Testing Limited Overlap

Xinwei Ma* Yuya Sasaki† Yulong Wang‡

January 31, 2022

Abstract

Limited overlap, reflected either by a large discrepancy in covariate distributions between the treatment and the control group, or by the presence of extreme propensity scores, can be a threat to the estimation of and inference on treatment effect parameters. In this paper, we propose a formal statistical test which helps assess the degree of limited overlap. Rejecting the null hypothesis in our test indicates either no or very mild degree of limited overlap, and hence reassures that standard treatment effect estimators will be well-behaved. One distinguishing feature of our test is that it only requires the use of a few extreme propensity scores, which is in stark contrast to other methods that require consistent estimates of some tail index. Without the need to extrapolate using observations far away from the tail, our procedure is expected to exhibit excellent size properties, a result that is also borne out in our simulation study.

Keywords: Limited overlap, Treatment effect, Propensity score, Inverse probability weighting, Extreme value theory.

JEL Codes: C12, C13, C21

*Department of Economics, University of California San Diego. 9500 Gilman Dr. #0508, La Jolla, CA 92093 x1ma@ucsd.edu

†Department of Economics, Vanderbilt University. 415 Calhoun Hall, Vanderbilt University, Nashville, TN 37240 yuya.sasaki@vanderbilt.edu

‡Department of Economics, Syracuse University. 127 Eggers Hall, Syracuse University, Syracuse, NY, 13244 ywang402@syr.edu

1 Introduction

Following the seminal work by Rubin (1974, 1997), Rosenbaum and Rubin (1983), and Rosenbaum (1989), there is an extensive literature on estimating treatment effects with observational data – see Imbens and Rubin (2015) and Abadie and Cattaneo (2018). A key assumption for estimating the average treatment effect is the *strong overlap*, which requires that the probability of receiving treatment conditional on the covariates is bounded away from zero and one. In some applications, however, one may observe extreme propensity scores.

Limited overlap can be detrimental to commonly used statistical procedures, as they may converge at a slower rate (Khan and Tamer, 2010; Hong, Leung, and Li, 2020) and their limiting distributions may not be Gaussian (Ma and Wang, 2020). A common empirical strategy is to restrict the target population by trimming observations in the region of poor overlap, so that the overlap assumption is satisfied for the subpopulation obtained after trimming (Crump, Hotz, Imbens, and Mitnik, 2009; Chaudhuri and Hill, 2014; Ma and Wang, 2020; Sasaki and Ura, 2021).

To assess the degree of limited overlap and to decide if trimming is needed, visual diagnosis, such as plotting a histogram or a nonparametric density curve of the estimated propensity score, is commonly used in applied studies. To our best knowledge, however, the existing literature lacks a formal method to test the overlap assumption. We fill this gap by proposing a novel statistical test of limited overlap. In addition, our procedure accounts for the fact that the propensity score needs to be estimated in a first step. With this new device, researchers may for instance (i) test whether the assumption of overlap holds for estimating the average treatment effect (Hirano, Imbens, and Ridder, 2003; Cattaneo, 2010; Farrell, 2015; Belloni, Chernozhukov, Chetverikov, Hansen, and Kato, 2021; Farrell, Liang, and Misra, 2021), and (ii) test whether a subpopulation obtained after trimming satisfies the overlap condition for estimating subpopulation average treatment effects. If the testing result supports the violation of the overlap condition, researchers may continue with

alternative statistical inference methods that are robust to limited overlap. The following paragraph provides a short review on these methods.

Ma and Wang (2020) show that the standard inverse probability weighting estimator may not be asymptotically Gaussian if the propensity scores can be arbitrarily close to 0 or 1, and trimming may further complicate the limiting distribution. As a remedy, they recommend subsampling for conducting inference and constructing robust confidence intervals. The non-Gaussian limiting distribution also features in the doubly robust approach, as shown by Heiler and Kazak (2021). In the context of inverse density weighting, Khan and Nekipelov (2011) consider a local departure from the asymptotic Gaussian regime and study the properties of subsampling. Another approach to robust inference builds on parametric or shape restrictions on the underlying statistical model, see, for example, Rothe (2017) and Armstrong and Kolesár (2021). Our proposed testing procedure complements this literature as it allows researchers to rigorously test the severity of the overlap issue. Importantly, as we will demonstrate below, rejection of the null hypothesis indicates either no or very mild limited overlap, and hence justifies the use of standard $\sqrt{n}$-Gaussian inference/confidence intervals, eliminating the need of trimming or employing strong distributional assumptions.

2 Overview

We first present an overview of the proposed statistical test. Assume there is a random sample of size n consisting of (X_i, D_i), $i = 1, 2, \cdots, n$, where $X_i \in \mathbb{R}^p$ collects all covariates of the ith individual, and $D_i \in \{0, 1\}$ is a binary indicator (say, of treatment status). The propensity score is defined as the probability of receiving treatment conditional on an individual's covariates, that is, $\mathbb{P}[D_i = 1|X_i] = e(X_i) =: e_i$. To test, for example, the presence of small propensity scores, our procedure concerns the following competing hypotheses

$$\mathsf{H}_0 : \mathbb{E}\left[1/e_i\right] = \infty \quad \text{against} \quad \mathsf{H}_1 : \mathbb{E}\left[1/e_i\right] < \infty. \tag{2.1}$$

One distinguishing feature of our test is that rejecting the null indicates either no or very mild degree of limited overlap. In turn, this suggests that standard treatment effect estimators – such as inverse probability weighting – are expected to perform well, and that inference based on root-n Gaussian approximations should remain valid.

To better illustrate the connection between our hypotheses in (2.1) and the overlap issue, we consider the inverse probability weighting estimator, $n^{-1}\sum_{i=1}^{n} D_iY_i/e_i$. Here Y_i is some outcome variable of interest, and we denote the potential outcome by $Y_i(1)$, meaning that $D_iY_i = D_iY_i(1)$. Asymptotic Gaussianity requires that the ratio, D_iY_i/e_i, to have a finite second moment, which is equivalent to

$$\mathbb{E}\left[\left|\frac{D_iY_i}{e_i}\right|^2\right] = \mathbb{E}\left[\frac{1}{e_i}\mathbb{E}\big[|Y_i(1)|^2\big|X_i\big]\right] < \infty.$$

This should explain why we set our null and alternative hypotheses as the finiteness of the expectation of the inverse propensity score.[1]

Testing the hypotheses in (2.1) boils down to investigating tail properties of the propensity score. Hence, we employ the following self-normalized statistic

$$\mathbf{T} = \frac{1}{e_{(1)}^{-1} - e_{(k)}^{-1}}\left(1,\ e_{(2)}^{-1} - e_{(k)}^{-1},\ e_{(3)}^{-1} - e_{(k)}^{-1},\ \cdots,\ e_{(k-1)}^{-1} - e_{(k)}^{-1},\ 0\right)', \tag{2.2}$$

where $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(k)}$ correspond to the k smallest propensity scores. The limiting distribution of these statistics is solely characterized by a scalar parameter ξ that measures the tail heaviness of the distribution of e_i^{-1}. In particular, the null hypothesis in (2.1) corresponds to that $\xi \geq 1$. This elegant feature allows us to distinguish the two competing scenarios in (2.1) by using a generalized likelihood ratio approach

$$\text{reject } \mathsf{H}_0 \quad \text{if} \quad \left(\int_0^1 f_\xi\left(\mathbf{T}\right)\mathrm{d}W_1(\xi)\right) \Big/ \left(\int_1^{\bar{\xi}} f_\xi\left(\mathbf{T}\right)\mathrm{d}W_0(\xi)\right) > \mathrm{cv}_\alpha, \tag{2.3}$$

[1] Of course, an infinite second moment of the ratio can also be the result of heavy-tailed outcome variables (i.e., $\mathbb{E}[Y_i(1)^2] = \infty$), but we do not discuss this issue as our focus is on limited overlap.

where $f_\xi(\cdot)$ is the limiting (as $n \to \infty$ with fixed k) joint density of $\mathbf{T}$ parameterized by the tail index ξ of the propensity score distribution, and cv_α is the critical value which can be easily obtained via simulation. The weights W_1 and W_0 respectively transform the composite alternative and null hypotheses into simple ones by resorting to the weighted average power (Andrews and Ploberger, 1995) and the approximate least favorable distribution (Elliott, Müller, and Watson, 2015). The item $\bar{\xi}$ is the upper bound of the null space over which one would like to impose the size constraint. We use $\bar{\xi} = 2$ in later exercises, which can be easily extended.

