_i)-E_Pf(Z_i))^2}{\sigma_P^2(f)}\right|+\left|\frac{[(E_n-E_P)f(Z_i)]^2}{\sigma_P^2(f)}\right|. \qquad (8)
\end{aligned}$$

The first term is equal to

$$\left|\frac{(E_n-E_P)(f(Z_i)-E_Pf(Z_i))^2}{\mu_{2,P}([f-E_Pf(Z_i)]^2)\vee\sigma_n}\right|\frac{\mu_{2,P}([f-E_Pf(Z_i)]^2)\vee\sigma_n}{\sigma_P^2(f)}.$$

We have $\mu_{2,P}([f-E_Pf(Z_i)]^2)^2 = E_P[f(Z_i)-E_Pf(Z_i)]^4 \leq 4\overline{f}^2E_P[f(Z_i)-E_Pf(Z_i)]^2 = 4\overline{f}^2\sigma_P(f)^2$ so that

$$\frac{\mu_{2,P}([f-E_Pf(Z_i)]^2)\vee\sigma_n}{\sigma_P^2(f)}\leq\frac{[2\overline{f}\sigma_P(f)]\vee\sigma_n}{\sigma_P^2(f)}\leq\frac{2\overline{f}\vee1}{\sigma_P(f)}\leq\frac{2\overline{f}\vee1}{\sigma_n}$$

where the last two inequalities hold for $\sigma_P(f)\geq\sigma_n$. Thus, for any $\varepsilon>0$,

$$\begin{aligned}
&\sup_{P\in\mathcal{P}}P\left(\sup_{f\in\mathcal{F}_n,\sigma_P(f)\geq\sigma_n}\left|\frac{(E_n-E_P)(f(Z_i)-E_Pf(Z_i))^2}{\sigma_P^2(f)}\right|\geq\varepsilon\,\text{some } n\geq N\right) \\
&\leq\sup_{P\in\mathcal{P}}P\left(\sup_{f\in\mathcal{F}_n,\sigma_P(f)\geq\sigma_n}\frac{2\overline{f}\vee1}{\sigma_n}\left|\frac{(E_n-E_P)(f(Z_i)-E_Pf(Z_i))^2}{\{E[(f(Z_i)-E_Pf(Z_i))^2]^2\}^{(1/2)}\vee\sigma_n}\right|\geq\varepsilon\,\text{some } n\geq N\right) \\
&\leq\sup_{P\in\mathcal{P}}P\left(\sup_{f\in\mathcal{F}_n,\sigma_P(f)\geq\sigma_n}\frac{\sqrt{n}}{\sqrt{\log n}}\left|\frac{(E_n-E_P)(f(Z_i)-E_Pf(Z_i))^2}{\{E[(f(Z_i)-E_Pf(Z_i))^2]^2\}^{(1/2)}\vee\sigma_n}\right|\geq c\varepsilon/(2\overline{f}\vee1)\,\text{some } n\geq N\right)
\end{aligned}$$

where the last inequality holds for $\sigma_n\sqrt{n/\log n}\geq c$. By Lemma A.1, this will go to zero if c is large enough so that $c\varepsilon/(2\overline{f}\vee1)$ is greater than the B for which the conclusion of Lemma A.1 holds for the class $\mathcal{G}_{n,P}^2$.

The probability that the second term in the last line of Equation 8 is greater than $\varepsilon>0$ for some $n\geq N$ goes to zero uniformly in $P\in\mathcal{P}$ by Lemma A.1 with the class $\{f-E_Pf(Z_i)|f\in\mathcal{F}_n\}$ taking the place of $\mathcal{F}_{n,P}$ in that lemma. □

Combining these lemmas gives a consistency result for classes of functions weighted by their standard deviations. The conditions are the same as those for Lemma A.2.

Lemma A.3. *Let $Z_1, \ldots, Z_n$ be iid observations and let $\mathcal{F}_n$ be a sequence of classes of functions and $\mathcal{P}$ a set of probability distributions such that, for some $\overline{f}$, $f(Z_i) \leq \overline{f}$ with P-probability one for $P \in \mathcal{P}$ and $f \in \mathcal{F}_n$ for each n. Let $\sigma_P(f) = (E_P f(Z_i)^2 - (E_P f(Z_i))^2)^{1/2}$. Define $\mathcal{G}^1_{n,P} = \{(f - E_P f(Z_i))\sigma_n/(\sigma_P(f) \vee \sigma_n)\}$ and $\mathcal{G}^2_{n,P} = \{(f - E_P f(Z_i))^2\sigma_n/(\mu_{2,P}([f - E_P f(Z_i)]^2) \vee \sigma_n)\}$, and suppose that, for some positive constants A and W,*

$$\sup_{P\in\mathcal{P}} \sup_{n\in\mathbb{N}} \sup_{Q} N(\varepsilon, \mathcal{G}^i_{n,P}, L_1(Q)) \leq A\varepsilon^{-W}$$

for $0 < \varepsilon < 1$ and $i = 1, 2$, where the supremum over Q is over all probability measures.

Then, for some B and c that do not depend on N or P, if $\sigma_n\sqrt{n/\log n} \geq c$ for all n,

$$\sup_{P\in\mathcal{P}} P\left(\sup_{f\in\mathcal{F}_n} \frac{\sqrt{n}}{\sqrt{\log n}} \left|\frac{f(Z_i) - E_P(f(Z_i))}{\hat{\sigma}(f) \vee \sigma_n}\right| \geq B \text{ some } n \geq N\right) \overset{N\to\infty}{\to} 0.$$

Proof. We have

$$\begin{aligned}
&P\left(\sup_{f\in\mathcal{F}_n} \frac{\sqrt{n}}{\sqrt{\log n}} \left|\frac{f(Z_i) - E_P(f(Z_i))}{\hat{\sigma}(f) \vee \sigma_n}\right| \geq B \text{some } n \geq N\right) \\
&= P\left(\sup_{f\in\mathcal{F}_n} \frac{\sqrt{n}}{\sqrt{\log n}} \left|\frac{f(Z_i) - E_P(f(Z_i))}{\sigma_P(f) \vee \sigma_n}\right| \frac{\sigma_P(f) \vee \sigma_n}{\hat{\sigma}(f) \vee \sigma_n} \geq B \text{some } n \geq N\right) \\
&\leq P\left(\sup_{f\in\mathcal{F}_n} \frac{\sqrt{n}}{\sqrt{\log n}} \left|\frac{f(Z_i) - E_P(f(Z_i))}{\sigma_P(f) \vee \sigma_n}\right| \geq B/2 \text{some } n \geq N\right) \\
&+ P\left(\inf_{f\in\mathcal{F}_n} \frac{\hat{\sigma}(f) \vee \sigma_n}{\sigma_P(f) \vee \sigma_n} \leq 1/2 \text{some } n \geq N\right).
\end{aligned}$$

The second to last line goes to zero uniformly in $P \in \mathcal{P}$ by Lemma A.1 applied to the classes $\{f - E_P(f) | f \in \mathcal{F}_n, P \in \mathcal{P}\}$ (here, B must be chosen large enough so that the conclusion of this lemma holds with B replaced by $B/2$). Since $\frac{\hat{\sigma}(f)\vee\sigma_n}{\sigma_P(f)\vee\sigma_n} \geq 1 > 1/2$ when $\sigma_P(f) < \sigma_n$, the last line is bounded by

$$P\left(\inf_{f\in\mathcal{F}_n, \sigma_P(f)\geq\sigma_n} \frac{\hat{\sigma}(f)}{\sigma_P(f)} \leq 1/2 \text{some } n \geq N\right),$$

which goes to zero uniformly in $P \in \mathcal{P}$ if $\sigma_n\sqrt{n/\log n} \geq c$ for c large enough by Lemma A.2. □

A.2 Constants in the Rate Bounds

The constant B in Theorem A.1 can be calculated using a careful inspection of the arguments in the proof. If $\mu_{2,n}\sqrt{n/\log n}$ goes to infinity, the bound (7) will go to zero for $a = 8$ (its minimum value). Then, with this value of a, (6) will be bounded by a constant times $\exp(-[(B^2/(128 \cdot 64) - W]\log n)$. For our purposes, it will suffice to have this converge to zero at any rate, and for this it will suffice that $W < B^2/(128 \cdot 64)$, which can be rearranged as $B > 64\sqrt{2W}$. This bound can be applied throughout the arguments of the previous section to obtain the following.

Lemma A.4. *Suppose that the conditions of Lemma A.3 hold with $\sigma_n\sqrt{n/\log n} \to \infty$. Then*

$$\sup_{P\in\mathcal{P}} P\left(\sup_{f\in\mathcal{F}_n} \frac{\sqrt{n}}{\sqrt{\log n}}\left|\frac{f(Z_i) - E_P f(Z_i)}{\hat{\sigma}(f) \vee \sigma_n}\right| \geq 64\sqrt{2W} + \eta\right) \overset{n\to\infty}{\to} 0.$$

Proof. The result with $\hat{\sigma}(f)$ replaced by $\sigma(f)$ follows by replacing $B = 64\sqrt{2W}$ in the arguments of Theorem A.1 as discussed above. With $\sigma_n\sqrt{n/\log n} \to \infty$, $\hat{\sigma}(f)/\sigma(f)$ will converge to one uniformly over $\hat{\sigma}(f) \geq \sigma_n$ by Lemma A.2, which then gives the result. □

This lemma can be used to obtain a feasible critical value for the variance weighted KS statistic, which is stated in the following theorem.

Theorem A.2. *Suppose that the conditions of Theorem 3.1 hold with $S(t) = \|t\|_-$, and that $\sigma_n\sqrt{n/\log n} \to \infty$. Then the critical value $\hat{c}_n$ in Theorem 3.1 can be taken to be $64\sqrt{2V} + \varepsilon$ for any $\varepsilon > 0$:*

$$\inf_{P\in\mathcal{P}} P(\Theta_0(P) \subseteq \mathcal{C}_n(64\sqrt{2V} + \varepsilon)) \overset{n\to\infty}{\to} 1$$

where V is the covering number index in Assumption 3.2.

While this result gives a feasible critical value for the procedure proposed in this paper, the resulting critical value will typically be very conservative. For example, for pointwise inference (fixing a point in Θ rather than taking the supremum over θ) using the class of functions $\{I(s < X_i < s + t)|s, t \in \mathbb{R}^{d_X}\}$, the class of functions $\mathcal{F}_{j,1}$ in Assumption 3.2 has covering number $2 \cdot (2d_X + 2)$ (the class $\mathcal{G}$ is VC subgraph with VC index $2d_X$ and, by the properties of VC classes this means that the class $\mathcal{F}_{j,1}$ with θ fixed is VC subgraph with index $2d_X + 2$, which translates to an exponent of $2(2d_X + 2)$ in the covering number by Lemma 25 in Pollard (1984)). This gives a critical value of $64\sqrt{2 \cdot 2 \cdot (2d_X + 2)} = 128\sqrt{2d_X + 1}$

plus a small constant. Under additional conditions, Armstrong and Chan (2012) obtain an asymptotic distribution result with a critical value that is asymptotically no greater than $\sqrt{2d_X}$ at this scaling. Thus, in this case, the bound obtained using the methods in this paper is conservative by a factor of just over 128.

A.3 Upper Bounds for the Rate of Convergence

If σ_n is fixed, we will have a $\sqrt{n}$ rate of uniform convergence for the KS statistic. The $\sqrt{n/\log n}$ rate of convergence results used in Theorem 3.1 do not rule this out for the case where σ_n goes to zero, but another argument shows that the rate of convergence will be strictly slower than $\sqrt{n}$ in many situations. See also the recent results of Armstrong and Chan (2012), which imply that the $\sqrt{n/\log n}$ rate derived in the present paper is exact for a particular choice of $\mathcal{G}$.

Assumption A.1. *For some $\theta \in \Theta_0(P)$, some j, and some open set $\mathcal{X}$, the following hold. (i) $E_P(m_j(W_i,\theta)|X_i) = 0$ a.s. on $\mathcal{X}$ and X_i has a density $f_X(x)$ on $\mathcal{X}$ that is bounded from above and from below away from zero. (ii) $var(m(W_i,\theta)|X_i = x)$ is continuous as a function of x and bounded away from zero and infinity on $\mathcal{X}$. (iii) $\mathcal{G}$ contains the function $t \mapsto k((t-x)/h)$ for all x and all h less than some fixed positive constant where k satisfies Assumption 5.2 and is continuous at zero.*

The assumption on the set of functions $\mathcal{G}$ covers many commonly used cases, including indicator sets for d_X dimensional rectangles or boxes.

Theorem A.3. *If Assumption A.1 holds and S satisfies Assumption 3.3, then, if $\sigma_n \to 0$, $\sqrt{n}T_n(\theta)$ will diverge to ∞.*

Proof. Fix any points $x_1, \ldots, x_\ell \in \mathcal{X}$. For k from 1 to ℓ, let $g_{n,k}(t) = k((t-x_k)/h_n)$

$$Z_{n,k} = \frac{1}{\hat{\sigma}_{n,j}(\theta, g_{n,k}) \vee \sigma_n} E_n m_j(W_i,\theta) k((X_i - x_k)/h_n)$$

where h_n is a sequence going to zero such that $h_n^{d_X/2}/\sigma_n$ goes to infinity and $h_n^{d_X} \geq n^{-\alpha}$ for some $\alpha < 1$. By the assumption on S, $\sqrt{n}T_n(\theta)$ will diverge to ∞ if $\inf_{x,h,j} \frac{1}{\hat{\sigma}_{n,j}(\theta,x,h)\vee\sigma_n} E_n m_j(W_i,\theta)k((X_i - x)/h)$ diverges to $-\infty$, and, for this, it is sufficient to show that $\min_k Z_{n,k}$ can be made arbitrarily small asymptotically by making ℓ large enough. Using standard arguments, it can be shown that $\hat{\sigma}_{n,j}(\theta, g_{n,k})/\sigma_{P,j}(\theta, g_{n,k})$ converges in probability to one, and, since

$\sigma_{P,j}(\theta, g_{n,k})/h_n^{d/2}$ converges to a constant under these assumptions, we also have that

$$Z_{n,k} = \frac{1}{\hat{\sigma}_{n,j}(\theta, g_{n,k})} E_n m_j(W_i, \theta) k((X_i - x_k)/h_n)$$

with probability approaching one. By the Lindeberg central limit theorem, defining

$$\tilde{Z}_{n,k} \equiv \frac{1}{\sigma_{P,j}(\theta, g_{n,k})} E_n m_j(W_i, \theta) k((X_i - x_k)/h_n)$$

$(\sqrt{n}\tilde{Z}_{n,1}, \ldots, \sqrt{n}\tilde{Z}_{n,\ell})$ converges to a vector of independent standard normal variables, so, since each $Z_{n,k}$ is eventually equal to $\tilde{Z}_{n,k}$ times something that converges to one, $(\sqrt{n}Z_{n,1}, \ldots, \sqrt{n}Z_{n,\ell})$ also converges to a vector of independent standard normal variables. Thus, $\min_k \sqrt{n}Z_{n,k}$ converges to the minimum of ℓ independent standard normal variables, which can be made arbitrarily small by making ℓ large.

□

A.4 Rates of Convergence for Slope Parameters

In this section of the appendix, I present a counterexample that shows that a condition along the lines of part (iii) of Assumption 6.4 is necessary to obtain the rate of convergence in Theorem 6.2. As discussed below, a similar counterexample shows that a condition on the parameter space Θ such as Assumption 6.3 is necessary in Theorem 6.1. These counterexamples also show that the first display in Assumption 4.4 cannot be replaced with an assumption that only takes into account the magnitude of the derivative vector.

In the counterexample considered in this section, both the weighted KS statistic based estimator and the estimator based on the KS statistic with bounded weights converge at a slower rate. While both rates are slower than they are under Assumption 4.4, the rate for the weighted KS statistic based estimator is still faster than the rate for the estimator based on bounded weights. This suggests that, while different conditions are needed to derive rates power results for local alternatives typified by this counterexample, these types of alternatives still favor weighting by a truncated estimate of the variance.

Consider an example where $E_P(W_i^H|X_i = x) = x^2$, $E_P(W_i^L|X_i = x) = -x^2$, $var(W_i^H|X_i) = var(W_i^L|X_i) = 1$, and X_i is has a uniform distribution on $[-1/2, 1/2]$. Suppose that we use the set of functions $\{I(s < X_i < s + t)|s \in \mathbb{R}, t \geq 0\}$. In this case, the identified set is a single point $(0, 0)$. Consider the sequence of local alternatives given by $\theta_n = (0, b_n)$. We

have, for all s, t with $-1/2 \leq s \leq s+t \leq 1/2$,

$$E_P[(W_i^H - b_n X_i) I(s < X_i < s+t)] = E_P[(X_i^2 - b_n X_i) I(s < X_i < s+t)]$$
$$= \int_s^{s+t} (x^2 - b_n x)\, dx = \int_s^{s+t} [(x - b_n/2)^2 - b_n^2/4]\, dx$$
$$\geq \int_{-t/2}^{t/2} [u^2 - b_n^2/4]\, du = 2\left[\frac{1}{3}u^3 - \frac{b_n^2}{4}u\right]_{u=0}^{t/2} = 2t\left[\frac{1}{24}t^2 - \frac{b_n^2}{8}\right]$$

and

$$var_P[(W_i^H - b_n X_i) I(s < X_i < s+t)] \geq E_P\{var_P[(W_i^H - b_n X_i) I(s < X_i < s+t) | X_i]\}$$
$$= E_P[I(s < X_i < s+t)] = t.$$

Thus, for s, t such that $E_P[(W_i^H - b_n X_i) I(s < X_i < s+t)]$ is negative,

$$\left| \frac{E_P[(W_i^H - b_n X_i) I(s < X_i < s+t)]}{\{var_P[(W_i^H - b_n X_i) I(s < X_i < s+t)]\}^{1/2}} \right| \leq 2t^{1/2} \left| \frac{1}{24}t^2 - \frac{b_n^2}{8} \right|_- \leq \frac{3^{1/4}}{4} b_n^{1/2} b_n^2.$$

A symmetric argument applies to moments based on W_i^L. For some constant K, this sequence of local alternatives will be in $\mathcal{C}_n(\hat{c}_n)$ if $b_n^{5/2} \leq K((\log n)/n)^{1/2}$ iff. $b_n \leq K((\log n)/n)^{1/5}$. In contrast, convergence to the identified set for one sided regression will be at a $((\log n)/n)^{2/5}$ rate if the parameter space Θ is restricted so that the absolute value of the slope parameter cannot be too large.

Using similar arguments, it can be shown that the set estimator with bounded weights will converge at an even slower rate. To see this, note that, by the calculations above,

$$\left| E_P[(W_i^H - b_n X_i) I(s < X_i < s+t)] \right|_- = 2t \left| \frac{1}{24}t^2 - \frac{b_n^2}{8} \right|_- \leq \frac{3^{1/2}}{4} b_n^3,$$

so this sequence of alternatives will be in $\mathcal{C}_{n,\omega}(\hat{c}_n)$ as long as, for a small enough constant C, $b_n^3 \leq C\hat{c}_n/n^{1/2}$, which can be rewritten as $b_n \leq C\hat{c}_n^{1/3}/n^{1/6}$. Thus, the set estimator based on a KS statistic with bounded weights will converge at the even slower $n^{-1/6}$ rate, in contrast to the $n^{-1/3}$ rate achieved when Assumption 4.4 holds.

Now consider the one sided regression model of Section 6.1 with $E_P(W_i^H | X_i = x) = x^2$ and the parameter space Θ given by $[0, \infty) \times \mathbb{R}$. That is, the parameter space Θ incorporates the prior knowledge that the intercept is nonnegative. Again, the identified set is the point $(0, 0)$, and the Hausdorff distance between the set estimate $\mathcal{C}_n(\hat{c}_n)$ and the identified set will

be at least b_n if $\mathcal{C}_n(\hat{c}_n)$ contains the point $(0, b_n)$. By the same argument used above, $(0, b_n)$ will be in $\mathcal{C}_n(\hat{c}_n)$ for some sequence b_n going to zero at a $((\log n)/n)^{1/5}$ rate, so that the rate of convergence of $\mathcal{C}_n(\hat{c}_n)$ to the identified set will be no faster than $((\log n)/n)^{1/5}$, which is slower than the $((\log n)/n)^{2/5}$ rate given by Theorem 6.1 when the intercept is not restricted. Note that, in the case where the intercept parameter is not restricted a priori, the sequence of local alternatives $(0, b_n)$ will still be in the estimate $\mathcal{C}_n(\hat{c}_n)$, but the distance of these points to the identified set will no longer be equal to b_n, since the identified set will contain a point $(\theta'(b_n), b_n)$ for some $\theta'(b_n)$ that is smaller in magnitude than b_n.

A.5 Covering Number Conditions

In this section, I state some simple sufficient conditions under which Assumption 3.2 holds. I first prove that Assumption 3.2 holds under individual bounds on the complexity of the classes $\mathcal{G}$ and $\{w \mapsto m(w, \theta) | \theta \in \Theta\}$. The proof of this result uses Lemma A.5, stated and proved at the end of the section. I then provide examples of classes $\mathcal{G}$ that satisfy these bounds, and show that the class $\{w \mapsto m(w, \theta) | \theta \in \Theta\}$ satisfies these bounds in each of the applications covered in Section 6. Throughout this section, I define $\mathcal{F}_m \equiv \{w \mapsto m(w, \theta) | \theta \in \Theta\}$ to be the class of moment functions indexed by θ.

The following theorem translates bounds on the covering numbers of the classes $\mathcal{G}$ and $\{w \mapsto m(w, \theta) | \theta \in \Theta\}$ to the conditions of Assumption 3.2.

Theorem A.4. *Suppose that the classes $\mathcal{F}_m \equiv \{w \mapsto m(w, \theta) | \theta \in \Theta\}$ and $\mathcal{G}$ are uniformly bounded and satisfy $\sup_Q N(\varepsilon, \mathcal{F}_m, L_1(Q)) \leq A\varepsilon^{-W}$ and $\sup_Q N(\varepsilon, \mathcal{G}, L_1(Q)) \leq A\varepsilon^{-W}$ for some $A, W > 0$ where the supremum is over all probability measures Q. Then Assumption 3.2 holds.*

Proof. The result follows immediately from Lemma A.5, since the classes of functions in Assumption 3.2 are sums and products of these bounded classes and bounded classes of constant functions, which also have polynomial uniform covering numbers. □

With this result in hand, we can verify Assumption 3.2 for a particular model and choice of $\mathcal{G}$ using results stated in Pollard (1984), van der Vaart and Wellner (1996) and other sources. For convenience, I do this here for some choices of $\mathcal{G}$.

Theorem A.5. *Suppose that $\mathcal{F}_m \equiv \{w \mapsto m(w, \theta) | \theta \in \Theta\}$ and $\mathcal{G}$ are uniformly bounded $\sup_Q N(\varepsilon, \mathcal{F}_m, L_1(Q)) \leq A\varepsilon^{-W}$. Then Assumption 3.2 will hold for the following classes of functions $\mathcal{G}$:*

(i) The class of indicator functions $\mathcal{G} = \{x \mapsto I(x \in V) | V \in \mathcal{V}\}$ *for any VC class of sets* $\mathcal{V}$.

(ii) The class of dilations of a kernel function k *given by* $\mathcal{G} = \{x \mapsto k((x-t)/h) | x \in \mathbb{R}^{d_X}, h \in \mathbb{R}_+\}$ *for any kernel function* k *given by* $k(x) = r(\|x\|)$ *for a decreasing, bounded function* r *on* $\mathbb{R}_+$.

Proof. The covering number bound for $\mathcal{G}$ in Theorem A.4 holds by Lemma 25 in Pollard (1984) (since a VC class of sets has polynomial discrimination) for part (i), and by problem 18 in Chapter 2 of Pollard (1984) for part (ii). □

See Pollard (1984) for the definition of a VC class and examples of VC classes of sets. The class of all d_X dimensional rectangles falls into this category. The condition that the class of functions $\mathcal{F}_m = \{w \mapsto m(w,\theta) | \theta \in \Theta\}$ satisfy the covering number bound $\sup_Q N(\varepsilon, \mathcal{F}_m, L_1(Q)) \leq A\varepsilon^{-W}$ can be verified on a case by case basis using general results such as those in Pollard (1984) and van der Vaart and Wellner (1996). I do this for the examples in this paper in the next theorem.

Theorem A.6. *The class of moment functions* $\mathcal{F}_m = \{w \mapsto m(w,\theta) | \theta \in \Theta\}$ *satisfies the covering number bound* $\sup_Q N(\varepsilon, \mathcal{F}_m, L_1(Q)) \leq A\varepsilon^{-W}$ *in all of the models of Section 6 as long as the data are bounded and* Θ *is compact in the conditional mean models of Sections 6.1, 6.2 and B.3.*

Proof. The class $\{w \mapsto m(w,\theta) | \theta \in \Theta\}$ has VC subgraph for all of the models of Section 6, so the result follows from Lemma 25 in Pollard (1984). □

The proof of Theorem A.4 uses the following lemma, which modifies an argument from van der Vaart and Wellner (1996).

Lemma A.5. *Let* $\mathcal{F}$, $\mathcal{G}$ *and* $\mathcal{H}$ *be classes of functions bounded by a fixed constant* B, *and let* $\mathcal{F} \cdot \mathcal{G} + \mathcal{H} = \{f \cdot g + h | f \in \mathcal{F}, g \in \mathcal{G}, h \in \mathcal{H}\}$. *Suppose that, for some* $A, W > 0$, $\sup_Q N(\varepsilon, \mathcal{F}, L_1(Q)) \leq A\varepsilon^{-W}$, *where the supremum is taken over all probability measures, and that the same statement holds with* $\mathcal{F}$ *replaced by* $\mathcal{G}$ *and* $\mathcal{H}$. *Then* $\sup_Q N(\varepsilon, \mathcal{F} \cdot \mathcal{G} + \mathcal{H}, L_1(Q)) \leq A^3(2B+1)^{3W}\varepsilon^{-3W}$, *where the supremum is again taken over all probability measures.*

Proof. The result follows from an argument similar to the proof of Theorem 2.10.20 in van der Vaart and Wellner (1996). Given $\varepsilon > 0$ and a probability measure Q, let $k_{\mathcal{F},Q} =$

$N(\varepsilon, \mathcal{F}, L_1(Q)) \leq \sup_{Q'} N(\varepsilon, \mathcal{F}, L_1(Q'))$ and let $f_{1,Q}, \ldots, f_{k_{\mathcal{F},Q},Q}$ be such that, for all $f \in \mathcal{F}$, there exists a $f_{i,Q}$ such that $E_Q|f_{i,Q}(Z_i) - f(Z_i)| \leq \varepsilon$ (here, the notation $E_Q f(Z_i)$ refers to the expectation $\int f(z)\, dQ(z)$ of $f(Z_i)$ for Z_i a random variable with distribution Q). Define $k_{\mathcal{G},Q}$, $k_{\mathcal{H},Q}$, $g_{1,Q}, \ldots, g_{k_{\mathcal{G},Q},Q}$ and $h_{1,Q}, \ldots, h_{k_{\mathcal{H},Q},Q}$ similarly. For any $fg + h \in \mathcal{F} \cdot \mathcal{G} + \mathcal{H}$, there is some $j_\mathcal{F}$, $j_\mathcal{G}$ and $j_\mathcal{H}$ such that $E_Q|f_{j_\mathcal{F},Q}(Z_i) - f(Z_i)| \leq \varepsilon$, $E_Q|g_{j_\mathcal{F},Q}(Z_i) - g(Z_i)| \leq \varepsilon$ and $E_Q|h_{j_\mathcal{H},Q}(Z_i) - h(Z_i)| \leq \varepsilon$. We have, for all z,

$$
\begin{aligned}
&|f(z)g(z) + h(z) - (f_{j_\mathcal{F},Q}(z)g_{j_\mathcal{G},Q}(z) + h_{j_\mathcal{H},Q}(z))| \\
&= |(f(z) - f_{j_\mathcal{F},Q}(z))g(z) + (g(z) - g_{j_\mathcal{G},Q}(z))f_{j_\mathcal{F},Q}(z) + h(z) - h_{j_\mathcal{H},Q}(z))| \\
&\leq |f(z) - f_{j_\mathcal{F},Q}(z)| \cdot |g(z)| + |g(z) - g_{j_\mathcal{G},Q}(z)| \cdot |f_{j_\mathcal{F},Q}(z)| + |h(z) - h_{j_\mathcal{H},Q}(z))| \\
&\leq |f(z) - f_{j_\mathcal{F},Q}(z)| \cdot B + |g(z) - g_{j_\mathcal{G},Q}(z)| \cdot B + |h(z) - h_{j_\mathcal{H},Q}(z))|
\end{aligned}
$$

so that

$$
\begin{aligned}
&E_Q|f(Z_i)g(Z_i) + h(Z_i) - (f_{j_\mathcal{F},Q}(Z_i)g_{j_\mathcal{G},Q}(Z_i) + h_{j_\mathcal{H},Q}(Z_i))| \\
&\leq (E_Q|f(Z_i) - f_{j_\mathcal{F},Q}(Z_i)| + E_Q|g(Z_i) - g_{j_\mathcal{G},Q}(Z_i)|)B + E_Q|h(Z_i) - h_{j_\mathcal{H},Q}(Z_i))| \leq (2B+1)\varepsilon.
\end{aligned}
$$

Since Q was arbitrary, it follows that $\sup_Q N((2B+1)\varepsilon, \mathcal{F}\cdot\mathcal{G}+\mathcal{H}, L_1(Q)) \leq (\sup_Q N(\varepsilon, \mathcal{F}, L_1(Q))) \cdot (\sup_Q N(\varepsilon, \mathcal{G}, L_1(Q))) \cdot (\sup_Q N(\varepsilon, \mathcal{G}, L_1(Q))) \leq A^3\varepsilon^{-3W}$. Replacing ε with $\varepsilon/(2B+1)$ gives the result. □

A.6 Proofs

This section of the appendix contains proofs of the results stated in the body of the paper and in Section B of the appendix.

proof of Theorem 3.1. If $\Theta_0(P) \nsubseteq \mathcal{C}_n(\hat{c}_n)$, then, for some $\theta_0 \in \Theta_0(P)$, $\sqrt{n/\log n} T_n(\theta_0) \geq \hat{c}_n$ so that for some $g \in \mathcal{G}$,

$$
S\left(\frac{\hat{\mu}_{n,1}(\theta, g)}{\hat{\sigma}_{n,1}(\theta, g) \vee \sigma_n}, \ldots, \frac{\hat{\mu}_{n,d_Y}(\theta, g)}{\hat{\sigma}_{n,d_Y}(\theta, g) \vee \sigma_n}\right) \geq \frac{\hat{c}_n \sqrt{\log n}}{\sqrt{n}}
$$

so that, for some j, $\frac{\hat{\mu}_{n,j}(\theta,g)}{\hat{\sigma}_{n,j}(\theta,g)\vee\sigma_n} \leq -\frac{\hat{c}_n\sqrt{\log n}}{\sqrt{n}} K_{S,1}$. Since $\theta_0 \in \Theta_0(P)$, $E_P m(W_i, \theta_0) g(X_i) \geq 0$, so this implies that

$$
\frac{\sqrt{n}}{\sqrt{\log n}} \frac{(E_n - E_P) m(W_i, \theta_0) g(X_i)}{\hat{\sigma}_{n,j}(\theta, g) \vee \sigma_n} \leq -\hat{c}_n K_{S,1}.
$$

Thus, $\Theta_0(P) \nsubseteq \mathcal{C}_n(\hat{c}_n)$ implies that the above display holds for some θ_0, g, and j. If K is large enough so that the conclusion of Lemma A.3 holds for $B = K \cdot K_{S,1}$. and c from that lemma equal to K, the probability that there exist some $\theta_0 \in \Theta_0(P)$ and $g \in \mathcal{G}$ such that this event holds and $\hat{c}_n$ and $\sigma_n\sqrt{n/\log n}$ are greater than K will be bounded by a sequence that goes to zero uniformly in $P \in \mathcal{P}$.