While we leave further details to Section 3, it is worth mentioning that our procedure does not require k to diverge to infinity, which is in stark contrast to other methods that require consistent estimates of some tail index (e.g., Hill, 1975). Precisely, we characterize ahead in (3.2) the limiting density $f_\xi(\cdot)$ for any fixed k, and hence the generalized likelihood ratio approach in (2.3) will remain valid even if k is small relative to the sample size. Allowing for a small k is crucial in samples of moderate size, as researchers can focus on a few extreme propensity scores in their analysis, avoiding any extrapolation using observations far away from the tail. In other words, our procedure is expected to be more robust with respect to the choice of k and will exhibit better size properties. This type of robustness benefiting from fixed order k has been similarly explored in Müller and Wang (2017) for inference about extreme quantiles and tail conditional expectations[2], and in Sasaki and Wang (2020) for testing the bounded moment conditions in extremum estimation and generalized method of moments.

To operationalize the generalized likelihood approach in (2.2) and (2.3), one usually needs to estimate the propensity score in a first step. As we will demonstrate below, estimating the propensity score will not affect the asymptotic properties of the test, provided that the estimated propensity score, denoted by $\hat{e}_i$, satisfies a uniform consistency requirement: $\max_{1\le i\le n} |e_i/\hat{e}_i| \to 1$ in probability. Given that this is a nontrivial assumption, and is

[2]Müller and Wang (2017) require observing the extreme values, which are not available in our setup since the propensity scores are estimated. See Assumption 2 ahead for more details.

generally not implied by standard extremum estimation results (for example, Newey and McFadden 1994), we provide primitive sufficient conditions in Section 4. In particular, we first consider two widely adopted parametric propensity score specifications, Logit and Probit, and show that the strong uniform consistency requirement holds as long as the covariates have a few finite moments. Of course, a parametric model can be restrictive, and hence we also propose a semiparametric propensity score estimator following Bierens (2014). Together with our generalized likelihood approach, this paper offers a comprehensive toolkit for propensity score estimation and the assessment of limited overlap.

3 Main Results

Recall that the propensity score is defined as the conditional probability of receiving a treatment, that is, $\mathbb{P}[D_i = 1|X_i] = e(X_i) =: e_i$. The statistical decision of rejecting the null hypothesis in (2.1) depends solely on the left-tail (close-to-zero) heaviness of the propensity score. To proceed, we assume that the distribution of the propensity score has a regularly varying tail at zero.

Assumption 1. The distribution of e_i satisfies

$$\lim_{t \downarrow 0} \frac{\mathbb{P}[e_i \leq tu]}{\mathbb{P}[e_i \leq t]} = u^{\frac{1}{\xi}} \quad \forall u > 0$$

for some tail index ξ.

We make three remarks. First, the condition that e_i is regularly varying at 0 is equivalent to that $1/e_i$ is regularly varying at infinity. In particular,

$$u^{-\frac{1}{\xi}} = \lim_{t \downarrow 0} \frac{\mathbb{P}\left[e_i \leq tu^{-1}\right]}{\mathbb{P}\left[e_i \leq t\right]} = \lim_{t \uparrow \infty} \frac{1 - F_{e(X)^{-1}}(tu)}{1 - F_{e(X)^{-1}}(t)},$$

where $F_{e(X)^{-1}}(\cdot)$ denotes the distribution function of the inverse propensity score. Then, it follows that $\mathbb{E}\left[1/e(X_i)\right]$ is finite/infinite if ξ is below/above 1, and hence the hypotheses in

(2.1) can be rewritten as

$$\mathsf{H}_0 : \xi \in [1, \bar{\xi}\,] \quad \text{against} \quad \mathsf{H}_1 : \xi \in (0, 1)\,. \tag{3.1}$$

In the above, $\bar{\xi}$ denotes the upper bound of the null space that collects all values of ξ on which we will require size control. It can be ∞ in principle, but for numerical practice we use a large but finite value, say $\bar{\xi} = 2$ in simulations. It turns out that our procedure is not sensitive to the choice of $\bar{\xi}$, as our numerical experiments suggest that $\xi = 1$ is the "least favorable" model in the null space. Second, the regular variation assumption on the propensity score distribution is mild and is satisfied by many commonly used distributions, such as the Pareto, Student-t, Beta, and F distributions. Third, the alternative space can be easily extended to cover negative values of ξ, which imply that e_i^{-1} has a finite right end-point.

To test the hypotheses in (3.1), one possibility is to estimate the tail index (e.g., Hill, 1975). This approach, however, requires using k smallest propensity scores with $k \to \infty$ at a certain rate. The choice of k is often delicate, as employing a large k corresponds to extrapolation based on observations that are far away from the tail. Therefore, instead of relying on some consistent estimate of the tail index, we directly consider the large-sample distribution of the k smallest propensity scores, and hence our approach is valid for any fixed k.

For ease of exposition, we first assume that the true propensity scores are observed, and later discuss the impact of estimating the propensity score. Consider the k smallest propensity scores, $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(k)}$. By the extreme value theory (e.g., de Haan and Ferreira, 2007, Chapter 1), Assumption 1 implies that there exist sequences of constants a_n and b_n such that[3]

$$\tilde{\mathbf{T}} = \frac{1}{a_n}\left(e_{(1)}^{-1} - b_n,\ e_{(2)}^{-1} - b_n,\ \cdots,\ e_{(k)}^{-1} - b_n\right)' \Rightarrow \mathbf{V},$$

[3]In the example with the standard Pareto distribution, a_n is n^{ξ} and b_n is zero.

where $\mathbf{V}$ has the following density ,

$$f_{\mathbf{V}|\xi}(v_1, ..., v_k) = G_\xi(v_k) \prod_{i=1}^{k} g_\xi(v_i)/G_\xi(v_i) \text{ on } v_k \leq v_{k-1} \leq \ldots \leq v_1$$

with $G_\xi(v) = \exp(-(1+\xi v)^{-1/\xi})$ and $g_\xi(v) = \mathrm{d}G_\xi(v)/\mathrm{d}v$. If the scaling and centering sequences, a_n and b_n, were known, the above distributional approximation can be used for testing our hypotheses in (2.1) and (3.1). Unfortunately, a_n and b_n depend on the distribution $F_{e(X)^{-1}}(\cdot)$ and are usually difficult to estimate.

To avoid estimating the centering and scaling in $\tilde{\mathbf{T}}$, we consider the self-normalized statistic in (2.2). Specifically, let

$$\mathbf{T} = \frac{\tilde{\mathbf{T}} - \tilde{\mathbf{T}}_k}{\tilde{\mathbf{T}}_1 - \tilde{\mathbf{T}}_k},$$

where $\tilde{\mathbf{T}}_1$ and $\tilde{\mathbf{T}}_k$ are the first and last elements in $\tilde{\mathbf{T}}$. It is easy to establish that $\mathbf{T}$ is maximal invariant with respect to the group of location and scale transformations (e.g., Lehmann and Romano, 2005, Chapter 6). In other words, the test statistic as a function of $\mathbf{T}$ remains unchanged if the data is shifted and multiplied by any non-zero constant. Such invariance property is desirable for our purposes as tail features should preserve no matter how the data is linearly transformed.

It follows from the continuous mapping theorem that

$$\tilde{\mathbf{T}} \;\Rightarrow\; \left(1,\; \frac{V_2 - V_k}{V_1 - V_k},\; \frac{V_3 - V_k}{V_1 - V_k},\; \cdots,\; 0\right)'$$

whose density function becomes

$$f_\xi(\mathbf{t}) = \Gamma(k) \int_0^\infty u^{k-2} \exp\left(-(1+1/\xi)\left(\sum_{i=2}^{k-1} \log(1+\xi t_i u) + \log(1+\xi u)\right)\right) \mathrm{d}u, \qquad (3.2)$$

where $\Gamma(\cdot)$ is the gamma function. (Recall that the first and the last elements of $\mathbf{T}$ are

respectively one and zero by construction.)

Given a random draw $\mathbf{T}$ from the density (3.2), the hypotheses in (2.1) and (3.1) can be tested by the generalized likelihood ratio statistic in (2.3). We now give more details about the weights $W_1(\cdot)$ and $W_0(\cdot)$. First, $W_1(\cdot)$ is a weighting function specified by the analyst, which reflects the importance allocated to different values of ξ in the alternative space. We employ the uniform distribution in our simulation study and the empirical applications, which can be easily modified. The weights, $W_0(\cdot)$, can be understood as the least favorable distribution (e.g., Lehmann and Romano, 2005, Chapter 3), which are defined on the null space $\left[1, \bar{\xi}\right]$ to maintain the asymptotic size control that for any fixed $\xi \in \left[1, \bar{\xi}\right]$

$$\lim_{n\to\infty} \mathbb{P}[\text{Reject } \mathsf{H}_0] \leq \alpha. \tag{3.3}$$

To determine $W_0(\cdot)$ and the critical value cv_α, we resort to the numerical algorithm developed by Elliott, Müller, and Watson (2015) and Müller and Wang (2017). The output of this numerical algorithm is an *approximate* least favorable distribution, which turns out to be the point mass allocated solely on $\xi = 1$ in our simulations. Across various values of k and α, $W_0(\cdot)$ and cv_α only need to be computed once. We provide the Matlab algorithms and tabulate the results of cv_α in the supplementary material.