□

proof of Theorem 4.1. If $d_H(\Theta_0(P), \mathcal{C}_n(\hat{c}_n)) > \varepsilon$ and $\Theta_0(P) \subseteq \mathcal{C}_n(\hat{c}_n)$, then there exists some $\theta \in \mathcal{C}_n(\hat{c}_n)$ such that $d_H(\theta, \Theta_0(P)) > \varepsilon$. Letting δ be such that, for all $P \in \mathcal{P}$, $E_P m_j(W_i, \theta) g_j(X_i) < -\delta$ for some j and $g \in \mathcal{G}$, this implies that, once $\hat{\sigma}_{n,j}(\theta', g)$ is bounded uniformly in (θ', g) by some $\bar{\sigma}$ (this happens with probability approaching one uniformly in $P \in \mathcal{P}$ by Lemma A.2),

$$-T_n(\theta) \leq \frac{1}{K_{S,2}\bar{\sigma}} \left(E_n m_j(W_i, \theta) g_j(X_i) \vee 0\right) \leq -\frac{1}{K_{S,2}\bar{\sigma}} \left(\delta - \sup_{\theta', g, k} |(E_n - E_P) m_k(W_i, \theta') g_k(X_i)|\right).$$

The probability that $\sup_{\theta', g, k} |(E_n - E_P) m_k(W_i, \theta') g_k(X_i)| \leq \delta/2$ goes to one uniformly in $P \in \mathcal{P}$ by Lemma A.1, and once this holds, the above display will imply $T_n(\theta) \geq \delta/(2K_{S,2}\bar{\sigma})$. This cannot hold for $\theta \in \mathcal{C}_n(\hat{c}_n)$ for $\hat{c}_n\sqrt{(\log n)/n} \leq \delta/(2K_{S,2}\bar{\sigma})$, and the probability of this holding goes to zero uniformly in $P \in \mathcal{P}$.

□

proof of Theorem 4.2. If $d_H(\Theta_0(P), \mathcal{C}_n(\hat{c}_n)) > B\left(\frac{\hat{c}_n^2 \log n}{n}\right)^{\gamma/2}$, $\Theta_0(P) \subseteq \mathcal{C}_n(\hat{c}_n)$ and $d_H(\mathcal{C}_n(\hat{c}_n, \Theta_0(P)) \leq \delta$ (the latter two events hold with probability approaching one uniformly in $P \in \mathcal{P}$ by Theorems 3.1 and 4.1), then there exists some $\theta \in \mathcal{C}_n(\hat{c}_n)$ such that $d_H(\theta, \Theta_0(P)) > B\left(\frac{\hat{c}_n^2 \log n}{n}\right)^{\gamma/2}$. For this θ (and P), there will be, by Assumption 4.2, a $g^* \in \mathcal{G}$ and j^* such that

$$\frac{\mu_{P,j^*}(\theta, g^*)}{\sigma_{P,j^*}(\theta, g^*) \vee \left[B^{1/\gamma}\left(\frac{\hat{c}_n^2 \log n}{n}\right)^{\psi/2}\right]} \leq -(C/2)B^{1/\gamma}\left(\frac{\hat{c}_n^2 \log n}{n}\right)^{1/2}$$

(replacing C with $C/2$ takes care of the possibility that the infimum in the assumption is not achieved) and, by part (ii), for some constant $\eta > 0$ that does not depend on P, this will eventually imply

$$\frac{\mu_{P,j^*}(\theta, g^*)}{\sigma_{P,j^*}(\theta, g^*) \vee (\eta\sigma_n)} \leq -(C/2)B^{1/\gamma}\left(\frac{\hat{c}_n^2 \log n}{n}\right)^{1/2}$$

so that, letting $C_1 = (C/2)(\eta \wedge 1)$, we will have $\frac{\mu_{P,j^*}(\theta,g^*)}{\sigma_{P,j^*}(\theta,g^*)\vee\sigma_n} \leq -C_1 B^{1/\gamma}\left(\frac{\hat{c}_n^2 \log n}{n}\right)^{1/2}$. Since $\theta \in \mathcal{C}_n(\hat{c}_n)$, we will also have $T_n(\theta) \leq \hat{c}_n\left(\frac{\log n}{n}\right)^{1/2}$, so that, for all $g \in \mathcal{G}$ and all j, $\frac{\hat{\mu}_{n,j}(\theta,g)}{\hat{\sigma}_{n,j}(\theta,g)\vee\sigma_n} \geq -K_{S,2}\hat{c}_n\left(\frac{\log n}{n}\right)^{1/2}$. By Lemma A.2, this will also imply $\frac{\hat{\mu}_{n,j}(\theta,g)}{\sigma_{P,j}(\theta,g)\vee\sigma_n} \geq -\frac{K_{S,2}\hat{c}_n}{2}\left(\frac{\log n}{n}\right)^{1/2}$ with probability approaching one uniformly in $P \in \mathcal{P}$. When these events all hold, we will have

$$\frac{\hat{\mu}_{n,j^*}(\theta,g^*)}{\sigma_{P,j^*}(\theta,g^*)\vee\sigma_n} - \frac{\mu_{P,j^*}(\theta,g^*)}{\sigma_{P,j^*}(\theta,g^*)\vee\sigma_n} \geq -\frac{K_{S,2}\hat{c}_n}{2}\left(\frac{\log n}{n}\right)^{1/2} + C_1 B^{1/\gamma}\left(\frac{\hat{c}_n^2 \log n}{n}\right)^{1/2}$$

so that

$$\sup_{\theta\in\Theta, g\in\mathcal{G}, j\in\{1,\ldots,j\}} \frac{\sqrt{n}}{\sqrt{\log n}}\left|\frac{\hat{\mu}_{n,j}(\theta,g)}{\sigma_{P,j}(\theta,g)\vee\sigma_n} - \frac{\mu_{P,j}(\theta,g)}{\sigma_{P,j}(\theta,g)\vee\sigma_n}\right| \geq \hat{c}_n(B^{1/\gamma}C_1 - K_{S,2}/2).$$

Since $\hat{c}_n$ is bounded away from zero, we can choose B large so that $\hat{c}_n(B^{1/\gamma}C_1 - K_{S,2}/2)$ is large enough so that the conclusion of Lemma A.1 holds with B from that lemma replaced by $\hat{c}_n(B^{1/\gamma}C_1 - K_{S,2}/2)$. For this value of B, the probability of the last display holding will go to zero uniformly in $P \in \mathcal{P}$ so that the desired conclusion will hold. □

proof of Theorem 4.3. It is sufficient to find a C such that, given θ and P, there exists a $\theta_0(\theta,P)$, $j_0(\theta,P)$, and a $g \in \mathcal{G}$ such that

$$\frac{\mu_{P,j}(\theta,g)}{\sigma_{P,j}(\theta,g)\vee d(\theta,\theta_0(P))^{\psi/\gamma}} \leq -C\|\theta - \theta_0(\theta,P)\|^{1/\gamma}.$$

Given θ and P, let $\theta_0(\theta,P)$ and $j_0(\theta,P)$ be chosen as in Assumption 4.4. To avoid cumbersome notation, I will use θ_0 and j_0 to denote $\theta_0(\theta,P)$ and $j_0(\theta,P)$ when the dependence on θ and P is clear. For this θ_0 and j_0, we will have, for $\|x - x_0\| < \eta$,

$$\begin{aligned}
&\bar{m}_{j_0}(\theta,x,P) = \bar{m}_{j_0}(\theta,x,P) - \bar{m}_{j_0}(\theta_0,x_0,P)\\
&= [\bar{m}_{j_0}(\theta,x,P) - \bar{m}_{j_0}(\theta_0,x,P)] + [\bar{m}_{j_0}(\theta_0,x,P) - \bar{m}_{j_0}(\theta_0,x_0,P)]\\
&\leq \bar{m}_{\theta,j_0}(\theta^*,x,P)(\theta - \theta_0) + C\|x - x_0\|^\alpha
\end{aligned}$$

for some θ^* between θ and θ_0. By Assumptions 4.3 and 4.4, for $\|\theta - \theta_0\|$ and $\|x - x_0\|$ smaller than some constant that does not depend on P or θ, this will be less than or equal to

$$-(\eta/2)\|\theta - \theta_0\| + C\|x - x_0\|^\alpha.$$

For $\|x - x_0\| \leq [\eta/(4C)]^{1/\alpha}\|\theta - \theta_0\|^{1/\alpha}$, this is less than or equal to $-(\eta/4)\|\theta - \theta_0\|$. Thus, letting $g \in \mathcal{G}$ be as in Assumption 4.6 with $s = x_0$ and $t = [\eta/(4C)]^{1/\alpha}\|\theta - \theta_0\|^{1/\alpha}$ so that $g(x) \leq I(\|x - x_0\| \leq [\eta/(4C)]^{1/\alpha}\|\theta - \theta_0\|^{1/\alpha})$ and $g(x) \geq C_{\mathcal{G},1} I(\|x - x_0\| \leq [\eta/(4C)]^{1/\alpha}\|\theta - \theta_0\|^{1/\alpha} C_{\mathcal{G},2})$, we will have

$$\mu_{P,j_0}(\theta, g) = E_P \bar{m}_{j_0}(\theta, X_i, P) g(X_i) \leq -(\eta/4)\|\theta - \theta_0\| E_P g(X_i) \tag{9}$$

and

$$\begin{aligned}
&\sigma_{P,j_0}(\theta, g) = \{var_P[m_{j_0}(W_i, \theta) g(X_i)]\}^{1/2} \leq \left\{E_P[m_{j_0}(W_i, \theta) g(X_i)]^2\right\}^{1/2} \\
&\leq \overline{Y}\overline{g}^{1/2} \left\{E_P g(X_i)\right\}^{1/2}.
\end{aligned}$$

The lower bound on g implies that $\{E_P g(X_i)\}^{1/2}$ is greater than some constant that does not depend on P times $\|\theta - \theta_0\|^{d_X/(2\alpha)} \geq d(\theta, \theta_0(P))^{d_X/(2\alpha)}$. Thus, for some constant K that does not depend on P, $\sigma_{P,j_0}(\theta, g) \vee d(\theta, \theta_0(P))^{d_X/(2\alpha)} \leq K\{E_P g(X_i)\}^{1/2}$. Thus,

$$\begin{aligned}
&\frac{\mu_{P,j_0}(\theta, g)}{\sigma_{P,j_0}(\theta, g) \vee d(\theta, \theta_0(P))^{d_X/(2\alpha)}} \leq \frac{-(\eta/4)}{K}\|\theta - \theta_0\|[E_P g(X_i)]^{1/2} \\
&\leq \frac{-(\eta/4)}{K}\|\theta - \theta_0\| C_{\mathcal{G},1}^{1/2} P\left\{\|x - x_0\| \leq [\eta/(4C)]^{1/\alpha}\|\theta - \theta_0\|^{1/\alpha} C_{\mathcal{G},2}\right\}^{1/2} \\
&\leq \frac{-(\eta/4)}{K}\|\theta - \theta_0\| C_{\mathcal{G},1}^{1/2} \eta^{1/2} \left\{[\eta/(4C)]^{1/\alpha}\|\theta - \theta_0\|^{1/\alpha} C_{\mathcal{G},2}\right\}^{d_X/2}
\end{aligned}$$

where the second inequality follows from the lower bound on g. This is equal to a negative constant that does not depend on P times $\|\theta - \theta_0\|^{(d_X+2\alpha)/(2\alpha)}$, so that Assumption 4.2 holds with $\gamma = 2\alpha/(d_X + 2\alpha)$ and $\psi = d_X/(d_X + 2\alpha)$.

□

proof of Theorem 6.1. Assumption 4.3 holds because $\bar{m}(\theta, x)$ is linear, so it remains to verify Assumption 4.4. Given $\theta \in \Theta$ and $P \in \mathcal{P}$, let $x_0(\theta, P)$ minimize $E_P(W_i^H | X_i = x) - \theta_1 - x'\theta_{-1}$ over the support of X_i, and let $t(\theta, P)$ be the minimum (the minimum is taken since $E(W_i^H | X_i = x) - \theta_1 - x'\theta_{-1}$ is continuous). Let $\theta_0(\theta, P) = (\theta_1 + t(\theta, P), \theta_{-1})$. Then $\bar{m}(\theta_0(\theta, P), x, P) = E(W_i^H | X_i = x) - \theta_1 - t(\theta, P) - x'\theta_{-1}$ so that $\theta_0(\theta, P) \in \Theta_0(P)$ and $\bar{m}(\theta_0(\theta, P), x_0(\theta, P), P) = 0$. We have

$$\begin{aligned}
&\bar{m}_\theta(\theta_0(\theta, P), x_0(\theta, P), P)(\theta - \theta_0(\theta, P)) = -(1, x_0(\theta, P)')(-t(\theta, P), 0, \ldots, 0)' \\
&= t(\theta, P) = -\|(t(\theta, P), 0, \ldots, 0)'\| = -\|\theta - \theta_0(\theta, P)\|
\end{aligned}$$

where the second to last equality holds because $t(\theta, P)$ is negative by definition of the identified set. The Hölder continuity part of Assumption 4.4 is immediately implied by Assumption 6.1. Under Assumption 6.2, $x_0(\theta, P)$ must be on the interior of the support of X_i by part (ii) of this assumption. Thus, $x_0(\theta, P)$ is an interior minimum of the twice differentiable function $x \mapsto E_P(W_i^H | X_i = x) - \theta_1 - x'\theta_{-1}$, so the first derivative of this function at $x_0(\theta, P)$ is zero. This and a second order mean value expansion of this function around $x_0(\theta, P)$ imply the Hölder continuity part of Assumption 4.4 with C a bound on the norm of the second derivative matrix.

□

proof of Theorem 6.2. Everything is the same as in the proof of Theorem 6.1 except for the verification of the first part of Assumption 4.4. For any θ, either (θ_1', θ_2) is in $\Theta_0(P)$ for some θ', in which case the same argument to verify Assumption 4.4 goes through, or $\theta_2 > \overline{\theta}_2(P)$ or $\theta_2 < \underline{\theta}_2(P)$. Let $(\theta_1', \overline{\theta}_2) = (\theta_1^u(P), \overline{\theta}_2(P))$, $x_{0,1} = x_{0,1}^u(P)$ and $x_{0,2} = x_{0,2}^u(P)$ (using the notation defined before Assumption 6.4). so that $x_{0,2} < x_{0,1}$ and $E(W_i^H | X_i = x_{0,1}) = \theta_1' + x_{0,1}\overline{\theta}_2$ and $E(W_i^L | X_i = x_{0,2}) = \theta_1' + x_{0,2}\overline{\theta}_2$. We have $\bar{m}_{\theta,1}(\theta, x, P) = -(1, x)$ and $\bar{m}_{\theta,2}(\theta, x, P) = (1, x)$, so that

$$\bar{m}_{\theta,1}(\theta, x_{0,1}, P)(\theta - (\theta_1', \overline{\theta}_2)) = -(1, x_{0,1})(\theta - (\theta_1', \overline{\theta}_2))$$

and

$$\bar{m}_{\theta,2}(\theta, x_{0,2}, P)(\theta - (\theta_1', \overline{\theta}_2)) = (1, x_{0,2})(\theta - (\theta_1', \overline{\theta}_2)).$$

If the sum of the expressions in these two displays is less than $-2\eta\|\theta - (\theta_1', \overline{\theta}_2)\|$, at least one of them must be less than $-\eta\|\theta - (\theta_1', \overline{\theta})\|$, so it suffices to bound

$$[(1, x_{0,2}) - (1, x_{0,1})](\theta - (\theta_1', \overline{\theta}_2))/\|\theta - (\theta_1', \overline{\theta}_2)\| = -\frac{(x_{0,1} - x_{0,2})(\theta_2 - \overline{\theta}_2)}{\left[(\theta_1 - \theta_1')^2 + (\theta_2 - \overline{\theta}_2)^2\right]^{1/2}}.$$

For this, it suffices to bound $x_{0,1} - x_{0,2}$ away from zero and $|\theta_1 - \theta_1'|/|\theta_2 - \overline{\theta}_2|$ away from infinity.

$x_{0,1} - x_{0,2}$ is bounded away from zero by part (iii) of Assumption 6.4. For parameter

values where $|\theta_1 - \theta_1'|/|\theta_2 - \overline{\theta}_2|$ is large, we can use another argument. Note that

$$\frac{-(1, x_{0,1})(\theta - (\theta_1', \overline{\theta}_2))}{\|\theta - (\theta_1', \overline{\theta}_2)\|} = -\frac{(\theta_1 - \theta_1') + x_{0,1}(\theta_2 - \overline{\theta}_2)}{\|\theta - (\theta_1', \overline{\theta}_2)\|} = -\frac{(\theta_1 - \theta_1')/(\theta_2 - \overline{\theta}_2) + x_{0,1}}{\left[(\theta_1 - \theta_1')^2/(\theta_2 - \overline{\theta}_2)^2 + 1\right]^{1/2}}$$

and, similarly,

$$\frac{(1, x_{0,2})(\theta - (\theta_1', \overline{\theta}_2))}{\|\theta - (\theta_1', \overline{\theta}_2)\|} = \frac{(\theta_1 - \theta_1')/(\theta_2 - \overline{\theta}_2) + x_{0,2}}{\left[(\theta_1 - \theta_1')^2/(\theta_2 - \overline{\theta}_2)^2 + 1\right]^{1/2}}.$$

For $|\theta_1 - \theta_1'|/|\theta_2 - \overline{\theta}_2| > 2\max\{|x_{0,1}|, |x_{0,2}|, 1\}$, one of these displays will be less than $-1/4$. □

proof of Theorem B.1. For Assumption 4.3, note that

$$\bar{m}_\theta(\theta, x, P) = \frac{d}{d\theta} E_P[\tau - I(W_i^H \leq \theta_1 + X_i'\theta_{-1})|X_i = x] = -\frac{d}{d\theta} P(W_i^H \leq \theta_1 + X_i'\theta_{-1}|X_i = x)$$
$$= -f_{W_i^H|X_i}(\theta_1 + x'\theta_{-1}|x)(1, x').$$

This is continuous as a function of θ uniformly in (θ, x, P) by Assumption B.3 and the bound on the support of X_i.

To verify the first part of Assumption 4.4, let $x_0(\theta, P)$, $t(\theta, P)$ and $\theta_0(\theta, P)$ be defined as in the proof of Theorem 6.1, but with $E_P(W_i^H|X_i = x)$ replaced by $q_{\tau,P}(W_i^H|X_i = x)$. Then $\theta_0(\theta, P) \in \Theta_0(P)$ and

$$\bar{m}(\theta_0(\theta, P), x_0(\theta, P), P) = \tau - P(W_i^H \leq \theta_1 + t(\theta, P) + X_i'\theta_{-1}|X_i = x_0(\theta, P)) = 0$$

since $q_{\tau,P}(W_i^H|X_i = x_0(\theta, P)) = \theta_1 + t(\theta, P) + x_0(\theta, P)'\theta_{-1}$. We also have

$$m_\theta(\theta_0(\theta, P), x_0(\theta, P), P)(\theta - \theta_0(\theta, P)) = -f_{W_i^H|X_i}(\theta_1 + x'\theta_{-1}|x)(1, x')(-t(\theta, P), 0, \ldots, 0)'$$
$$= f_{W_i^H|X_i}(\theta_1 + x'\theta_{-1}|x)t(\theta, P) = -f_{W_i^H|X_i}(\theta_1 + x'\theta_{-1}|x)\|\theta - \theta_0(\theta, P)\| \leq -\underline{f}\|\theta - \theta_0(\theta, P)\|.$$

For the second part of Assumption 4.4, note that, since $\theta_0 = \theta_0(\theta, P) \in \Theta_0(P)$,

$$\begin{aligned}
&\bar{m}(\theta_0, x, P) = \tau - P(W_i^H \leq \theta_{0,1} + X_i'\theta_{0,-1} | X_i = x) \\
&= \tau - P(W_i^H \leq q_{\tau,P}(W_i^H|X_i = x)|X_i = x) \\
&+ P(\theta_{0,1} + X_i'\theta_{0,-1} \leq W_i^H \leq q_{\tau,P}(W_i^H|X_i = x)|X_i = x) \\
&= P(\theta_{0,1} + X_i'\theta_{0,-1} \leq W_i^H \leq q_{\tau,P}(W_i^H|X_i = x)|X_i = x).
\end{aligned}$$

For $\|x - x_0\|$ small enough, the distance between $\theta_{0,1} + x'\theta_{0,-1}$ and $q_{\tau,P}(W_i^H|X_i = x)$ will be less than the η in Assumption B.3. For x such that this holds,

$$\begin{aligned}
&|\bar{m}(\theta_0, x, P) - \bar{m}(\theta_0, x_0, P)| = \bar{m}(\theta_0, x, P) \\
&= P(\theta_{0,1} + X_i'\theta_{0,-1} \leq W_i^H \leq q_{\tau,P}(W_i^H|X_i = x)|X_i = x) \\
&\leq \overline{f}[q_{\tau,P}(W_i^H|X_i = x) - \theta_{0,1} - x'\theta_{0,-1}] \\
&= \overline{f}\{[q_{\tau,P}(W_i^H|X_i = x) - \theta_{0,1} - x'\theta_{0,-1}] - [q_{\tau,P}(W_i^H|X_i = x_0) - \theta_{0,1} - x_0'\theta_{0,-1}]\}.
\end{aligned}$$

Under Assumption B.1, the second part of Assumption 4.4 then follows immediately since, for $\alpha \leq 1$ and $\|x - x_0\|$ small enough, $\|(x - x_0)'\theta_{0,-1}\| \leq \|\theta_{0,-1}\|\|x - x_0\| \leq \|\theta_{0,-1}\|\|x - x_0\|^\alpha$ so that the expression in the above display is bounded by $\overline{f}(C + \|\theta_{0,-1}\|)\|x - x_0\|^\alpha$. Under Assumption B.2, Assumption 4.4 follows from a second order mean value expansion of $q_{\tau,P}(W_i^H|X_i = x_0)$ since x_0 is on the interior of the support of X_i.

□

proof of Theorem B.2. Everything is the same as in the proof of Theorem B.1 except for the verification of the first part of Assumption 4.4. Verifying this condition uses a similar argument to the one in Theorem 6.2 for mean regression. For any θ, either $(\theta_1', \theta_2) \in \Theta_0(P)$ for some θ', in which case the same argument to verify Assumption 4.4 goes through, or $\theta_2 > \overline{\theta}_2(P)$ or $\theta_2 < \underline{\theta}_2(P)$, where $\overline{\theta}_2(P)$ and $\underline{\theta}_2(P)$ are defined as in the assumptions of Theorem 6.2 ($\overline{\theta}_2(P) \equiv \sup\{\theta_2|(\theta_1, \theta_2) \in \Theta_0(P) \text{ some } \theta_1\}$ and $\underline{\theta}_2(P) \equiv \inf\{\theta_2|(\theta_1, \theta_2) \in \Theta_0(P) \text{ some } \theta_1\}$). If $\theta_2 > \overline{\theta}_2(P)$ (a symmetric argument applies when $\theta_2 < \underline{\theta}_2(P)$), then, for $(\theta_1', \overline{\theta}_2) = (\theta_1^u(P), \overline{\theta}_2(P))$, $x_{0,1} = x_{0,1}^{u,q}(P)$ and $x_{0,2} = x_{0,2}^{u,q}(P)$, we have $x_{0,2} < x_{0,1}$ and $q_{\tau,P}(W_i^H|X_i = x_{0,1}) = \theta_1' + x_{0,1}\overline{\theta}_2$ and $q_{\tau,P}(W_i^L|X_i = x_{0,2}) = \theta_1' + x_{0,2}\overline{\theta}_2$. We have $\bar{m}_{\theta,1}(\theta, x_{0,1}, P) = -f_{W_i^H|X_i}(\theta_1 + x_{0,1}'\theta_2|x_{0,1})(1, x_{0,1})$ and $\bar{m}_{\theta,2}(\theta, x_{0,2}, P) = f_{W_i^L|X_i}(\theta_1 + x_{0,2}'\theta_2|x_{0,1})(1, x_{0,1})$, so

$$\bar{m}_{\theta,1}(\theta, x_{0,1}, P)(\theta - (\theta_1', \overline{\theta}_2)) = -f_{W_i^H|X_i}(\theta_1 + x_{0,1}'\theta_2|x_{0,1})(1, x_{0,1})(\theta - (\theta_1', \overline{\theta}_2))$$

and

$$\bar{m}_{\theta,2}(\theta, x_{0,2}, P)(\theta - (\theta_1', \overline{\theta}_2)) = f_{W_i^L|X_i}(\theta_1 + x_{0,2}'\theta_2|x_{0,1})(1, x_{0,2})(\theta - (\theta_1', \overline{\theta}_2)).$$

Letting a_1 be the expression in the first display above, and a_2 the expression in the second display above, note that, if

$$[f_{W_i^H|X_i}(\theta_1 + x_{0,1}'\theta_2|x_{0,1})]^{-1} \cdot a_1 + [f_{W_i^L|X_i}(\theta_1 + x_{0,2}'\theta_2|x_{0,1})]^{-1} \cdot a_2 \leq -\frac{2\eta}{\underline{f}}\|\theta - (\theta_1', \overline{\theta}_2)\|,$$

then either $a_1 \leq -\eta\|\theta - (\theta_1', \overline{\theta}_2)\|$ or $a_2 \leq -\eta\|\theta - (\theta_1', \overline{\theta}_2)\|$. Thus, it suffices to bound the expression on the left hand side of the above display divided by $\|\theta - (\theta_1', \overline{\theta}_2)\|$ away from zero from above. The left hand side of the above display divided by $\|\theta - (\theta_1', \overline{\theta}_2)\|$ is equal to

$$[(1, x_{0,2}) - (1, x_{0,1})](\theta - (\theta_1', \overline{\theta}_2))/\|\theta - (\theta_1', \overline{\theta}_2)\| = -\frac{(x_{0,1} - x_{0,2})(\theta_2 - \overline{\theta}_2)}{[(\theta_1 - \theta_1')^2 + (\theta_2 - \theta_2')^2]^{1/2}}.$$

By the same argument as in the proof of Theorem 6.2, this is bounded away from zero from above for $|\theta_1 - \theta_1'|/|\theta_2 - \overline{\theta}_2|$ bounded away from infinity since $x_{0,1} - x_{0,2}$ is bounded away from zero, and, for $|\theta_1 - \theta_1'|/|\theta_2 - \overline{\theta}_2|$ large enough, either $\bar{m}_{\theta,1}(\theta, x_{0,1}, P)(\theta - (\theta_1', \overline{\theta}_2))$ or $\bar{m}_{\theta,2}(\theta, x_{0,2}, P)(\theta - (\theta_1', \overline{\theta}_2))$ will be less than the same negative constant for all $P \in \mathcal{P}$.

□

proof of Theorem B.3. The result follow immediately from Theorem 3.1. □

proof of Theorem B.4. For the case where Assumption B.7 holds, the result follows by verifying the conditions of Theorem 4.2 with g a function that is positive only on $[\underline{x}, \overline{x}]$. For the other cases, the result will follow by verifying the conditions of Theorem 4.3 once we show that these models can be transformed so that Assumption B.6 holds with ϕ_x in the transformed model equal to zero and, under Assumption B.6 on the original model, ϕ_m in the transformed model equal to $\phi_m/(\phi_x + 1)$ and, under Assumption B.5 (and $d_X = 1$) on the original model, ϕ_m in the transformed model equal to $\phi_m/(\phi_x - 1)$. (Assumption 4.6 is invariant to taking the same invertible monotonic transformation of each element of X_i, since we can replace $\|\cdot\|$ in that assumption with the supremum norm, and then the sets involved are d_X dimensional boxes, and the set of all d_X-dimensional boxes is invariant to such transformations. This holds even for the transformations used under Assumption B.5 in which infinity is taken to a finite support point by taking t in Assumption 4.6 to be large

enough so that the largest value of any component of X_i in the sample is contained in the d_X-dimensional box.)

Suppose that Assumption B.6 holds for some ϕ_m and ϕ_x. Then, for any $t \in \mathbb{R}$ with each element less than η_X

$$P(0 < x_{0,k} - X_{i,k} < t_k \text{ all } k) = P(x_0 - t < X_i < x_0)$$
$$\geq \frac{1}{C} \int_{x_{0,1}-t_1}^{x_0} \cdots \int_{x_{0,d_X}-t_{d_X}}^{x_0} \prod_{k=1}^{d} |x_{0,k} - x_k|^{\phi_x} \, dx_1 \cdots dx_{d_X} = \frac{1}{C} \prod_{k=1}^{d} \frac{t_k^{\phi_x+1}}{\phi_x + 1}$$

so that

$$P(x_{0,k} - t_k < x_{0,k} - (x_{0,k} - X_{i,k})^{(\phi_x+1)} < x_{0,k} \text{ all } k) = P(0 < x_{0,k} - X_{i,k} < t_k^{1/(\phi_x+1)} \text{ all } k)$$
$$\geq \frac{1}{C} \prod_{k=1}^{d} \frac{t_k}{\phi_x + 1}.$$

Thus, the random variable V_i defined to have kth element $x_{0,k} - (x_{0,k} - X_{i,k})^{(\phi_x+1)}$ for $x_0 - \eta_X < X_i < x_0$ and $X_{i,k}$ otherwise will satisfy part (ii) of Assumption B.6 (for a different value of η_X) with ϕ_x equal to zero for the transformed variable. To get the conditional mean of the transformed model, note that, for $x_0 - \eta_X < X_i < x_0$,

$$E_P(W_i^H | V_i = v) = E_P(W_i^H | x_{0,k} - (x_{0,k} - X_{i,k})^{(\phi_x+1)} = v_k \text{ all } k)$$
$$= E_P(W_i^H | x_{0,k} - X_{i,k} = (x_{0,k} - v_k)^{1/(\phi_x+1)} \text{ all } k) = E_P(W_i^H | X_{i,k} = x_{0,k} - (x_{0,k} - v_k)^{1/(\phi_x+1)} \text{ all } k)$$
$$\leq C \|((x_{0,1} - v_1)^{1/(\phi_x+1)}, \ldots, (x_{0,d_X} - v_{d_X})^{1/(\phi_x+1)})\|^{\phi_m} \leq C d_X^{\phi_m} \|x_0 - v\|^{\phi_m/(\phi_x+1)}.$$

Thus, Assumption B.6 will hold for the transformed model with X_i replaced with V_i and ϕ_m in the transformed model equal to $\phi_m/(\phi_x + 1)$ and ϕ_x in the transformed model equal to zero.