Figure 1 presents the rejection probability of the test (2.3), averaged from 10,000 random draws simulated from (3.2) with $\xi \in [0.01, 2]$. The figure is plotted in the reverse order of ξ so that $\xi \in [1, 2]$ corresponds to the null hypothesis. For any k, the test controls size for any ξ under the null hypothesis and obtains a monotonically increasing power as ξ decreases to zero. Comparing the results across different values of k, the power of the test increases as k increases. Unfortunately a theoretically optimal choice of k is very challenging to obtain, if possible at all. This is also the reason we adopt the fixed k asymptotics so that our test controls size for any pre-determined k as long as n is sufficiently large. We recommend practitioners to report results with a variety of k for the sake of sensitivity analysis, as we

will do in Section 6. Additional discussions and a rule-of-thumb choice are given in the Section D to conserve space.

[FIGURE 1 HERE]

The only remaining challenge is to obtain some feasible analog of $\mathbf{T}$, as the propensity score is usually unknown in practice. To this end, we first construct some consistent estimator $\hat{e}(X_i) =: \hat{e}_i$ of the propensity score $e(X_i)$. Then we take the smallest k estimated propensity scores $\hat{e}_{(1)} \leq \hat{e}_{(2)} \leq \cdots \leq \hat{e}_{(k)}$ and construct the self-normalized statistic $\hat{\mathbf{T}}$ similarly to (2.2). Fortunately, $\hat{\mathbf{T}}$ will have the same asymptotic distribution as $\mathbf{T}$ under the following high-level uniform consistency assumption. Primitive sufficient conditions will be provided in Section 4.

Assumption 2.

$$\max_{1\leq i\leq n}\left|\frac{e_i}{\hat{e}_i}-1\right| = o_p(1).$$

Assumption 2 indicates that the estimation error of the propensity score is asymptotically dominated in magnitude by the true large order statistics of $1/e_i$. There are many estimators in the existing literature that satisfy this assumption. We discuss two commonly used methods and their primitive assumptions for Assumption 2 in detail in the next section.

Under Assumptions 1 and 2, the following theorem presents the main result of this article.

Theorem 1. *Suppose that (D_i, X_i) is i.i.d. and Assumptions 1 and 2 are satisfied. Then* (3.3) *holds for any fixed k.*

We close this section by discussing some features of the test in (2.3). First, our testing procedure controls size over all values of ξ under the null hypothesis. This feature can be appealing because a practitioner may not know *ex ante* the tail heaviness ξ for data in use.

Second, our fixed-k asymptotic framework differs from the literature where it is typically assumed that $k = k_n \to \infty$ and $k_n/n \to 0$. We adopt this fixed-k design instead of the

increasing-k design for two reasons. On the one hand, the uniform size control property holds for any pre-determined k, while the methods based on an increasing k_n inevitably involve the delicate balance between the two restrictions on how fast k_n can grow. Such delicacy may lead to a poor finite sample performance when the sample size is only moderate (Müller and Wang 2017). On the other hand, the fixed tuning parameter substantially weakens the primitive conditions for Assumption 2 – see Section 4 below. The intuition follows from the fact that $\hat{e}_i^{-1} - e_i^{-1}$ is $o_p(1)$ while the largest order statistics of e_i^{-1} are $O_p(n^\xi)$.

Third, while we present our test for the left tail of the propensity score, the same technique applies to the right tail by employing order statistics of $1 - \hat{e}_i$. This is useful, for example, when the goal is to estimate the treatment effect on the treated where limited overlap arises if the propensity scores can be close to 1. As another example, it is possible to simultaneously test the presence of both small and large propensity scores by considering the hypotheses $\mathsf{H}_0 : \mathbb{E}[1/(e_i(1-e_i))] = \infty$ vs. $\mathsf{H}_1 : \mathbb{E}[1/(e_i(1-e_i))] < \infty$. This problem is essentially identical to jointly testing that at least one of $\mathbb{E}[1/e_i]$ and $\mathbb{E}[1/(1-e_i)]$ is infinite – see Lemma 1 in Appendix A for a formal result and more discussion. The test statistic will then employ the smallest k_1 order statistics of $\hat{e}_i$ and the smallest k_2 order statistics of $1 - \hat{e}_i$ for some fixed k_1 and k_2. They are asymptotically independent, and therefore the joint density of the self-normalized statistics $\mathbf{T}$ for both the left and the right tails is simply the product of their marginal densities (e.g. Arnold, Balakrishnan, and Nagaraja, 2008, Chapter 8).

Fourth, as another extension of our test, it is possible to consider a procedure that employs the treated sample only, that is, one can construct the test statistic, $\mathbf{T}$, using the smallest k order statistics of $\hat{e}_i$ from the $D_i = 1$ subgroup. Such a test can be appealing if the researcher believes the propensity score is more precisely estimated for the treated group. To see how the hypotheses change, we note that the conditional distribution of the propensity score still admits a regularly varying tail. More precisely, Assumption 1 implies

that (e.g., Ma and Wang, 2020, Lemma 1)

$$\lim_{t \downarrow 0} \frac{\mathbb{P}\left[e_i \leq tu | D_i = 1\right]}{\mathbb{P}\left[e_i \leq t | D_i = 1\right]} = u^{\frac{1}{\varsigma}}, \qquad \varsigma = \frac{\xi}{1+\xi}.$$

Therefore, we may test the null hypothesis $\mathsf{H}_0 : \varsigma \in [0.5, \bar{\varsigma}]$ versus the alternative $\mathsf{H}_1 : \varsigma \in (0, 0.5)$, where $\bar{\varsigma}$ denoes the upper bound of the parameter space. As before, rejection of the null hypothesis will imply either no or very mild degree of limited overlap, and hence can be taken as the statistical evidence that standard inference methods based on Gaussian approximations are expected to perform well.

Finally, although we focus on the fixed-k design for the above practical and theoretical advantages, it is also possible to consider a diverging $k = k_n$ and derive the consistency of the test. In particular, one can construct some root-k consistent estimator of ξ (e.g., Hill, 1975) and the corresponding confidence interval. In this sense, our fixed-k framework can be considered as a local analysis where we aim for a powerful test against local alternatives under the uniform size control constraint.

4 Estimation of the propensity score

This section justifies Assumption 2 in a couple of commonly used models for the propensity score estimation. Section 4.1 presents the case of parametric propensity scores, covering the Logit and Probit models as special cases. Section 4.2 postulates a more flexible semiparametric setup.

4.1 Parametric estimation of the propensity score

Practitioners employing the inverse probability weighting approach routinely estimate the propensity score with parametric models. To start, consider the following specification

$$e(X_i) = G(X_i'\beta_0), \tag{4.1}$$

where $G(\cdot)$ is a known link function, and hence the propensity score is parameterized by a finite dimensional vector β_0. Depending on the specific form of the link function, various estimation methods are available, such as the maximum likelihood and the nonlinear least squares. In this subsection, we focus on the maximum likelihood approach, where an estimate of β_0 can be obtained by solving

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \sum_i D_i \ln\big(G(X_i'\beta)\big) + (1 - D_i) \ln\big(1 - G(X_i'\beta)\big), \tag{4.2}$$

which means that the estimated propensity score is $\hat{e}(X_i) = G(X_i'\hat{\beta})$. Although standard large-sample techniques (see Newey and McFadden 1994 and references therein) can be invoked to prove the consistency of the estimated propensity score, such a result will generally not be strong enough for our purpose. In particular, $|\hat{\beta} - \beta_0| = o_p(1)$ does not imply Assumption 2.

To show that not only the estimated propensity score is consistent, but also the estimation error is negligible with respect to the tails of the propensity score, we employ the following high-level assumption, and lower-level sufficient conditions for which will be discussed ahead for the Logit and Probit models in Remarks 1 and 2, respectively.

Assumption 3. Let β_0 and $\hat{\beta}$ be given by (4.1) and (4.2), respectively.

(i) $\|\hat{\beta} - \beta_0\| = O_p(1/\sqrt{n})$.

(ii) $G(\cdot)$ is continuously differentiable. There exists a vanishing sequence c_n, such that for all $\epsilon > 0$,

$$\max_{1 \le i \le n} \sup_{\|\beta - \beta_0\| \le \frac{\epsilon}{\sqrt{n}}} \left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| = O_p\left(\sqrt{n} c_n\right).$$

Part (i) requires that the estimated parameter converges at the usual $\sqrt{n}$-rate. This high-level condition is standard, and can be easily verified using extremum estimation theories. In Remarks 1 and 2 we provide further discussions on this assumption for two widely used parametric propensity score specifications: the Logit and Probit models. Part (ii) is the key regularity condition that we need to establish Assumption 2 (which, in turn, allows one to use estimated propensity scores in our testing procedure). This condition, partially motivated by Ma and Wang (2020), bridges the gap between the estimation error in $\hat{\beta}$ and the tail behavior of the link function $G(\cdot)$. In particular, the faster c_n tends to zero, the easier it is to bound the discrepancy $|e(X_i)/\hat{e}(X_i) - 1|$ (Theorem 2 below). Although Assumption 3 (ii) seems complicated, we show in remarks below that it holds in both Logit and Probit models under very mild moment conditions on the covariates.