If Assumption B.5 holds for some ϕ_m and ϕ_x, then, for t greater than K_X (here $d_X = 1$),

$$P(X_i \geq t) \geq \frac{1}{C} \int_t^{\infty} x^{-\phi_x} \, dx = \frac{1}{C(\phi_x - 1)} t^{1-\phi_x}.$$

Thus,

$$
\begin{aligned}
&P(K_X + 1 - 1/(X_i - K_X + 1) \geq K_X + 1 - t) = P(-1/(X_i - K_X + 1) \geq -t) \\
&= P(1/(X_i - K_X + 1) \leq t) = P(1/t \leq X_i - K_X + 1) \\
&= P(K_X - 1 + 1/t \leq X_i) \geq \frac{1}{C(\phi_x - 1)}(K_X - 1 + 1/t)^{1-\phi_x} \geq \frac{2^{1-\phi_x}}{C(\phi_x - 1)} t^{\phi_x - 1}
\end{aligned}
$$

where the last inequality holds for t small enough so that $1/t \geq K_X - 1$. It follows that part (ii) of Assumption B.6 holds with ϕ_x in that assumption replaced by $\phi_x - 2$ for the transformed random variable V_i given by $V_i = K_X + 1 - 1/(X_i - K_X + 1)$ for $X_i > K_X$ and $V_i = X_i$ otherwise. Here, x_0 from Assumption B.6 is equal to $K_X + 1$ in the transformed model. As for the conditional mean of the transformed model, we have, for v close enough to $K_X + 1$,

$$
\begin{aligned}
&E_P(W_i^H | V_i = v) = E_P(W_i^H | K_X + 1 - 1/(X_i - K_X + 1) = v) \\
&= E_P(W_i^H | -1/(X_i - K_X + 1) = v - 1 - K_X) = E_P(W_i^H | X_i - K_X + 1 = -1/(v - 1 - K_X)) \\
&= E_P(W_i^H | X_i = -1/(v - 1 - K_X) + K_X - 1) \leq C(-1/(v - 1 - K_X) + K_X - 1)^{-\phi_m} \\
&\leq 2C(1/(1 + K_X - v))^{-\phi_m} = 2C(1 + K_X - v)^{\phi_m}
\end{aligned}
$$

so that part (i) of Assumption B.6 holds with the same ϕ_m.

□

proof of Theorem 5.1. Let θ_n be a sequence converging to θ_0 such that, for some $\varepsilon > 0$, $d_H(\theta_n, \Theta_0(P)) = n^{-\alpha/(2d_X + 2\alpha)}\varepsilon$, for large enough n, (conditions on how small ε is will be stated below). Such a sequence exists by part (iv) of Assumption 5.1. For each n, let $\theta_0'(n) \in \delta\Theta_0(P)$ be such that $d_H(\theta_n, \theta_0'(n)) \leq 2n^{-\alpha/(2d_X + 2\alpha)}\varepsilon$ (doubling the distance to the identified set covers the possibility that the infimum is not achieved). For each j, we have, for some $x_0 \in \mathcal{X}_0(\theta_0'(n))$ and some θ_n^* between θ_n and $\theta_0'(n)$,

$$
\begin{aligned}
&\bar{m}_j(\theta_n, x, P) = \bar{m}_j(\theta_n, x, P) - \bar{m}_j(\theta_0'(n), x_0, P) \\
&= [\bar{m}_j(\theta_n, x, P) - \bar{m}_j(\theta_0'(n), x, P)] + [\bar{m}_j(\theta_0'(n), x, P) - \bar{m}_j(\theta_0'(n), x_0, P)] \\
&= \bar{m}_{\theta,j}(\theta_n^*, x, P)(\theta_n - \theta_0'(n)) + [\bar{m}_j(\theta_0'(n), x, P) - \bar{m}_j(\theta_0'(n), x_0, P)] \\
&\geq -2Kn^{-\alpha/(2d_X + 2\alpha)}\varepsilon + \eta \min_{x_0 \in \mathcal{X}_0(\theta_0'(n))} (\|x - x_0\|^\alpha \wedge \eta) \qquad (10)
\end{aligned}
$$

where K is a bound on the derivative. For n large enough, the last line of the above display

is negative only for x such that, for some $x_0 \in \mathcal{X}_0(\theta_0'(n))$, $\|x - x_0\| < \left(\frac{2K\varepsilon}{\eta}\right)^{1/\alpha} n^{-1/(2d_X+2\alpha)}$. This will imply, letting $\overline{g}$ be an upper bound for functions in $\mathcal{G}$ and K_1 an upper bound for the number of elements in $\mathcal{X}_0(\theta_0'(n))$,

$$\begin{aligned}
&\mu_{P,j}(\theta_n, g) = E_P \bar{m}_j(\theta_n, X_i, P) g(X_i) \geq -2Kn^{-\alpha/(2d_X+2\alpha)}\varepsilon\overline{g} P(\bar{m}_j(\theta_n, X_i, P) < 0) \\
&\geq -2Kn^{-\alpha/(2d_X+2\alpha)}\varepsilon\overline{g} \sum_{x_0 \in \mathcal{X}_0(\theta_0'(n))} P\left(\|X_i - x_0\| < \left(\frac{2K\varepsilon}{\eta}\right)^{1/\alpha} n^{-1/(2d_X+2\alpha)}\right) \\
&\geq -2Kn^{-\alpha/(2d_X+2\alpha)}\varepsilon K_1 \overline{g}\overline{f} 2^{d_X} \left(\frac{2K\varepsilon}{\eta}\right)^{d_X/\alpha} n^{-d_X/(2d_X+2\alpha)} \\
&= -2K\varepsilon K_1 \overline{g}\overline{f} 2^{d_X} \left(\frac{2K\varepsilon}{\eta}\right)^{d_X/\alpha} n^{-1/2}.
\end{aligned}$$

Here, the first inequality follows for large enough n since $\bar{m}_j(\theta_n, x, P) \geq -2Kn^{-\alpha/(2d_X+2\alpha)}\varepsilon$ eventually by the argument above.

If $d_H(\mathcal{C}_{n,\omega}(\hat{c}_n), \Theta_0(P)) < \varepsilon n^{-\alpha/(2d_X+2\alpha)}$, then $\theta_n \notin \mathcal{C}_{n,\omega}(\hat{c}_n)$, so that $T_{n,\omega}(\theta_n) > \hat{c}_n n^{-1/2} \geq \underline{c} n^{-1/2}$ where $\underline{c}$ is a lower bound for $\hat{c}_n$. Then, for some j and g, we will have, letting $K_{S,1}$ be as in Assumption 3.3, $\omega_n(\theta_n, g)\hat{\mu}_{n,j}(\theta_n, g) \leq -K_{S,1}\underline{c} n^{-1/2}$ so that, letting $\overline{\omega}$ be an upper bound for $\omega_n(\theta, g)$, $n^{1/2}\hat{\mu}_{n,j}(\theta_n, g) \leq -K_{S,1}\underline{c}/\overline{\omega}$. For large enough n, we will also have $n^{1/2}\mu_{P,j}(\theta_n, g) \geq -2K\varepsilon K_1 \overline{g}\overline{f} 2^{d_X} \left(\frac{2K\varepsilon}{\eta}\right)^{d_X/\alpha}$. This will imply

$$n^{1/2}\left\{\hat{\mu}_{n,j}(\theta_n, g) - [\mu_{P,j}(\theta_n, g) \wedge 0]\right\} \leq -K_{S,1}\underline{c}/\overline{\omega} + 2K\varepsilon K_1 \overline{g}\overline{f} 2^{d_X} \left(\frac{2K\varepsilon}{\eta}\right)^{d_X/\alpha}$$

so that $n^{1/2}\left\{\hat{\mu}_{n,j}(\theta_n, g) - [\mu_{P,j}(\theta_n, g) \wedge 0]\right\}$ is bounded away from zero from above by a negative constant when this event holds for small enough ε. Thus, it suffices to show that, for any $\delta > 0$, $n^{1/2} \inf_{g \in \mathcal{G}} \left\{\hat{\mu}_{n,j}(\theta_n, g) - [\mu_{P,j}(\theta_n, g) \wedge 0]\right\} > -\delta$ with probability approaching one.

We have, for any $r > 0$,

$$\begin{aligned}
&n^{1/2} \inf_{g \in \mathcal{G}} \left\{\hat{\mu}_{n,j}(\theta_n, g) - [\mu_{P,j}(\theta_n, g) \wedge 0]\right\} \\
&\geq n^{1/2} \inf_{g \in \mathcal{G}} \hat{\mu}_{n,j}(\theta_n, g) I(\mu_{P,j}(\theta_n, g) > r) + n^{1/2} \inf_{g \in \mathcal{G}} \left\{\hat{\mu}_{n,j}(\theta_n, g) - \mu_{P,j}(\theta_n, g)\right\} I(\mu_{P,j}(\theta_n, g) \leq r).
\end{aligned}$$

The first term is greater than zero with probability approaching one since $\hat{\mu}_{n,j}(\theta, g)$ converges to $\mu_{P,j}(\theta, g)$ at a root-n rate uniformly over (θ, g) by standard arguments (e.g. Theorem 2.5.2 in van der Vaart and Wellner (1996)).

As for the second term, note that, for any $\delta_1, \delta_2 > 0$ with $\delta_1^\alpha \leq \eta$, $\bar{m}_j(\theta_n, x, P)$ will be greater than or equal to $-\delta_2 I(d_H(x, \mathcal{X}_0(\theta_0'(n))) < \delta_1) + (\eta\delta_1^\alpha - \delta_2) I(d_H(x, \mathcal{X}_0(\theta_0'(n))) \geq \delta_1)$ for large enough n by (10). To simplify notation, define the sets $A_{n,\delta_1} = \{x | d_H(x, \mathcal{X}_0(\theta_0'(n))) < \delta_1\}$. Using this notation, the above observation implies that, for n greater than some constant that depends on δ_1,

$$\mu_{P,j}(\theta_n, g) = E_P \bar{m}_j(\theta_n, X_i, P) g(X_i) \geq -\delta_2 E_P g(X_i) I(X_i \in A_{n,\delta_1}) + (\eta\delta_1^\alpha - \delta_2) E_P g(X_i) I(X_i \notin A_{n,\delta_1}).$$

If $\mu_{P,j}(\theta_n, g) \leq r$, then this means that

$$(\eta\delta_1^\alpha - \delta_2) E_P g(X_i) I(X_i \notin A_{n,\delta_1}) \leq \delta_2 E_P g(X_i) I(X_i \in A_{n,\delta_1}) + r$$

where, as above, K_1 is an upper bound for the number of elements in $\mathcal{X}_0(\theta_0'(n))$. Thus, for $\mu_{P,j}(\theta_n, g) \leq r$, and n larger than some constant that depends only on δ_1, letting $\overline{g}$ be a bound for $g(X_i)$ and M a bound for $m_j(W_i, \theta)$,

$$\begin{aligned}
&E_P[m_j(W_i, \theta_n) g(X_i)]^2 \leq \overline{g} M^2 E_P g(X_i) = \overline{g} M^2 [E_P g(X_i) I(X_i \notin A_{n,\delta_1}) + E_P g(X_i) I(X_i \in A_{n,\delta_1})] \\
&\leq \overline{g} M^2 \{[\delta_2 E_P g(X_i) I(X_i \in A_{n,\delta_1}) + r] / (\eta\delta_1^\alpha - \delta_2) + E_P g(X_i) I(X_i \in A_{n,\delta_1})\} \\
&= \overline{g} M^2 \left[\left(\frac{\delta_2}{\eta\delta_1^\alpha - \delta_2} + 1 \right) E_P g(X_i) I(X_i \in A_{n,\delta_1}) + \frac{r}{\eta\delta_1^\alpha - \delta_2} \right] \\
&\leq \overline{g} M^2 \left[\left(\frac{\delta_2}{\eta\delta_1^\alpha - \delta_2} + 1 \right) \overline{g} K_1 (2\delta_1)^{d_X} + \frac{r}{\eta\delta_1^\alpha - \delta_2} \right]
\end{aligned}$$

By choosing r, δ_1, and δ_2 so that δ_1, $r/(\eta\delta_1^\alpha - \delta_2)$ and $\delta_2/(\eta\delta_1^\alpha - \delta_2)$ are small, we can make the last line of the display less than any $\delta_3 > 0$. Then, for n large enough, $\mu_{P,j}(\theta_n, g) \leq r$ will imply $var_P[m(W_i, \theta_n) g(X_i)] \leq \delta_3$, so that

$$\begin{aligned}
&n^{1/2} \inf_{g \in \mathcal{G}} \{\hat{\mu}_{n,j}(\theta_n, g) - \mu_{P,j}(\theta_n, g)\} I(\mu_{P,j}(\theta_n, g) \leq r). \\
&\geq n^{1/2} \inf_{g \in \mathcal{G}} \{\hat{\mu}_{n,j}(\theta_n, g) - \mu_{P,j}(\theta_n, g)\} I(var_P[m(W_i, \theta_n) g(X_i)] \leq \delta_3).
\end{aligned}$$

This can be made arbitrarily small in magnitude by the stochastic asymptotic equicontinuity of $n^{1/2}(E_n - E_P) m(W_i, \theta) g(X_i)$ with respect to the covariance semimetric $\rho((\theta, g), (\theta', g')) = var_P[m(W_i, \theta) g(X_i) - m(W_i, \theta') g'(X_i)]$ as a sequence of processes indexed by (θ, g). Letting $\tilde{g}(x) = 0$ be the zero function and $\tilde{\theta}$ an arbitrary value in Θ, the last line of the above display

is equal to

$$n^{1/2} \inf_{g \in \mathcal{G}} \left\{ (E_n - E) m_j(W_i, \theta_n) g_j(X_i) - (E_n - E) m_j(W_i, \tilde{\theta}) \tilde{g}_j(X_i) \right\} I(\rho((\theta_n, g), (\tilde{\theta}, \tilde{g})) \leq \delta_3).$$

By making δ_3 small, the probability of this being less than any negative constant can be made arbitrarily small by equicontinuity of $n^{1/2}(E_n - E) m_j(W_i, \theta_n) g_j(X_i)$ in ρ.

□

proof of Theorem 5.2. By the same argument that gives (9) in the proof of Theorem 4.3, we will have, for θ with $d_H(\theta, \Theta_0(P))$ smaller than some constant that does not depend on P, there exists a $\theta_0 \in \Theta_0(P)$, j_0 and $g \in \mathcal{G}$ with $g(x) \geq C_{\mathcal{G},1} I(\|x - x_0\| \leq [\eta/(4C)]^{1/\alpha} \|\theta - \theta_0\|^{1/\alpha} C_{\mathcal{G},2})$ such that

$$\mu_{P,j_0}(\theta, g) = E_P \bar{m}_{j_0}(\theta, X_i, P) g(X_i) \leq -(\eta/4) \|\theta - \theta_0\| E_P g(X_i).$$

This, and the lower bound on g gives

$$\begin{aligned}
&\mu_{P,j_0}(\theta, g) \leq -(\eta/4) \|\theta - \theta_0\| C_{\mathcal{G},1} \left\{ [\eta/(4C)]^{1/\alpha} \|\theta - \theta_0\|^{1/\alpha} C_{\mathcal{G},2} \right\}^{d_X} \eta \\
&= -(\eta/4) \|\theta - \theta_0\|^{(\alpha + d_X)/\alpha} C_{\mathcal{G},1} \left\{ [\eta/(4C)]^{1/\alpha} C_{\mathcal{G},2} \right\}^{d_X} \eta \\
&\leq -(\eta/4) d_H(\theta, \Theta_0(P))^{(\alpha + d_X)/\alpha} C_{\mathcal{G},1} \left\{ [\eta/(4C)]^{1/\alpha} C_{\mathcal{G},2} \right\}^{d_X} \eta.
\end{aligned}$$

Thus, the conditions of Lemma A.6 hold with $\gamma = \alpha/(d_X + \alpha)$.

□

The proof of Theorem 5.2 uses the following lemma, which is analogous to Theorem 4.2 for set estimates based on variance weighted KS statistics.

Lemma A.6. *Suppose that, for some positive constants C, γ, and δ, we have, for all $P \in \mathcal{P}$ and θ with $d_H(\theta, \Theta_0(P)) < \delta$,*

$$\inf_{g,j} \mu_{P,j}(\theta, g) \leq -C d_H(\theta, \Theta_0(P))^{1/\gamma}$$

where the infimum is taken over $g \in \mathcal{G}$ and $j \in \{1, \ldots, d_Y\}$. Suppose that Assumptions 3.1, 3.2, 3.3, and 4.1 hold, and that the weight function $\omega_n(\theta, g)$ satisfies $\underline{\omega} \leq \omega_n(\theta, g) \leq \overline{\omega}$ for some $0 < \underline{\omega} \leq \overline{\omega} < \infty$, and suppose that $\hat{c}_n \to \infty$ with $\hat{c}_n / \sqrt{n} \to 0$. Then,

$$\inf_{P \in \mathcal{P}} P(\Theta_0(P) \subseteq \mathcal{C}_{n,\omega}(\hat{c}_n)) \overset{n \to \infty}{\to} 1$$

and, for some large B,

$$\sup_{P\in\mathcal{P}} P\left(\left(n/\hat{c}_n^2\right)^{\gamma/2} d_H(\mathcal{C}_n(\hat{c}_n),\Theta_0(P)) > B\right) \overset{n\to\infty}{\to} 0.$$

Proof. First, note that, for all j, $\sup_{\theta,g}\sqrt{n}|(E_n - E)m_j(W_i,\theta)g_j(X_i)| = \mathcal{O}_P(1)$ uniformly in P by Theorem 2.14.1 in van der Vaart and Wellner (1996) (the constant function equal to $\overline{Y}$ does not depend on P and can be used as an envelope function). This, along with Assumption 3.3 and the bound on the weight function, implies the first claim.

For the second claim, once $\Theta_0(P) \subseteq \mathcal{C}_{n,\omega}(\hat{c}_n)$, if $\left(n/\hat{c}_n^2\right)^{\gamma/2} d_H(\mathcal{C}_n(\hat{c}_n),\Theta_0(P)) > B$, there will be a $\theta \in \mathcal{C}_{n,\omega}(\hat{c}_n)$ such that $d_H(\theta,\Theta_0(P)) > B\frac{\hat{c}_n^\gamma}{n^{\gamma/2}}$. If $d_H(\mathcal{C}_{n,\omega}(\hat{c}_n),\Theta_0(P)) < \delta$, which happens with probability approaching one uniformly in $P \in \mathcal{P}$ by arguments similar to the proof of Theorem 4.1, then, for this θ and P, there will be a g^* and j^* such that, for n greater than some constant that does not depend on P, $\mu_{P,j^*}(\theta,g^*) \leq -(C/2)\left(\hat{c}_n^2/n\right)^{1/2} B^{1/\gamma}$. Since $\theta \in \mathcal{C}_{n,\omega}(\hat{c}_n)$, we will also have $T_{n,\omega}(\theta) \leq \hat{c}_n n^{-1/2}$, so that $\hat{\mu}_{n,j^*}(\theta,g^*)\omega_{n,j^*}(\theta,g^*) \geq -\hat{c}_n n^{-1/2} K_{S,2}$. By the lower bound on the weight function, this implies $\hat{\mu}_{n,j^*}(\theta,g^*) \geq -\hat{c}_n n^{-1/2} K_{S,2}/\underline{\omega}$. Thus,

$$\sqrt{n}[\hat{\mu}_{n,j^*}(\theta,g^*) - \mu_{P,j^*}(\theta,g^*)] \geq \hat{c}_n\left[-K_{S,2}/\underline{\omega} + (C/2)B^{1/\gamma}\right].$$

For B large enough, the right hand side will go to infinity. Since the left hand side is $\mathcal{O}_P(1)$ uniformly in $P \in \mathcal{P}$, this gives the desired result.

□

proof of Theorem 5.3. Let θ_n and $\theta_0'(n)$ be as in the proof of Theorem 5.1, but with $d_H(\theta_n,\Theta_0(P)) = \varepsilon\left(\frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}} \vee h_n^\alpha\right)$.

If $d_H(\mathcal{C}_n^{\text{kern}}(\hat{c}_n),\Theta_0(P)) < \varepsilon\left(\frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}} \vee h_n^\alpha\right)$, then $\theta_n \notin \mathcal{C}_n^{\text{kern}}(\hat{c}_n)$ so that $T_{n,k,h_n}^{\text{kern}}(\theta_n) \geq \hat{c}_n$. Then, letting $K_{S,1}$ be as in Assumption 3.3, we will have, for some j and x, $\frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}}\hat{\bar{m}}_j(x,\theta_n) \leq -K_{S,1}\hat{c}_n$. By Lemmas A.7 and A.8, for large enough a we will have, for some constant K,

$$\sup_{x\in\mathbb{R}^{d_X},\theta\in\Theta} \frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}} \left|\frac{(E_n - E_P)m_j(W_i,\theta)k((X_i-x)/h_n)}{E_n k((X_i-x)/h_n)}\right| \leq K \tag{11}$$

with probability approaching one (Lemma A.7 allows $E_P k((X_i-x)/h_n)$ to be replaced by

its sample analogue in Lemma A.8). When $T^{\text{kern}}_{n,k,h_n}(\theta_n) \geq \hat{c}_n$, we will have

$$\frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}}\left[\frac{(E_n - E_P)m_j(W_i,\theta_n)k((X_i-x)/h_n)}{E_n k((X_i-x)/h_n)} + \frac{E_P m_j(W_i,\theta_n)k((X_i-x)/h_n)}{E_n k((X_i-x)/h_n)}\right]$$
$$= \frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}}\hat{\bar{m}}_j(x,\theta_n) \leq -K_{S,1}\hat{c}_n,$$

so that, when (11) holds, we will have

$$\frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}}\frac{E_P m_j(W_i,\theta_n)k((X_i-x)/h_n)}{E_n k((X_i-x)/h_n)} \leq -K_{S,1}\hat{c}_n + K.$$

Appealing again to Lemma A.7, if a is large enough, this will imply

$$\frac{\sqrt{nh_n^{d_X}}}{\sqrt{\log n}}\frac{E_P m_j(W_i,\theta_n)k((X_i-x)/h_n)}{E_P k((X_i-x)/h_n)} \leq \frac{-K_{S,1}\hat{c}_n + K}{2}.$$

Letting η be as in Assumption 4.5 letting $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ be such that $k(t) \geq \varepsilon_1$ for $\|t\| \leq \varepsilon_2$ and defining $K_1 = \eta\varepsilon_1\varepsilon_2^{d_X}$, we have $E_P k((X_i-x)/h_n) \geq \varepsilon_1 P(\|X_i - x\| \leq h_n\varepsilon_2) \geq \eta\varepsilon_1\varepsilon_2^{d_x}h_n^{d_X} = K_1 h_n^{d_X}$ by Assumption 4.5, so that the above display implies

$$E_P m_j(W_i,\theta_n)k((X_i-x)/h_n) \leq K_1 h_n^{d_X}\frac{-K_{S,1}\hat{c}_n + K}{2}\frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}} = \frac{K_1\left(-K_{S,1}\hat{c}_n + K\right)}{2}\frac{\sqrt{h_n^{d_X}}\sqrt{\log n}}{\sqrt{n}}.$$

Let $\hat{c}_n$ be large enough so that $K_1\left(-K_{S,1}\hat{c}_n + K\right)/2 \leq -\delta$ for some fixed constant $\delta > 0$. Then the above display implies

$$E_P m_j(W_i,\theta_n)k((X_i-x)/h_n) \leq -\delta\frac{\sqrt{h_n^{d_X}}\sqrt{\log n}}{\sqrt{n}}. \tag{12}$$

When this holds, the right hand side will be negative, so that, by Lemma A.9, $h_n \leq B[d_H(\theta_n,\Theta_0(P))]^{1/\alpha}$. If $h_n^\alpha \geq \frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}}$, this will imply $h_n \leq \varepsilon^{1/\alpha}Bh_n$, which is a contradiction for ε small enough.

Now suppose $h_n^\alpha \leq \frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}}$. By the same argument as in the proof of Theorem 5.1, we have, for some constant K_2 that does not depend on n, $\bar{m}_j(\theta_n,x) \geq -K_2 d_H(\theta_n,\Theta_0(P))$ so that, if $h_n^\alpha \leq \frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}}$, $\bar{m}_j(\theta_n,x) \geq -\varepsilon K_2\frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}}$ so that the left hand side of (12) is greater

than or equal to

$$-\varepsilon K_2 \frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}} E_P k((X_i - x)/h_n) \geq -\varepsilon K_2 \frac{\sqrt{\log n}}{\sqrt{nh_n^{d_X}}} \overline{f} h_n^{d_X} = -\varepsilon \overline{f} K_2 \frac{\sqrt{h_n^{d_X}} \sqrt{\log n}}{\sqrt{n}}$$

so that (12) implies $\varepsilon \overline{f} K_2 \geq \delta$, a contradiction for ε small enough.

□

The proof of Theorem 5.3 uses the lemmas stated and proved below.

Lemma A.7. *Suppose that Assumption 5.2 holds, and that Assumption 4.5 and part (iii) of Assumption 5.1 hold, with the upper bound on the density in the latter assumption uniform in $P \in \mathcal{P}$. Then, for any ε, there exists an a such that, if $h_n^{d_X} n / \log n \geq a$ eventually,*

$$\sup_{P \in \mathcal{P}} P \left(\sup_{x \in supp_P(X_i)} \left| \frac{E_n k((X_i - x)/h)}{E_P k((X_i - x)/h)} - 1 \right| > \varepsilon \right) \overset{n \to \infty}{\to} 0$$

for all $\varepsilon > 0$.

Proof. We have

$$\begin{aligned} &\left| \frac{E_n k((X_i - x)/h_n)}{E_P k((X_i - x)/h_n)} - 1 \right| = \frac{\{E_P[k((X_i - x)/h_n)]^2\}^{1/2}}{E_P k((X_i - x)/h_n)} \left| \frac{(E_n - E_P)k((X_i - x)/h_n)}{\{E_P[k((X_i - x)/h_n)]^2\}^{1/2}} \right| \\ &\leq \overline{k}^{1/2} \left| \frac{(E_n - E_P)k((X_i - x)/h_n)}{\{E_P[k((X_i - x)/h_n)]^2\}^{1/2}} \right| \cdot \frac{1}{[E_P k((X_i - x)/h_n)]^{1/2}} \end{aligned}$$

where $\overline{k}$ is an upper bound for the kernel function k. By Theorem A.1,

$$\sup_{P \in \mathcal{P}} P \left(\sup_{x \in \text{supp}_P(X_i)} \frac{\sqrt{n}}{\sqrt{\log n}} \left| \frac{(E_n - E_P)k((X_i - x)/h_n)}{\{E_P[k((X_i - x)/h_n)]^2\}^{1/2}} \right| > K \right) \to 0$$

for large enough K (the lower bound on the denominator follows from Assumption 4.5), so the result will follow if we can show that $[E_P k((X_i - x)/h_n)]^{1/2} \sqrt{n} / \sqrt{\log n}$ can be made arbitrarily large by choosing a large in the assumptions of the lemma. By Assumptions 5.2 and 4.5, we have, for some $\delta > 0$ and all x on the support of X_i under P,

$$[n/(\log n)] E_P k((X_i - x)/h_n) \geq [n/(\log n)] \delta h_n^{d_X},$$

and taking the square root of this expression gives something that can be made arbitrarily large by choosing a large. □

Lemma A.8. *Suppose that Assumption 5.2 holds, and that Assumption 4.5 and part (iii) of Assumption 5.1 hold, with the upper bound on the density in the latter assumption uniform in $P \in \mathcal{P}$. Then, if $h_n^{d_X} n / \log n \geq a$ eventually for a large enough, we will have*

$$\sup_{P \in \mathcal{P}} P\left(\sup_{x \in supp_P(X_i), \theta \in \Theta} \frac{\sqrt{n h_n^{d_X}}}{\sqrt{\log n}} \left| \frac{(E_n - E_P) m_j(W_i, \theta) k((X_i - x)/h_n)}{E_P k((X_i - x)/h_n)} \right| > B \right) \overset{n \to \infty}{\to} 0$$

for some B.

Proof. We have

$$\begin{aligned}
&\frac{\sqrt{n h_n^{d_X}}}{\sqrt{\log n}} \left| \frac{(E_n - E_P) m_j(W_i, \theta) k((X_i - x)/h_n)}{E_P k((X_i - x)/h_n)} \right| \\
&= \frac{\sqrt{n}}{\sqrt{\log n}} \left| \frac{(E_n - E_P) m_j(W_i, \theta) k((X_i - x)/h_n)}{\sqrt{var_P[m_j(W_i, \theta) k((X_i - x)/h_n)] \vee \sqrt{h_n^{d_X}}}} \right| \\
&\cdot \frac{\sqrt{h_n^{d_X}} \left\{ \sqrt{var_P[m_j(W_i, \theta) k((X_i - x)/h_n)] \vee \sqrt{h_n^{d_X}}} \right\}}{E_P k((X_i - x)/h_n)}
\end{aligned}$$

Since

$$\begin{aligned}
&var_P[m_j(W_i, \theta) k((X_i - x)/h_n)] \leq \overline{Y} E_P[k((X_i - x)/h_n)]^2 \leq \overline{Yf} \int_{t \in \mathbb{R}^{d_X}} [k((t - x)/h_n)]^2 \, dt \\
&= h_n^{d_X} \int_{u \in \mathbb{R}^{d_X}} [k(u)]^2 \, du,
\end{aligned}$$

the last line is bounded by a constant times

$$\frac{\sqrt{n}}{\sqrt{\log n}} \left| \frac{(E_n - E_P) m_j(W_i, \theta) k((X_i - x)/h_n)}{\sqrt{var_P[m_j(W_i, \theta) k((X_i - x)/h_n)] \vee \sqrt{h_n^{d_X}}}} \right| \cdot \frac{h_n^{d_X}}{E_P k((X_i - x)/h_n)}.$$

By Assumptions 5.2 and 4.5, we have, for some $\delta > 0$ and x on the support of X_i under P, $E_P k((X_i - x)/h_n) \geq \delta h_n^{d_X}$, so that this is bounded by

$$\frac{\sqrt{n}}{\sqrt{\log n}} \left| \frac{(E_n - E_P) m_j(W_i, \theta) k((X_i - x)/h_n)}{\sqrt{var_P[m_j(W_i, \theta) k((X_i - x)/h_n)] \vee \sqrt{h_n^{d_X}}}} \right| \cdot (1/\delta).$$

The claim now follows from Theorem A.1, with $\sqrt{h_n^{d_X}}$ playing the role of the cutoff point σ_n.