Theorem 2. *Let the true and estimated propensity score be given by* (4.1) *and* (4.2)*, respectively. Assume Assumption 3 holds. Then Assumption 2 holds with*

$$\max_{1\leq i\leq n}\left|\frac{e(X_i)}{\hat{e}(X_i)} - 1\right| = O_p(c_n) = o_p(1).$$

Theorem 2 not only provides a formal justification for Assumption 2 for parametrically estimated propensity scores, but also establishes an order at which the difference $|e(X_i)/\hat{e}(X_i) - 1|$ shrinks uniformly. At this level of generality, however, it seems quite difficult to make the order c_n explicit. We therefore consider the Logit and Probit models.

Remark 1. Assume the Logit propensity score model, that is, $G(X_i'\beta) = e^{X_i'\beta}/(1 + e^{X_i'\beta})$, and that the following primitive assumptions hold: (i) the population moment condition $\mathbb{E}[D_i \ln\big(G(X_i'\beta)\big) + (1 - D_i)\ln\big(1 - G(X_i'\beta)\big)]$ is uniquely maximized at β_0, which is in the interior of a compact parameter space $\mathcal{B}$; (ii) $\mathbb{E}[\|X_i\|^{2+\epsilon}] < \infty$ for some $\epsilon > 0$. Then, Assumption 3 holds with $c_n = n^{-\epsilon/(4+2\epsilon)}$.

Remark 2. Assume the Probit propensity score model, that is, $G(\cdot)$ is the standard normal distribution function, and that the following primitive assumptions hold: (i) the population

moment condition $\mathbb{E}[D_i \ln\big(G(X_i'\beta)\big) + (1 - D_i) \ln\big(1 - G(X_i'\beta)\big)]$ is uniquely maximized at β_0, which is in the interior of a compact parameter space $\mathcal{B}$; (ii) $\mathbb{E}[\|X_i\|^{6+\epsilon}] < \infty$ for some $\epsilon > 0$. Then, Assumption 3 holds with $c_n = n^{-\epsilon/(12+2\epsilon)}$.

4.2 Semiparametric estimation of the propensity score

We next consider a more flexible semiparametric propensity score model

$$e(X_i) = F_0(X_i'\beta_0),$$

where the link function F_0 is unknown and is allowed to be nonparametric.

Let G be the logistic link function. Suppose that the unknown nonparametric link function F_0 can be written as $F_0 = H_0 \circ G$, where H_0 is an unknown distribution function on $[0, 1]$. By Bierens (2014, Theorem 3.1), H_0 can be written in terms of the Fourier representation

$$H_0(u) = H(u; \delta_0) := u + \frac{\Upsilon(u; \delta_0)}{1 + \sum_{j=1}^{\infty} \delta_{0j}^2},$$

where

$$\begin{aligned}\Upsilon(u; \delta) =& 2\sqrt{2} \sum_{j=1}^{\infty} \delta_j \frac{\sin(j\pi u)}{j\pi} + \sum_{j=1}^{\infty} \delta_j^2 \frac{\sin(2j\pi u)}{2j\pi} \\ &+ 2 \sum_{j=2}^{\infty} \sum_{m=1}^{j-1} \delta_j \delta_m \frac{\sin((j+m)\pi u)}{(j+m)\pi} + 2 \sum_{j=2}^{\infty} \sum_{m=1}^{j-1} \delta_j \delta_m \frac{\sin((j-m)\pi u)}{(j-m)\pi}.\end{aligned}$$

With this representation, the propensity score model can be summarized by the sieve parameter $\psi_0 = (\beta_0', \delta_0')'$. However, this representation is over-parameterized and ψ cannot be identified without further restrictions, such as the two-quantile restrictions

$$H(u_1) = u_1 \qquad \text{and} \qquad H(u_2) = u_2$$

for $u_1, u_2 \in (0,1)$ with $u_1 \neq u_2$ (cf. Bierens, 2014, Section 2.2). For example, we let $u_1 = 0.25$ and $u_2 = 0.75$. We impose this assumption formally as Assumption 4.(iv) ahead. Define the parameter space by

$$\Psi = \left\{ \psi = (\beta', \delta')' \;\middle|\; \sum_{j=1}^{p} |\beta_j| + \sum_{j=1}^{\infty} j^2 |\delta_j| \leq M \right\},$$

for some M. We consider both the metric d on Ψ defined by $d(\psi_1, \psi_2) = \|\psi_1 - \psi_2\|_1 + \|\psi_1 - \psi_2\|_2$, and the metric $d_{(2)}$ induced by the norm $\|\psi\|_{(2)} = \sum_{j=1}^{\infty} j^2 |\psi_j|$.

For estimation of ψ_0, we consider the sieve space

$$\Psi_n = \{ \Xi_{\ell_n} \psi \mid \psi \in \Psi \},$$

where Ξ_{ℓ_n} denotes the projection on the first $\ell_n > p$ coordinates. Define the penalized log-likelihood

$$g(D, X; \psi) = D \ln(H(G(X'\beta); \delta)) + (1 - D) \ln(1 - H(G(X'\beta); \delta)) - \Pi(\delta),$$

where Π denotes a penalty function defined by

$$\Pi(\delta) = (u_1 - H(u_1; \delta))^4 + (u_2 - H(u_2; \delta))^4.$$

We define the constrained maximum likelihood sieve estimator $\hat{\psi}$ of ψ_0 by

$$\hat{\psi} = \arg\max_{\psi \in \Psi_n} \hat{Q}(\psi),$$

where $\hat{Q}(\psi) = n^{-1} \sum_{i=1}^{n} g(D_i, X_i; \psi)$. We also define its population counterpart by $Q(\psi) = \mathbb{E}[g(D, X; \psi)]$. With the sieve estimator $\hat{\psi} = (\hat{\beta}', \hat{\delta}')'$, we in turn estimate the propensity score $e(X_i)$ by

$$\hat{e}(X_i) = H(G(X'\hat{\beta}); \hat{\delta}).$$

We now collect some primitive assumptions in Assumption 4. For convenience of writing those tailored conditions, we introduce some notation. Let ∂_{ψ_j} denote the partial derivative with respect to the j-th coordinate ψ_j of ψ. For any $j, \ell \in \mathbb{N}$, let

$$B_{j,\ell}(\psi_0) = \begin{pmatrix} \mathbb{E}[\partial_{\psi_1}\partial_{\psi_1} g(D_i, X_i; \psi_0)] & \cdots & \mathbb{E}[\partial_{\psi_1}\partial_{\psi_\ell} g(D_i, X_i; \psi_0)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\partial_{\psi_j}\partial_{\psi_1} g(D_i, X_i; \psi_0)] & \cdots & \mathbb{E}[\partial_{\psi_j}\partial_{\psi_\ell} g(D_i, X_i; \psi_0)] \end{pmatrix}.$$

Let $K_{j,\ell}(\psi_0)$ and $L_\ell(\psi_0)$ denote a $j \times \ell$ orthogonal matrix and an $\ell \times \ell$ lower-triangular matrix, respectively, such that $\text{diag}(2^{-1}, 2^{-2}, ..., 2^{-j}) B_{j,\ell}(\psi_0) = K_{j,\ell}(\psi_0) L_\ell(\psi_0)$ obtained by the Gram-Schmidt orthogonalization procedure. Let $L_m^{(j,\ell)}(\psi_0)$ be the upper-left $m \times m$ block of $L_{j,\ell}(\psi_0)$ from this decomposition.

Assumption 4.

(i) (D_i, X_i) is i.i.d.

(ii) $\mathbb{E}[\|X_i\|^{2+\epsilon}] < \infty$ for some $\epsilon > 0$.

(iii) If $p = 1$, then the distribution of X has support $\mathbb{R}$ and $\beta_0 \neq 0$. If $p \geq 2$, then there exists j such that the conditional distribution of X_j given X_{-j} has support $\mathbb{R}$, $\beta_j \neq 0$ and $\mathbb{V}\text{ar}(X_i)$ is nonsingular.