□

Lemma A.9. *Suppose that Assumptions 4.5, 5.1 and 5.2 hold. Let θ_0 be as in Assumption 5.1 and let θ_n be a sequence in $\Theta \backslash \Theta_0(P)$ converging to θ_0. Then, for some constant B that does not depend on n and some $N \in \mathbb{N}$, $E_P m_j(W_i, \theta_n) k((X_i - x)/h)$ will be nonnegative for $h_n \geq B[d_H(\theta_n, \Theta_0(P))]^{1/\alpha}$ and $n \geq N$ for x on the support of X_i.*

Proof. Let $b_n = d_H(\theta_n, \Theta_0(P))$. By an argument similar to the one leading up to Equation (10), we will have, for each j,

$$\bar{m}_j(\theta_n, x, P) \geq -Cb_n + \eta \min_{x_0 \in \mathcal{X}_0(\theta_0'(n))} (\|x - x_0\|^\alpha \wedge \eta)$$

for some C that depends only on the bound on the derivative $\bar{m}_{\theta,j}(\theta, x, P)$ in Assumption 5.1 and some $\theta_0'(n) \in \Theta_0(P)$. Thus, for x such that $\bar{m}_j(\theta_n, x, P) \leq Cb_n$, we will have, for some $x_0 \in \mathcal{X}_0(\theta_0'(n))$, $Cb_n \geq -Cb_n + \eta(\|x - x_0\|^\alpha \wedge \eta)$ so that $2Cb_n \geq \eta(\|x - x_0\|^\alpha \wedge \eta)$. For b_n small enough, this implies that $\|x - x_0\| \leq (2Cb_n/\eta)^{1/\alpha}$. This means that, letting K be a bound for the number of elements in $\mathcal{X}_0(\theta_0'(n))$ and $\overline{f}$ an upper bound for the density of X_i,

$$P\left(\bar{m}_j(\theta_n, X_i, P) \leq Cb_n\right) \leq K\overline{f}(2Cb_n/\eta)^{d_X/\alpha}. \tag{13}$$

This, and the lower bound on $\bar{m}_j(\theta_n, x, P)$ imply, letting $\overline{k}$ be an upper bound on the kernel k,

$$\begin{aligned} &E_P m_j(W_i, \theta_n) k((X_i - x)/h_n) I(\bar{m}_j(\theta_n, X_i, P) \leq Cb_n) \geq -\overline{k} Cb_n P\left(\bar{m}_{\theta,j}(\theta, X_i, P) \leq Cb_n\right) \\ &\geq -\overline{k} Cb_n \cdot K\overline{f}(2Cb_n/\eta)^{d_X/\alpha}. \end{aligned}$$

We also have, for x on the support of X_i, letting ε and K_1 be such that $k(t) \geq K_1$ for $\|t\| \leq \varepsilon$,

$$\begin{aligned} &E_P m_j(W_i, \theta_n) k((X_i - x)/h_n) I(\bar{m}_j(\theta_n, X_i, P) > Cb_n) \\ &\geq Cb_n E_P k((X_i - x)/h_n) I(\bar{m}_j(\theta_n, X_i, P) > Cb_n) \\ &\geq K_1 Cb_n E_P I(\|(X_i - x)/h_n\| \leq \varepsilon) I(\bar{m}_j(\theta_n, X_i, P) > Cb_n) \\ &\geq K_1 Cb_n [P(\|(X_i - x)/h_n\| \leq \varepsilon) - P(\bar{m}_j(\theta_n, X_i, P) \leq Cb_n)] \\ &\geq K_1 Cb_n [\eta \varepsilon^{d_X} h_n^{d_X} - K\overline{f}(2Cb_n/\eta)^{d_X/\alpha}]. \end{aligned}$$

The last inequality follows from Assumption 4.5 and from the inequality (13) above (here the

two ηs come from different conditions, but they can be chosen to be the same by decreasing one). Combining this with the bound in the previous display gives

$$E_P m_j(W_i, \theta_n) k((X_i - x)/h_n) \geq K_1 C b_n [\eta \varepsilon^{d_X} h_n^{d_X} - K \overline{f} (2Cb_n/\eta)^{d_X/\alpha}] - \overline{k} C b_n \cdot K \overline{f} (2Cb_n/\eta)^{d_X/\alpha}$$
$$= b_n (K_2 h_n^{d_X} - K_3 b_n^{d_X/\alpha})$$

where $K_2 = K_1 C \eta \varepsilon^{d_X}$ and $K_3 = K_1 C K \overline{f} (2C/\eta)^{d_X/\alpha} + \overline{k} C K \overline{f} (2C/\eta)^{d_X/\alpha}$ are both positive constants that do not depend on n. For $h_n \geq (K_3/K_2)^{1/d_X} b_n^{1/\alpha}$, this will be nonnegative.

□

B Additional Applications

This section of the appendix derives rates of convergence to the identified set for several applications not considered in Section 6 by verifying the conditions of this paper. Section B.1 considers a one sided quantile regression model. Section B.2 considers an interval quantile regression model. Section B.3 considers bounds in a selection model. Proofs of the theorems in this section are given in Section A.6.

B.1 One Sided Quantile Regression

In this and the next section, I treat quantile versions of the regression models considered above. Here, we have a model for a conditional quantile of the unobserved variable W_i^* rather than the mean. The results are essentially the same, but, in addition to smoothness conditions on the quantile itself, conditions are needed on the joint density of the observed variables near the conditional quantile to translate these into the conditions on $\bar{m}(\theta, x, P)$.

First, consider the one sided case in which we observe (X_i, W_i^H) with $W_i^H \geq W_i^*$. For a random variable Z_i, define $q_{\tau,P}(Z_i|X_i)$ to be the τth quantile of Z_i conditional on X_i under P. Suppose that, for some known τ, the conditional τth quantile of W_i^* satisfies $q_{\tau,P}(W_i^*|X_i) = \theta_1 + X_i'\theta_{-1}$ for some θ. Then $E_P[\tau - I(W_i^* \leq \theta_1 + X_i'\theta_{-1})|X_i] = 0$ so that $E_P[\tau - I(W_i^H \leq \theta_1 + X_i'\theta_{-1})|X_i] \geq 0$. Thus, this fits into the framework of this paper with $W_i = (X_i, W_i^H)$ and $m(W_i, \theta) = \tau - I(W_i^H \leq \theta_1 + X_i'\theta_{-1})$.

In many situations, models for quantiles of an outcome variable given covariates can be more informative under interval data than models for the conditional mean. If W_H can be infinite with positive probability conditional on any value of X_i, the identified set for a conditional mean model will be the entire parameter space. If W_H has a low probability of

being large or infinite, and is usually close to W_i^*, a model for conditional quantiles of the unobserved variable will still give informative bounds with interval data.

Smoothness conditions that lead to Assumptions 4.3 and 4.4 for the quantile model are similar to those for the conditional mean considered above, but with smoothness assumptions placed on the conditional quantile $q_{\tau,P}(W_i^H|X_i)$ rather than the conditional mean, and additional assumptions on the joint density of (X_i, W_i^H). The first two assumptions are exactly the same as Assumptions 6.1 and 6.2, but with the conditional mean replaced by the conditional τth quantile.

Assumption B.1. *For some $C > 0$ and $\alpha \leq 1$, $\|q_{\tau,P}(W_i^H|X_i = x) - q_{\tau,P}(W_i^H|X_i = x')\| \leq C\|x - x'\|^\alpha$ for x and x' on the support of X_i for all $P \in \mathcal{P}$.*

Assumption B.2. *(i) $q_{\tau,P}(W_i^H|X_i = x)$ has a second derivative that is bounded uniformly in P and x and (ii) for any $P \in \mathcal{P}$, $\theta_0 \in \Theta_0(P)$, $q_{\tau,P}(W_i^H|X_i = x)$ is bounded away from $\theta_{0,1} + x'\theta_{0,-1}$ on the boundary of the support of X_i.*

The next assumption states that W_i^H has a density near its τth quantile conditional on X_i. One type of interval data that will frequently lead to this assumption holding is if (X_i, W_i^*) has a well behaved joint density, and W_i^H is equal to W_i^* with high probability and much larger than W_i^* with some small probability. For example, suppose that (X_i, W_i^*) has a joint density, and, W_i^H is either equal to ∞ or W_i^*, with $P(W_i^H = \infty|X_i = x, W_i^* = w)$ a smooth function of (x, w) that is bounded from above by some constant strictly less than $1 - \tau$. Then (X_i, W_i^H) will have a joint density near the τth conditional quantile of W_i^H. This type of situation arises naturally with missing data on an outcome variable. However, other types of interval data will not lead to this assumption holding. If W_i^H is the upper end of an interval from a survey in which W_i^* is always reported in the same interval, W_i^H will not have a density conditional on X_i.

Assumption B.3. *For some $\eta > 0$, $W_i^H|X_i$ has a conditional density $f_{W_i^H|X_i}(w|x)$ on $\{(x, w)|q_{\tau,P}(W_i^H|X_i = x) - \eta \leq w \leq q_{\tau,P}(W_i^H|X_i = x) + \eta\}$ that is continuous as a function of w uniformly in (w, x, P) and satisfies $\underline{f} \leq f_{W_i^H|X_i}(w|x) \leq \overline{f}$ for some $0 < \underline{f} < \overline{f} < \infty$.*

Under these conditions, Assumptions 4.3 and 4.4 will hold for the one sided quantile regression model. The proof is similar to the proof of Theorem 6.1 in the one sided regression model. The only difference is that some additional steps are needed to translate smoothness conditions on the τth quantile into smoothness conditions on the objective function using the assumptions on the conditional density of W_i^H given X_i.

Theorem B.1. *Suppose that the support of X_i is bounded uniformly in $P \in \mathcal{P}$, and that Assumptions 6.3 and B.3 hold in the one sided quantile regression model, with $E_P(W_i^H|X_i = x)$ replaced by $q_P(W_i^H|X_i = x)$ in assumption 6.3. Then, if Assumption B.1 holds, Assumptions 4.3 and 4.4 will hold for α specified in Assumption B.1. If Assumption B.2 holds, Assumptions 4.3 and 4.4 will hold for $\alpha = 2$.*

B.2 Interval Quantile Regression with a Scalar Regressor

Now consider a quantile regression model with two sided interval data in which, in addition to W_i^H, we observe a variable W_i^L that is known to satisfy $W_i^L \leq W_i^*$. This leads to $E_P[I(W_i^L \leq \theta_1 + X_i'\theta_{-1}) - \tau|X_i] \geq E_P[I(W_i^* \leq \theta_1 + X_i'\theta_{-1}) - \tau|X_i] = 0$ so that the interval quantile regression fits into the conditional moment inequality framework with $W_i = (X_i, W_i^L, W_i^H)$ and $m(W_i, \theta) = (\tau - I(W_i^H \leq \theta_1 + X_i'\theta_{-1}), I(W_i^L \leq \theta_1 + X_i'\theta_{-1}) - \tau)$.

As with the case of mean regression, the condition on the angle of the derivative and path in Assumption 4.4 will not hold in general in the quantile regression model with two sided interval data because of cases where alternatives are closest to a point in the identified set where the regression line is rotated around a contact point. Sufficient conditions to rule this out in the case of a scalar regressor are similar as well. Bounding the conditional quantiles of the upper and lower endpoints of the interval away from each other rules out these cases when the regressors include only a constant and a scalar. The next assumption is the same as Assumption 6.4, but with conditional expectations replaced by conditional τth quantiles. In the following, $x_{0,1}^{u,q}(P)$ is defined in the same way as $x_{0,1}^{u}(P)$ in Assumption 6.4, but with $E_P(\cdot|X_i = x)$ replaced by $q_{\tau,P}(\cdot|X_i = x)$, and similarly for $x_{0,2}^{u,q}(P)$, $x_{0,1}^{\ell,q}(P)$ and $x_{0,2}^{\ell,q}(P)$.

Assumption B.4. *(i) The support of X_i is bounded uniformly in $P \in \mathcal{P}$. (ii) The absolute value of the slope parameter θ_2 is bounded uniformly on the identified sets $\Theta_0(P)$ of $P \in \mathcal{P}$. (iii) $x_{0,1}^{u,q}(P) - x_{0,2}^{u,q}(P)$ and $x_{0,1}^{\ell,q}(P) - x_{0,2}^{\ell,q}(P)$ are bounded from below away from zero uniformly over $P \in \mathcal{P}$.*

The next theorem states that KS statistic based set estimators will have the same rate of convergence as in the one sided model with a scalar regressor under these conditions, and the assumption stated earlier on the density of the observed variables. The proof is similar to the proof of the analogous result for mean regression, Theorem 6.2, but with additional steps to translate conditions on quantiles and densities into conditions on the conditional mean of the objective function.

Theorem B.2. *In the interval regression example with $d_X = 1$, suppose that Assumptions B.3 and B.4 hold, and that Assumption B.3 also holds with W_i^H replaced by W_i^L. Then, if Assumption B.1 holds as stated and with W_i^H replaced by W_i^L, Assumptions 4.3 and 4.4 will hold for α specified in Assumption B.1 (and $d_X = 1$). If Assumption B.2 holds as stated and with W_i^H replaced with W_i^L, Assumptions 4.3 and 4.4 will hold for $\alpha = 2$ (and $d_X = 1$).*

B.3 Selection Model and Identification at the Boundary

In this section, I treat a class of models in which the conditional moment inequalities give the most identifying information when conditioning on a set where X_i may not have a density that is bounded away from zero and infinity. That is, as θ approaches the identified set, the moment inequality $E_P(m(W_i, \theta)|X_i = x) \geq 0$ is violated on a region in which the density of X_i goes to zero or infinity, or in which X_i does not have a density with respect to the Lebesgue measure. This covers cases of conditional moment inequalities leading to point or set identification at infinity or at a finite boundary. While I motivate the conditions in this section with a selection model, the results apply more generally to other cases of set identification at the boundary.

The selection model is particularly interesting in that it leads naturally to different shapes of the conditional mean of $m(W_i, \theta)$ and distribution of X_i, since set identification at the boundary of the support of X_i appears to be a common case. For cases where the conditioning variable has a density function that goes to zero or infinity near a (possibly infinite) support point, a transformation of the conditioning variable leads to a model for which the smoothness assumptions for rates of convergence given in this paper can be verified. The resulting value of the Hölder constant α depends on the shape of both the density and the conditional mean.

This is related to cases of point identification at infinity, such as the estimator proposed by Andrews and Schafgans (1998) for a selection model similar to the one treated in this section, but under conditions that lead to point identification. As with the estimator proposed in that paper, the estimators I consider based on KS statistics for conditional moment inequalities and possible set identification have rates of convergence that depend on the tail behavior of the random variables in the model. The behavior of distributions of random variables at the tails determines which functions in $\mathcal{G}$ correspond to the region of the tail of the conditioning variable with the most identifying power. The truncated variance weighting I propose allows the KS statistic to automatically find these functions.

We are interested in the marginal distribution of a random variable Y_i^*, but we do not always observe this variable. Instead, we observe (Y_i, D_i) where D_i is an indicator for being

observed in the sample and $Y_i \equiv Y_i^* \cdot D_i$. For example, suppose we are interested in the distribution of wage offers for a population of individuals, but we only observe wages of people who decide to work. In this case, Y_i^* is the wage individual i is offered, and D_i is an indicator for employment. In what follows, Y_i and D_i are scalars, but the results described below can be extended to multiple partially observed outcomes. In the treatment effects literature, potential outcomes under different treatment programs are typically modeled as latent variables, with the observed variable being the actual treatment. In this case, we can consider each possible treatment separately, each time defining Y_i^* and D_i to be potential outcomes and indicators for the treatment group in question. Bounds on the marginal distribution for each treatment will follow from methods described in this section, and these bounds can be combined to give bounds on treatment effects defined as differences between statistics of the unobserved distribution of each outcome.

If Y_i^* is not independent of D_i and $D_i = 0$ with positive probability, the distribution of Y_i will be different from the distribution of Y_i^* conditional on entry. However, it is often possible to obtain informative bounds. Suppose that we observe a random variable X_i that shifts participation in the sample, but is exogenous to outcomes in the sense that Y_i^* is independent of X_i. If Y_i is known to lie in some interval $[\underline{Y}, \overline{Y}]$, we can bound the distribution of Y_i^* following Manski (1990). In this section, I consider estimation of bounds for the mean of the distribution of Y_i^*, but bounds on quantiles can be estimated using similar methods. For the same reasons as those described in Section B.1, bounds on quantiles will often be tighter than bounds on the mean when the difference between $\underline{Y}$ and $\overline{Y}$ is large or infinite.

To see how this model fits into the framework of this paper, note that $Y_i \cdot D_i + \underline{Y} \cdot (1 - D_i) \leq Y_i^* \leq Y_i \cdot D_i + \overline{Y} \cdot (1 - D_i)$, so that, letting $\gamma = E_P(Y_i^*) = E_P(Y_i^*|X)$, we have $E_P(Y_i \cdot D_i + \underline{Y} \cdot (1 - D_i)|X) \leq \gamma \leq E_P(Y_i \cdot D_i + \overline{Y} \cdot (1 - D_i)|X)$. Define $W_i^L = Y_i \cdot D_i + \underline{Y} \cdot (1 - D_i)$ and $W_i^H = Y_i \cdot D_i + \overline{Y} \cdot (1 - D_i)$. The problem of estimating the identified set for γ fits into the framework of this paper with $W_i = (W_i^L, W_i^H, X_i)$ and $m(W_i, \gamma) = (\gamma - W_i^L, W_i^H - \gamma)'$.

Typically, the best upper and lower bounds on γ will come from values of X_i for which the probability of participation is high. If participation is monotonic, these points will be near the boundary of the support of X_i. The support of X_i could be infinite or finite, and there is typically no reason to impose any conditions on how the distribution of X_i behaves near its support points (whether it has a density, whether the density approaches zero, infinity, a positive constant, or oscillates wildly) or how $E_P(W_i^H|X_i)$ and $E_P(W_i^L|X_i)$ behave near these points. In addition, while identification at the boundary of the support seems likely, it is best not to impose this either.

The results in this section show that estimates of the identified set using weighted KS statistics defined above are robust to all of these types of set identification in the sense of controlling the probability that the set estimate fails to contain the identified set uniformly in a set of underlying distributions that contains these types of distributions and many more. In addition, for a wide variety of shapes of the density and conditional mean, the weighted KS statistic based set estimate obtains a better rate of convergence than estimates that do not weight the KS statistic.

Uniform coverage of the identified set follows immediately from Theorem 3.1, and is stated in the next theorem. Throughout this section, $\Theta_0(P)$ denotes the identified set for γ in the selection model under P, and $\mathcal{C}_n(\hat{c}_n)$ denotes an estimate of this set as described above.

Theorem B.3. *Let $\mathcal{P}$ be any class of probability measures on the random variables in the selection model described above such that W_i^H and W_i^L are bounded uniformly over $P \in \mathcal{P}$. If Assumptions 3.1, 3.2 and 3.3 hold and σ_n and $\hat{c}_n$ are chosen so that the assumptions of Theorem 3.1 hold, then*

$$\inf_{P \in \mathcal{P}} P(\Theta_0(P) \subseteq \mathcal{C}_n(\hat{c}_n)) \overset{n \to \infty}{\to} 1.$$

Rates of convergence to the identified set will depend on the shape of the conditional mean and the distribution of X_i. Note, however, that the set estimate based on the standard deviation weighted KS statistic can be calculated in the same manner regardless of these aspects of the data, so the researcher does not have to impose any restrictions on the shapes of these objects when performing inference. In this sense, inference based on these statistics adapts to the shapes of the conditional means of W_i^H and W_i^L and the distribution of X_i. In what follows, I consider several alternative assumptions. These include different types of set identification at the boundary, as well as set identification on a positive probability set.

In the following assumptions, $[\underline{\gamma}, \overline{\gamma}]$ is the identified set for γ, so that it is implicitly assumed that $E_P(W_i^H | X_i) \geq \overline{\gamma}$ and $E_P(W_i^L | X_i) \leq \underline{\gamma}$ with probability one. Here, $\underline{\gamma}$ and $\overline{\gamma}$ could be equal, leading to point identification. This will be the case when the probability of selection into the sample conditional on $X_i = x$ converges to one as x approaches some point on the support of X_i. These assumptions are stated so that the same type of identification holds for the upper and lower support of the identified set, but the same results will hold (with possibly different rates of convergence to the upper and lower support points) if different types of identification hold for the upper and lower support. When these assumptions are

invoked for a class of probability distributions $\mathcal{P}$, the constants C, K_X, and η_X are assumed not to depend on P.

Assumption B.5 (Set Identification at Infinity with Polynomial Tails)**.** *$d_X = 1$ and, for some positive constants K_X and C, we have, for all $x \geq K_X$, (i) $E_P(W_i^H | X_i = x) - \overline{\gamma} \leq Cx^{-\phi_m}$ and (ii) X_i has a density $f_X(x)$ such that $f_X(x) \geq x^{-\phi_x}/C$ for some $\phi_m > 0$ and $\phi_x > 1$. In addition, part (i) holds with $W_i^H - \overline{\gamma}$ replaced by $\underline{\gamma} - W_i^L$.*

Assumption B.6 (Set Identification at Finite Support with Polynomial Tails)**.** *For some $x_0 \in \mathbb{R}^{d_X}$ and $\eta_X > 0$, we have, for $x_0 - \eta_X \iota \leq x \leq x_0$ (where ι is a vector of ones and $\leq$ is elementwise if $d_X > 1$) (i) $E_P(W_i^H | X_i = x) - \overline{\gamma} \leq C|x_0 - x|^{\phi_m}$ and (ii) X_i has a density $f_X(x)$ such that $f_X(x) \geq \prod_{k=1}^{d_X} |x_{0,k} - x_k|^{\phi_x}/C$ for some $\phi_m > 0$ and some $\phi_x > -1$. In addition, parts (i) and (ii) hold with $W_i^H - \overline{\gamma}$ replaced by $\underline{\gamma} - W_i^L$ for some possibly different x_0.*

Assumption B.7 (Set Identification on a Positive Probability Set)**.** *For some interval $[\underline{x}, \overline{x}]$, $E_P(W_i^H | X_i) - \overline{\gamma} = 0$ P-a.s. for all $P \in \mathcal{P}$ and $P(\underline{x} \leq X_i \leq \overline{x})$ is bounded away from zero uniformly in $P \in \mathcal{P}$. In addition, the same assumption holds with $W_i^H - \overline{\gamma}$ replaced by $\underline{\gamma} - W_i^L$ for some possibly different interval $[\underline{x}, \overline{x}]$.*

All cases of Assumption B.5 and B.6 can be transformed into Assumption B.6 with $\phi_x = 0$ and some ϕ_m by monotonic transformations of each element of X_i. The case where Assumption B.6 holds with $\phi_x = 0$ fits into the framework of Theorem 4.3, so this can be applied to the transformed model.

Theorem B.4. *Let $\mathcal{P}$ be any class of probability measures on the random variables in the selection model described above such that W_i^H and W_i^L are bounded uniformly over $P \in \mathcal{P}$. Suppose that Assumptions 3.1, 3.2 and 3.3 hold and σ_n and $\hat{c}_n$ are chosen so that the assumptions of Theorem 3.1 hold, and Assumption 4.6 holds.*

If, in addition to these conditions, one of Assumptions B.5 or B.6 holds, then, for some B,

$$\sup_{P \in \mathcal{P}} P\left(\left(\frac{n}{\hat{c}_n^2 \log n} \right)^{\alpha/(d_X + 2\alpha)} d_H(\mathcal{C}_n(\hat{c}_n), \Theta_0(P)) > B \right) \overset{n \to \infty}{\to} 0$$

where $\alpha = \phi_m/(\phi_x + 1)$ if Assumption B.6 holds and $\alpha = \phi_m/(\phi_x - 1)$ (and $d_X = 1$) if

Assumption B.5 holds. If Assumption B.7 holds, then, for some B,

$$\sup_{P\in\mathcal{P}} P\left(\left(\frac{n}{\hat{c}_n^2 \log n}\right)^{1/2} d_H(\mathcal{C}_n(\hat{c}_n), \Theta_0(P)) > B\right) \overset{n\to\infty}{\to} 0.$$

The rate of convergence in Theorem B.4 shows that, for a given selection process conditional on X_i, the rate of convergence will be faster when X_i has more mass near the point x_0 or region $[\underline{x}, \overline{x}]$ where the conditional moment inequalities give the most identifiying information. The rate of convergence is fastest ($((\log n)/n)^{1/2}$) under Assumption B.7, when this region has a positive probability. Under identification at a finite point (Assumption B.6), the rate of convergence depends on whether the density of X_i approaches infinity, zero, or a finite nonzero value. If $-1 < \phi_x < 0$, the density will approach infinity at a rate that is faster when ϕ_x is closest to -1 (ϕ_x must be strictly greater than -1 in order for the density to integrate to a finite number). For $\phi_x = 0$, the density approaches a finite nonzero value, and, for $\phi_x > 0$ the density approaches zero at a rate that is faster for larger values of ϕ_x. The rate of convergence under Assumption B.6 will always be slower than $((\log n)/n)^{1/2}$, but it will be arbitrarily close to this rate when ϕ_x is close to -1 (when the density approaches infinity at close to the fastest possible rate). Under identification at infinity (Assumption B.5), the rate of convergence will be faster for thicker tails (smaller ϕ_x), and will be close to $((\log n)/n)^{1/2}$ for ϕ_x close to 1 (in this case, ϕ_x must be greater than one in order for the density to integrate to a finite number).

References

Andrews, D. W., S. Berry, and P. Jia (2004): "Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location," .

Andrews, D. W., and P. Guggenberger (2009): "Validity of Subsampling and ?plug-in Asymptotic? Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25(03), 669–709.

Andrews, D. W., and X. Shi (2009): "Inference Based on Conditional Moment Inequalities," *Unpublished Manuscript, Yale University, New Haven, CT*.

Andrews, D. W. K., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *SSRN eLibrary*.

Andrews, D. W. K., and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65(3), 497–517.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78(1), 119–157.

Armstrong, T., and H. P. Chan (2012): "Multiscale Adaptive Inference on Conditional Moment Inequalities," *Unpublished Manuscript*.

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

Bickel, P. J., and M. Rosenblatt (1973): "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, pp. 1071–1095.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20(1), 105–134.

Bugni, F. A. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78(2), 735–753.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2009): "Intersection bounds: estimation and inference," *Arxiv preprint arXiv:0907.3503*.

Chetverikov, D. (2012): "Adaptive Test of Conditional Moment Inequalities," .

Dumbgen, L., and V. G. Spokoiny (2001): "Multiscale Testing of Qualitative Hypotheses," *The Annals of Statistics*, 29(1), 124–152.

Galichon, A., and M. Henry (2009): "A test of non-identifying restrictions and confidence regions for partially identified parameters," *Journal of Econometrics*, 152(2), 186–196.

Ibragimov, I. A., and R. Z. Hasminskii (1981): *Statistical estimation–asymptotic theory /*. New York :.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Ingster, Y., and I. A. Suslina (2003): *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer.

Khan, S., and E. Tamer (2009): "Inference on endogenously censored regression models using conditional moment inequalities," *Journal of Econometrics*, 152(2), 104–119.

Kim, K. i. (2008): "Set estimation and inference with models characterized by conditional moment inequalities," .

Lehmann, E. L., and J. P. Romano (2005): *Testing statistical hypotheses*. Springer.

Lepski, O., and A. Tsybakov (2000): "Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point," *Probability Theory and Related Fields*, 117(1), 17–48.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80(2), 319–323.

Menzel, K. (2008): "Estimation and Inference with Many Moment Inequalities," *Preprint, Massachussetts Institute of Technology*.

Menzel, K. (2010): "Consistent Estimation with Many Moment Inequalities," *Unpublished Manuscript*.

Moon, H. R., and F. Schorfheide (2009): "Bayesian and Frequentist Inference in Partially Identified Models," *National Bureau of Economic Research Working Paper Series*, No. 14882.

Pollard, D. (1984): *Convergence of stochastic processes*. David Pollard.

Ponomareva, M. (2010): "Inference in Models Defined by Conditional Moment Inequalities with Continuous Covariates," .

Romano, J. P., and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.

Romano, J. P., and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10(4), 1040–1053.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77(4), 1299–1315.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak convergence and empirical processes*. Springer.

statistic		**weighted**	**weighted**	**weighted**	**unweighted**	**kernel**	**kernel**
σ_n or h_n		$\frac{1}{2}n^{-1/6}$	$\frac{1}{2}n^{-1/10}$	$\frac{1}{2}\left[\frac{(\log n)(\log\log n)}{n}\right]^{1/2}$	-	$(\overline{x}-\underline{x})n^{-1/3}$	$(\overline{x}-\underline{x})n^{-1/5}$
	$n=200$	0.48	0.45	0.49	0.44	0.7	0.53
(β_1,β_2)	$n=500$	0.34	0.32	0.37	0.29	0.52	0.37
	$n=1000$	0.28	0.26	0.3	0.23	0.42	0.3
	$n=200$	0.48	0.44	0.49	0.4	0.7	0.53
β_1	$n=500$	0.34	0.32	0.37	0.28	0.52	0.37
	$n=1000$	0.28	0.25	0.29	0.22	0.42	0.3
	$n=200$	0.3	0.31	0.31	0.41	0.34	0.31
β_2	$n=500$	0.21	0.2	0.23	0.27	0.24	0.2
	$n=1000$	0.16	0.15	0.18	0.2	0.19	0.15

Table 1: Median Distances for Set Estimators in Monte Carlos ($\hat{c}_n = c_{.95,n} \cdot \sqrt{\log\log n}$)

statistic σ_n $\hat{c}_n$		**weighted** $\frac{1}{2}n^{-1/6}$ $c_{.95,n}$	**weighted** $\frac{1}{2}n^{-1/6}$ $c_{.95,n} \cdot \sqrt{\log n}$
(β_1, β_2)	$n = 200$	0.39	2.7
	$n = 500$	0.27	0.54
	$n = 1000$	0.21	0.44
β_1	$n = 200$	0.39	0.96
	$n = 500$	0.27	0.54
	$n = 1000$	0.21	0.44
β_2	$n = 200$	0.23	2.42
	$n = 500$	0.15	0.4
	$n = 1000$	0.11	0.3

Table 2: Median Distances for Set Estimators Based on Variance Weighted Statistic (Other Critical Values)

σ_n or h_n $\hat{c}_n$	**weighted** $\frac{1}{2}n^{-1/6}$ $c_{.95,n} \cdot \sqrt{\log \log n}$	**unweighted** - $c_{.95,n} \cdot \sqrt{\log \log n}$
$n = 200$	0.3186	0.2405
$n = 500$	0.2092	0.1467
$n = 1000$	0.1623	0.1155

Table 3: Median Distances for Flat Conditional Mean