(iv) $H_0 > 0$ on $(0, 1)$. H_0 is three-times differentiable with uniformly continuous derivatives h_0, h_0' and h_0'' on $[0, 1]$. h_0 is uniformly bounded. $h_0 > 0$ on $[0, 1]$. $H_0(u_1) = u_1$ and $H_0(u_2) = u_2$ for $u_1, u_2 \in (0, 1)$ with $u_1 \neq u_2$. $x \mapsto H(G(x'\beta_0); \delta_0)/G(x'\beta_0)$ is bounded away from zero.

(v) $\psi_{0,n} = (\beta_{\ell_n}', \psi_0')' \in \text{int}(\Psi)$ with respect to $d_{(2)}$. $\lim_{n\to\infty} \sqrt{n} \sum_{j=\ell_n+1}^{\infty} j^2 |\psi_{0j}| = 0$.
$\sum_{j=1}^{\infty} \sum_{\ell=1}^{\infty} (j\ell)^{-2-\tau} \mathbb{E}[|\partial_{\psi_j}\partial_{\psi_\ell} g(D_i, X_i; \psi_0)|] < \infty$ for some $\tau \geq 0$.
$\lim_{\epsilon \downarrow 0} \sum_{j=1}^{\infty} \sum_{\ell=1}^{\infty} (j\ell)^{-2-\tau} \mathbb{E}[\sup_{\|\psi - \psi_0\|_{(2)} \leq \epsilon} |\partial_{\psi_j}\partial_{\psi_\ell} g(D_i, X_i; \psi) - \partial_{\psi_j}\partial_{\psi_\ell} g(D_i, X_i; \psi_0)|] = 0$.
$\mathbb{E}[\partial_{\psi_j}\partial_{\psi_\ell} g(D_i, X_i; \psi_0)] \neq 0$ for at least one pair $j, \ell \in \mathbb{N}$. $\text{rank}(B_{j,j}) = j$ for each $j \geq p$.
$\liminf_{j\to\infty} \lim_{\ell\to\infty} \det(L_m^{(j,\ell)}(\psi_0)) > 0$.

Part (i) requires random sampling. Parts (ii) and (iii) impose very mild restrictions on the

distribution of X_i. Part (iii) specifies regularity conditions of the unknown function H_0. Part (iv) imposes restrictions on the true parameter ψ_0 and the rate of the sieve dimension. Among others, it is worth discussing the condition in part (v) that $x \mapsto H(G(x'\beta_0); \delta_0)/G(x'\beta_0)$ is bounded away from zero. It requires that the true link function F_0 behaves similarly to the logistic link function G near the tails, although F_0 can be arbitrarily nonparametric and distant from G in the middle range. Under this set of conditions, we have the property as stated in the following theorem, which provides a formal justification for Assumption 2.

Theorem 3. *If Assumption 4 is satisfied, then Assumption 2 holds such that*

$$\max_{1\leq i\leq n}\left|\frac{e(X_i)}{\hat{e}(X_i)}-1\right| = o_p(1).$$

5 Simulations

This section evaluates the finite-sample performance of the test (2.3) with both the Logit and the semiparametric design studied in the previous section. To be precise, we generate the propensity score according to $e(X_i) = H(G(X_i'\beta_0); \delta_0)$, where $\beta_0 = (1, 0, -1)'$, and we consider various designs for δ_0 as follows:

Design 1 : $\delta_0 = (0,\ \cdots)'$;

Design 2 : $\delta_0 = \left(\frac{0.1}{1^{2.5}},\ \frac{0.1}{2^{2.5}},\ \frac{0.1}{3^{2.5}},\ 0,\ \cdots\right)'$;

Design 3 : $\delta_0 = \left(\frac{0.1}{1^{2.5}},\ \frac{0.1}{2^{2.5}},\ \frac{0.1}{3^{2.5}},\ \frac{0.1}{4^{2.5}},\ \frac{0.1}{5^{2.5}},\ \frac{0.1}{6^{2.5}}\ ,0,\ \cdots\right)'$.

Note that Design 1 corresponds to the parametric Logic model. We generate a random sample $\{X_i\}_{i=1}^n$ by the mixture of $(X_{i1}, X_{i2}, X_{i3}) = (V_{i1}, V_{i2}, V_{i1} + V_{i3})$ with probability 0.5 and $(X_{i1}, X_{i2}, X_{i3}) = -(V_{i1}, V_{i2}, V_{i1} + V_{i3})$ with probability 0.5, where $V_{i1} \sim N(0, 1)$, $V_{i2} \sim N(0, 1)$ and $V_{i3} \sim Exp(1/\xi)$ independently. We vary ξ across simulations. Notice that $\mathbb{P}(X_i'\beta_0 < -x) = 0.5\exp(-x/\xi)$ under this data generating design, and thus $\xi \geq 1$ (respectively, < 1) implies the null (respectively, alternative) hypothesis.

Table 1 collects simulation results. Displayed are the simulated average of $\|\hat{\beta} - \beta_0\|$, simulated average of $\max_{1\leq i\leq n} |e_i/\hat{e}_i - 1|$, and simulated frequency of rejecting H_0. For each design, the simulated average of $\|\hat{\beta} - \beta_0\|^2$ decreases proportionally to n^{-1}, which is consistent with the root-n consistency of $\hat{\beta}$ for β_0. Furthermore, the simulated average of $\max_{1\leq i\leq n} |e_i/\hat{e}_i - 1|$ is decreasing with the sample size n, which is consistent with our result that the estimated propensity scores are uniformly consistent in the sense of Assumption 2. These patterns are borne out across all the simulation designs.

[TABLE 1 HERE]

In the following, we summarize the findings in our simulation study regarding the rejection frequency of the test (2.3). First, our proposed test controls size very well under the null hypothesis, which corresponds to the cases with $\xi = 2$ and $\xi = 1$. However, we do note that, when the number of propensity scores used (i.e., k) is large, the extreme value approximation becomes less accurate. Again, this is because with a large k one is effectively extrapolating using observations far away from the tail. Even in this case, however, we note that the size distortion is only moderate. Second, the rejection probability of the test increases for smaller ξ. To be very precise, a small $\xi < 1$ corresponds to an alternative that is easier to distinguish from the null hypothesis. Overall, our testing procedures performs very well both in terms of size control and statistical power.

6 Applications

To illustrate the empirical applicability of our method, we revisit two datasets which might be prone to the limited overlap issue. It turns out that we reject the null hypothesis in one of these two applications, suggesting either no or only mild degree of limited overlap in this case. For the other application, however, we fail to reject the null hypothesis.

Our first illustration employs a dataset from the National Supported Work (NSW) program, which was implemented in 1970s with the aim of providing work experience to eco-

nomically disadvantaged workers lacking job skills. Since LaLonde (1986), this data set has been analyzed by many studies (Dehejia and Wahba, 1999, 2002; Smith and Todd, 2005). We consider one particular subsample from Dehejia and Wahba (1999), which consists of 185 treated individuals in the NSW experimental group ($D_i = 1$), and a nonexperimental comparison group of 2,490 individuals from the Panel Study of Income Dynamics (PSID, $D_i = 0$). As a baseline specification, the propensity score is estimated parametrically using a Logit model and the following covariates: `educ` (years of education) and its square, `age` and its square, `earn74` and `earn75` (earnings in 1974 and 1975) and their squares, indicators for `married`, `black` and `hispanic`, and the interaction term `black`×`u74`, where `u74` indicates unemployed in 1974. We refer interested readers to the aforementioned studies for detained information on variable definition, sample inclusion, and other specifications of the propensity score. We also estimate the propensity score using the semiparametric method with sieve dimensions of $\delta = 3$, and 6. Histograms of the estimated propensity scores are depicted in Figure 2.

[FIGURE 2 HERE]

To conduct a formal diagnosis, we implement our proposed test (2.3) with various choices of k to both the left tail and right tail of the distribution of the estimated propensity score. In particular, the left tail entails testing if $\mathbb{E}[1/e(X)] = \infty$, and the right tail entails testing if $\mathbb{E}[1/(1 - e(X))] = \infty$. The p-values of the tests are presented in Table 2. The results are coherent with the histograms. In particular, we reject the infinite mean in the right tail, but we fail to do so in the left tail. Such a heavy left tail may jeopardize the root-n asymptotic normality of the classical average treatment effect estimator. Consequently, researchers may resort to alternative methods that are robust to a heavy tail (Ma and Wang, 2020; Sasaki and Ura, 2021) or change the parameter of interest (say, employ a trimming strategy or instead estimate the treatment effect on the treated).

[TABLE 2 HERE]

Our second application examines the Right Heart Catheterization (RHC) dataset studied by Connors et al. (1996), and subsequently by Hirano and Imbens (2001) and Crump et al. (2009). The goal is to analyze the effectiveness of RHC using data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) with the propensity score weighting method. As reported in these studies, the estimated propensity score almost span the entire interval $(0, 1)$. Crump, Hotz, Imbens, and Mitnik (2009) thus propose to trim observations near the tails when conducting inverse propensity score weighting.

Our data consists of 5,735 observations, among which individuals fall into the treatment group if RHC was applied within 24 hours of admission. In the baseline Logit specification of the propensity score, 72 covariates are included, covering demographic, medical and clinical attributes. Summary statistics of the 72 covariates can be found in Connors et al. (1996) and Hirano and Imbens (2001). In addition to the parametric specification (denoted by $\delta = 0$), we also estimate using the semiparametric approach with sieve dimensions of $\delta = 3$, and 6. Figure 3 collects the histograms of the estimated propensity scores.

[FIGURE 3 HERE]

We implement our proposed test (2.3) with various choices of k. The p-values are reported in Table 3. Except for the case with $k = 50$ and $\delta = 3$ and 6, the test always rejects the null hypothesis of heavy tail, and therefore we expect that the root-n asymptotic Gaussian approximation of the classical average treatment effect estimator is reliable for this application.

[TABLE 3 HERE]

7 Concluding remarks

We proposed a formal statistical test which can help assess the degree of limited overlap in observational studies. This test takes the largest/smallest estimated propensity scores

as input, and controls size uniformly over a large range of underlying propensity score tail distributions. Rejecting the null hypothesis indicates either no or very mild degree of limited overlap. We illustrated our method both in a simulation study and by revisiting two datasets widely studied in the literature.

Appendix

A Additional details on testing both small and large propensity scores

In Section 3 we mention that our test can be extended for the hypotheses $\mathsf{H}_0 : \mathbb{E}[1/(e_i(1-e_i))] = \infty$ vs. $\mathsf{H}_1 : \mathbb{E}[1/(e_i(1-e_i))] < \infty$. The following lemma establishes the equivalence between these hypotheses and those for testing whether at least one of $\mathbb{E}[1/e_i]$ and $\mathbb{E}[1/(1-e_i)]$ is infinite.

Lemma 1. *The following equivalence holds:*

$$\mathbb{E}\left[1/\left(e_i\left(1-e_i\right)\right)\right] = \infty \quad \Leftrightarrow \quad \mathbb{E}\left[1/e_i\right] = \infty \text{ and/or } \mathbb{E}\left[1/\left(1-e_i\right)\right] = \infty.$$

A proof of this lemma is found in the following appendix section. Given this result and the assumption that both right tails of $1/e_i$ and $1/(1-e_i)$ are regularly varying with tail indices, say ξ_1 and ξ_2, respectively, the hypotheses can be reformulated as

$$\mathsf{H}_0 : \xi_1 \in [1, \bar{\xi}\,] \text{ and/or } \xi_2 \in [1, \bar{\xi}\,] \text{ vs. } \mathsf{H}_1 : \xi_1 \in (0,1) \text{ and } \xi_2 \in (0,1)\,. \tag{A.1}$$

To test (A.1), we take the smallest k_1 order statistics of $\hat{e}_i$ and the smallest k_2 order statistics of $1-\hat{e}_i$ for some fixed k_1 and k_2. Since they are asymptotically independent, the joint density of the self-normalized statistics $\mathbf{T}$ for both the left and the right tails is simply

the product of their marginal densities, respectively characterized by ξ_1 and ξ_2 (e.g. Arnold, Balakrishnan, and Nagaraja, 2008, Chapter 8). Therefore, we can construct the likelihood ratio statistic in a similar fashion to (2.3), where the integrals are taken with respect to both ξ_1 and ξ_2.

B Proofs

In this section we present the proofs of Lemma 1 and Theorem 1 and 2. The proof of Theorem 3, however, is quite involved, and hence is left to the supplementary material. There we also provide additional details for Remarks 1 and 2.

B.1 Proof of Lemma 1

For the "only if" part, note that

$$\mathbb{E}\left[\frac{1}{e_i\left(1-e_i\right)}\right]=\mathbb{E}\left[\frac{1}{e_i\left(1-e_i\right)}\mathbb{1}(e_i\leq 0.5)\right]+\mathbb{E}\left[\frac{1}{e_i\left(1-e_i\right)}\mathbb{1}(e_i>0.5)\right].$$

So that if $\mathbb{E}\left[1/\left(e_i\left(1-e_i\right)\right)\right]=\infty$, then at least one of the terms on the right-hand side is also infinite. Say it is the first one. Then

$$\infty=\mathbb{E}\left[\frac{1}{e_i\left(1-e_i\right)}\mathbb{1}(e_i\leq 0.5)\right]\leq 2\mathbb{E}\left[\frac{1}{e_i}\mathbb{1}(e_i\leq 0.5)\right]\leq 2\mathbb{E}\left[\frac{1}{e_i}\right].$$

Therefore, $\mathbb{E}\left[1/e_i\right]$ is ∞.

For the "if" part, note that

$$\mathbb{E}\left[1/\left(e_i\left(1-e_i\right)\right)\right]\geq\mathbb{E}\left[1/e_i\right]\quad\text{and}\quad\mathbb{E}\left[1/\left(e_i\left(1-e_i\right)\right)\right]\geq\mathbb{E}\left[1/\left(1-e_i\right)\right].$$

Therefore, $\mathbb{E}\left[1/\left(e_i\left(1-e_i\right)\right)\right]$ is infinite as long as one of $\mathbb{E}\left[1/e_i\right]$ and $\mathbb{E}\left[1/\left(1-e_i\right)\right]$ is ∞.

B.2 Proof of Theorem 1

To simplify the presentation, we define $Y_i =: e(X_i)^{-1}$ and $\hat{Y}_i =: \hat{e}(X_i)^{-1}$. Accordingly we denote the jth largest order statistic of $\{Y_i\}$ as $Y_{(j)}$ and similarly for $\{\hat{Y}_i\}$.

To start, Assumption 1 and the standard extreme value theory imply that there exist sequences of constants a_n and b_n such that

$$\frac{(Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)})' - b_n}{a_n} \Rightarrow \mathbf{V}, \tag{B.1}$$

where $\mathbf{V}$ has the following density ,

$$f_{\mathbf{V}|\xi}(v_1, ..., v_k) = G_\xi(v_k) \prod_{i=1}^{k} g_\xi(v_i)/G_\xi(v_i) \text{ on } v_k \leq v_{k-1} \leq \ldots \leq v_1$$

with $G_\xi(v) = \exp(-(1+\xi v)^{-1/\xi})$ and $g_\xi(v) = \mathrm{d}G_\xi(v)/\mathrm{d}v$. Besides, Theorem 1.2.1 in de Haan and Ferreira (2007) implies that $a_n = O(n^\xi)$ and $b_n = O(1)$. Therefore $Y_{(j)} = O_p(a_n)$ for $j \in \{1, ..., k\}$. Let $I = (I_1, \ldots, I_k) \in \{1, \ldots, n\}^k$ be the k random indices such that $Y_{(j)} = Y_{I_j}$, $j = 1, \ldots, k$, and let $\hat{I}$ be the corresponding indices such that $\hat{Y}_{(j)} = \hat{Y}_{\hat{I}_j}$. Then the convergence of $\hat{Y}_{(j)}$ follows from (B.1) once we establish $|\hat{Y}_{\hat{I}_j} - Y_{I_j}| = o_p(a_n)$ for $j = 1, \ldots, k$. We consider $k = 1$ for simplicity and the argument for a general k is very similar. Define $\eta_i = \hat{Y}_i - Y_i$. Assumption 2 implies that

$$\begin{aligned}
\max_i |\eta_i| &= \max_i \left| \hat{Y}_i - Y_i \right| \\
&\leq Y_{(1)} \max_i \left| \frac{\hat{Y}_i}{Y_i} - 1 \right| \\
&= o_p(a_n).
\end{aligned}$$

Given this, we have that, on the one hand, $\hat{Y}_{\hat{I}} = \max_i\{Y_i + \eta_i\} \leq Y_I + \max_i |\eta_i| = Y_I + o_p(a_n)$; and on the other hand, $\hat{Y}_{\hat{I}} = \max_i\{Y_i + \eta_i\} \geq \max_i\{Y_i + \min_i\{\eta_i\}\} \geq Y_I + \min_i\{\eta_i\} \geq Y_I - \sup_i |\eta_i| = Z_I - o_p(a_n)$. Therefore, $|\hat{Y}_{\hat{I}} - Y_I| \leq o_p(1) = o_p(a_n)$ as desired.

Then by the continuous mapping theorem, we have

$$\mathbf{T} \Rightarrow \mathbf{V}^* =: \frac{\mathbf{V} - \mathbf{V}_k}{\mathbf{V}_1 - \mathbf{V}_k}.$$

The rest of the proof follows from the construction of the weights W_0 and cv_α. See Section Section C ahead. In particular, denote our test (2.3) as $\varphi(\cdot)$. Since the null space is compact and the density of $\mathbf{V}^*$ is continuous, for any positive measure W_0 on $[1, \bar{\xi}]$, one can select a large enough cv_α so that $\sup_{\xi \in [1,\bar{\xi}]} \mathbb{P}[\varphi(\mathbf{V}^*) = 1] \leq \alpha$ under the null hypothesis. The continuous mapping theorem and (B.1) imply that $\lim_{n\to\infty} \mathbb{P}[\varphi(\mathbf{T}) = 1] \leq \alpha$ under the null hypothesis.

B.3 Proof of Theorem 2

We start with the high-level assumption that $|\hat{\beta} - \beta_0| = O_p(1/\sqrt{n})$. Using a Taylor expansion, we have

$$\frac{G(X_i'\beta_0)}{G(X_i'\hat{\beta})} - 1 = \left| \frac{G(X_i'\beta_0)}{G(X_i'\tilde{\beta})^2} \frac{\partial G(X_i'\tilde{\beta})}{\partial \beta} (\hat{\beta} - \beta_0) \right|,$$

where $\tilde{\beta}$ lies on the line segment between β_0 and $\hat{\beta}$. Let c_1 be some constant, we further decompose the above based on two events

$$\begin{aligned}
&\frac{G(X_i'\beta_0)}{G(X_i'\hat{\beta})} - 1 \\
&= \left| \frac{G(X_i'\beta_0)}{G(X_i'\tilde{\beta})^2} \frac{\partial G(X_i'\tilde{\beta})}{\partial \beta} (\hat{\beta} - \beta_0) \right| \mathbf{1}_{\|\hat{\beta}-\beta_0\| \leq \frac{c_1}{\sqrt{n}}} + \left| \frac{G(X_i'\beta_0)}{G(X_i'\tilde{\beta})^2} \frac{\partial G(X_i'\tilde{\beta})}{\partial \beta} (\hat{\beta} - \beta_0) \right| \mathbf{1}_{\|\hat{\beta}-\beta_0\| > \frac{c_1}{\sqrt{n}}} \\
&\leq \sup_{\|\beta-\beta_0\| \leq \frac{c_1}{\sqrt{n}}} \left| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} (\hat{\beta} - \beta_0) \right| \mathbf{1}_{\|\hat{\beta}-\beta_0\| \leq \frac{c_1}{\sqrt{n}}} \\
&\qquad + \left| \frac{G(X_i'\beta_0)}{G(X_i'\tilde{\beta})^2} \frac{\partial G(X_i'\tilde{\beta})}{\partial \beta} (\hat{\beta} - \beta_0) \right| \mathbf{1}_{\|\hat{\beta}-\beta_0\| > \frac{c_1}{\sqrt{n}}} \\
&\leq \frac{c_1}{\sqrt{n}} \sup_{\|\beta-\beta_0\| \leq \frac{c_1}{\sqrt{n}}} \left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| \mathbf{1}_{\|\hat{\beta}-\beta_0\| \leq \frac{c_1}{\sqrt{n}}} + \left| \frac{G(X_i'\beta_0)}{G(X_i'\tilde{\beta})^2} \frac{\partial G(X_i'\tilde{\beta})}{\partial \beta} (\hat{\beta} - \beta_0) \right| \mathbf{1}_{\|\hat{\beta}-\beta_0\| > \frac{c_1}{\sqrt{n}}}.
\end{aligned}$$

Now let c_2 be another constant. The goal is to find some sequence c_n, such that

$$\lim_{c_2 \uparrow \infty} \limsup_{n \to \infty} \mathbb{P}\left[\max_{1 \leq i \leq n} \left| \frac{G(X_i'\beta_0)}{G(X_i'\hat{\beta})} - 1 \right| \geq c_2 c_n \right] = 0.$$

The above probability is bounded by

$$\mathbb{P}\left[\|\hat{\beta} - \beta_0\| > \frac{c_1}{\sqrt{n}} \right] + \mathbb{P}\left[\frac{c_1}{\sqrt{n}} \max_{1 \leq i \leq n} \sup_{\|\beta - \beta_0\| \leq \frac{c_1}{\sqrt{n}}} \left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| \geq c_2 c_n \right].$$

Because the first probability above does not depend on c_2, and can be made arbitrarily small with suitable choices of c_1, it suffices to select c_n such that

$$\max_{1 \leq i \leq n} \sup_{\|\beta - \beta_0\| \leq \frac{c}{\sqrt{n}}} \left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| = O_p\left(\sqrt{n} c_n\right)$$

holds for all c.

C Computational details

This appendix provides computational details about constructing the test (2.3). The input is $\mathbf{V}^* =: \lim_{n \to \infty} \mathbf{T}$, whose density f_ξ is in (3.2) and computed by Gaussian Quadrature. To construct the test (2.3), we first specify the weight W_1 to be the uniform distribution over (0,1) for simplicity of exposition. The weight W_1 reflects the importance attached by the econometrician to different alternatives, which can be easily changed. Then, it remains to determine a suitable candidate for the weight W_0 and the critical value cv_α. This is achieved by employing the generic algorithm provided by Elliott, Müller, and Watson (2015) and Sasaki and Wang (2020, 2021). The idea is as follows.

First, we can discretize the null space $[1, \bar{\xi}]$ into a grid Ξ_a and determine W_0 accordingly as the point masses. To this end, we let $\tilde{W}_0 = \text{cv}_\alpha W_0$ to subsume the critical value. Denote $\varphi_{\tilde{W}_0}(\cdot)$ as the test (2.3) to emphasize the effect of $\tilde{W}_0$. Simulate N random draws of $\mathbf{V}^*$ from

$\xi \in \Xi_a$ and estimate the rejection probability under each value of ξ, denoted as $\mathbb{P}_\xi(\varphi_{\tilde{W}_0}(\mathbf{V}^*) = 1)$ by sample fractions. By iteratively increasing or decreasing the point masses as a function of whether the estimated $\mathbb{P}_\xi(\varphi_{\tilde{W}_0}(\mathbf{V}^*) = 1)$ is larger or smaller than the nominal level, we can always find a candidate $\tilde{W}_0$ that ensures the uniform size control on Ξ_a. This is because we allow $\mathbb{P}_\xi(\varphi_{\tilde{W}_0}(\mathbf{V}_*) = 1) < \alpha$ for some $\xi \in \Xi_a$. Once $\tilde{W}_0$ is obtained, we can numerically check if the test controls size on a finer grid than Ξ_a. If not, we repeat the algorithm based on such finer grid.

In practice, we can determine the point masses by the following concrete steps.

Algorithm:

1. Simulate $N = 10{,}000$ i.i.d. random draws from some proposal density with ξ drawn uniformly from Ξ_a, which is an equally spaced grid on $[1, 2]$ with 50 points.

2. Start with $\tilde{W}_0^{(0)} = \{1/50, 1/50, \ldots, 1/50\}'$ and $\text{cv}_\alpha = 1$. Calculate the (estimated) rejection probabilities $P_j =: \mathbb{P}_{\xi_j}(\varphi_{\tilde{W}_0^{(0)}}(\mathbf{V}^*) = 1)$ for every $\xi_j \in \Xi_a$ using importance sampling. Denote them by $P = (P_1, ..., P_{50})'$.

3. Update $\tilde{W}_0$ by setting $\tilde{W}_0^{(s+1)} = \tilde{W}_0^{(s)} + \eta(P - 0.05)$ with some step-length constant $\eta > 0$, so that the j-th point mass in $\tilde{W}_0$ is increased/decreased if the coverage probability for ξ_j is larger/smaller than the nominal level.

4. Keep the integration for 500 times. Then, the resulting $\tilde{W}_0^{(500)}$ is a valid candidate.

5. Numerically check if $\varphi_{\tilde{W}_0^{(500)}}$ indeed controls the size uniformly by simulating the rejection probabilities over a much finer grid on Ξ. If not, go back to step 2 with a finer Ξ_a.

The above algorithm takes a few seconds to run in a modern PC. The most time-consuming part is the calculation of the density f_ξ by Gaussian Quadrature. After conducting this algorithm, we find that W_0 always allocates all the weight to the single point $\xi = 1$. Accordingly, we present the logarithm of the critical values in Table 4. This finding

indicates that the least favorable distribution is indeed the point mass at $\xi = 1$. However, a theoretical justification eludes us due to the complicated expression of the density.

[TABLE 4 HERE]

D The Choice of k

The optimal choice of k has been a challenging question even without generated variables in the literature on extreme value theory. It is more difficult in our setup with generated variables since the propensity scores are estimated. In specific, Assumption 1 only characterizes the first-order approximation of the largest/smallest order statistics while a data-driven choice of k requires the knowledge of the higher-order approximation. This is close in spirit to the optimal choice of bandwidth in kernel regressions, which typically requires higher-order derivatives either by assumption or further estimation. Therefore, we would recommend practitioners to report our testing results with a range of plausible values of k for sensitivity analysis.

Next, we provide some rule-of-thumb guide for choosing a starting candidate of k under additional second-order conditions. We focus on the left tail of the score (equivalently the right tail of $1/e(X)$) for example. In addition to Assumption 1, we assume that the distribution of $1/e(X)$ satisfies that

$$1 - F_{e(X)^{-1}}(v) = c_1 v^{-1/\xi} + c_2 v^{-1/\xi+\rho/\xi}(1 + o(1)), \text{ as } v \to \infty,$$

for constants $c_1 > 0$, $c_2 \neq 0$, $\xi > 0$, and $\rho < 0$. This condition is sufficient for the regularly varying condition and also satisfied by many commonly used distributions, including, for example, Pareto, Student-t, F distributions. The second-order parameter ρ governs the approximation error from using the Pareto distribution for the true distribution. For the Student-t distribution with ν degrees of freedom, we have $\rho = -2\xi$ and $\xi = 1/\nu$. See de Haan and Ferreira (2007, Chapter 3) for more examples and review.

Based on the above condition, researchers have developed several data-driven choices of k for estimating the tail index ξ from the perspective of minimizing the asymptotic mean squared error. The second-order parameter ρ again plays a key role in the finite sample performance of these choices. Its value is unknown *a priori* and very challenging to estimate. As a rule-of-thumb guide, we may consider the Student-t distribution as the benchmark and use $\rho = -2\xi$. Then the optimal choice of k is $n^{4\xi/(1+4\xi)}$ (up to some constant). See Remark 3.2.7 in de Haan and Ferreira (2007) for further details. In practice, we may substitute some initial estimate (e.g. Hill, 1975) for ξ.

References

Abadie, A. and M. D. Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*, 465–503.

Andrews, D. W. K. and W. Ploberger (1995). Admissibility of the likelihood ratio test when a nuisance parameter is present only under the alternative. *Annals of Statistics 23*, 1609–1629.

Armstrong, T. and M. Kolesár (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica 89*(3), 1141–1177.

Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja (2008). *A first course in order statistics*. Society for Industrial and Applied Mathematics.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2021). High-dimensional econometrics and regularized GMM. *Handbook of Econometrics*, forthcoming.

Bierens, H. J. (2014). Consistency and asymptotic normality of sieve ml estimators under low-level conditions. *Econometric Theory*, 1021–1077.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–154.

Chaudhuri, S. and J. B. Hill (2014). Heavy tail robust estimation and inference for average treatment effects. Technical report.

Connors, A. F., T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus (1996). The effectiveness of right heart catheterization in the initial care of critically III patients. *Jama 276*(11), 889–897.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

de Haan, L. and A. Ferreira (2007). *Extreme Value Theory: An Introduction*. New York: Springer.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluations of training programs. *Journal of the American Statistical Association 94*, 1053–1062.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics 84*, 151–161.

Elliott, G., U. K. Müller, and M. W. Watson (2015). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica 83*, 771–811.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Heiler, P. and E. Kazak (2021). Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics 222*(2), 1083–1108.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics 3*(5), 1163–1174.

Hirano, K. and G. W. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology 2*(3-4), 259–278.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hong, H., M. P. Leung, and J. Li (2020). Inference on finite-population treatment effects under limited overlap. *Econometrics Journal 23*(1), 32–47.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Khan, S. and D. Nekipelov (2011). On uniform inference in nonlinear models with endogeneity. *Journal of Econometrics*, forthcoming.

Khan, S. and E. Tamer (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica 78*(6), 2021–2042.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 76*, 604–620.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypothesis*. New York: Springer.

Ma, X. and J. Wang (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association 115*, 1851–1860.

Müller, U. K. and Y. Wang (2017). Fixed-k asymptotic inference about tail properties. *Journal of the American Statistical Association 112*, 1134–1143.

Newey, W. K. and D. L. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics, Volume IV*, pp. 2111–2245. Elsevier Science B.V.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association 84*(408), 1024–1032.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica 85*(2), 645–660.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine 127*(8_Part_2), 757–763.

Sasaki, Y. and T. Ura (2021). Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory*, forthcoming.

Sasaki, Y. and Y. Wang (2020). Diagnostic testing of finite moment conditions for the consistency and root-n asymptotic normality of the gmm and m estimators. *arXiv preprint arXiv:2006.02541*.

Sasaki, Y. and Y. Wang (2021). Fixed-k inference for conditional extremal quantiles. *Journal of Business & Economic Statistics*. forthcoming.

Smith, J. A. and P. E. Todd (2005). Does matching overcome lalonde's critique of nonexperimental estimators. *Journal of Econometrics 125*, 305–353.

SUPPLEMENTARY MATERIAL

Supplementary material: This article includes a supplementary material, which contains additional proofs and details.

					Frequency of Rejecting H_0			
ξ	δ_0	n	$\|\hat{\beta}-\beta_0\|^2$	$\max_i \|e_i/\hat{e}_i-1\|$	$k=25$	50	75	100
2.0	Design 1	2000	0.042	7.973	0.002	0.001	0.000	0.000
		4000	0.021	2.756	0.003	0.000	0.000	0.000
		8000	0.010	1.557	0.001	0.000	0.000	0.000
	2	2000	0.031	5.539	0.002	0.001	0.000	0.000
		4000	0.015	2.025	0.003	0.000	0.000	0.000
		8000	0.007	1.321	0.001	0.000	0.000	0.000
	3	2000	0.032	5.247	0.002	0.001	0.000	0.000
		4000	0.015	1.856	0.003	0.000	0.000	0.000
		8000	0.008	1.197	0.001	0.000	0.000	0.000
1.0	Design 1	2000	0.059	1.651	0.055	0.066	0.077	0.090
		4000	0.034	1.173	0.044	0.059	0.069	0.073
		8000	0.018	0.810	0.048	0.043	0.054	0.055
	2	2000	0.042	1.450	0.057	0.062	0.078	0.088
		4000	0.021	1.070	0.045	0.058	0.068	0.069
		8000	0.011	0.781	0.048	0.046	0.053	0.055
	3	2000	0.042	1.392	0.058	0.064	0.083	0.090
		4000	0.021	1.020	0.045	0.062	0.071	0.073
		8000	0.011	0.740	0.050	0.046	0.056	0.058
0.67	Design 1	2000	0.083	1.243	0.156	0.281	0.397	0.496
		4000	0.052	0.937	0.145	0.266	0.408	0.494
		8000	0.030	0.699	0.146	0.257	0.397	0.507
	2	2000	0.058	1.124	0.162	0.289	0.429	0.510
		4000	0.030	0.841	0.157	0.282	0.426	0.523
		8000	0.016	0.644	0.153	0.260	0.423	0.540
	3	2000	0.057	1.108	0.168	0.293	0.435	0.519
		4000	0.030	0.821	0.161	0.286	0.433	0.535
		8000	0.016	0.621	0.155	0.267	0.429	0.550

Table 1: Simulated averages of $\|\hat{\beta}-\beta_0\|^2$ and $\max_i |e_i/\hat{e}_i-1|$, and rejection frequency. The results are based on 2,000 Monte Carlo repetitions.

	δ	$k = 25$	$k = 50$	$k = 75$	$k = 100$	$k = 125$	$k = 150$
Left tail	0	0.96	1.00	1.00	1.00	1.00	1.00
	3	0.71	0.52	0.64	0.89	1.00	1.00
	6	1.00	1.00	1.00	1.00	1.00	1.00
Right tail	0	0.40	0.30	0.00	0.00	0.00	0.00
	3	0.04	0.00	0.00	0.00	0.00	0.00
	6	0.03	0.00	0.00	0.00	0.00	0.00

Table 2: P-values of the fixed-k test in the NSW illustration. Left tail: testing if $\mathbb{E}[1/e(X)] = \infty$; Right tail: testing if $\mathbb{E}[1/(1 - e(X))] = \infty$. Rows with $\delta = 0$ correspond to our baseline parametric estimate with a Logit specification.

	δ	$k = 25$	$k = 50$	$k = 75$	$k = 100$	$k = 125$	$k = 150$
Left tail	0	0.27	0.00	0.00	0.00	0.00	0.00
	3	0.32	0.16	0.00	0.00	0.00	0.00
	6	0.35	0.16	0.00	0.00	0.00	0.00
Right tail	0	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: P-values of the fixed-k test in the RHS illustration. Left tail: testing if $\mathbb{E}[1/e(X)] = \infty$; Right tail: testing if $\mathbb{E}[1/(1 - e(X))] = \infty$. Rows with $\delta = 0$ correspond to our baseline parametric estimate with a Logit specification.

k	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	k	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5	0.426	0.490	0.567	80	0.074	0.649	2.004
10	0.711	0.926	1.256	85	0.041	0.574	1.998
15	0.806	1.159	1.707	90	0.028	0.593	1.880
20	0.801	1.229	2.037	95	-0.011	0.547	1.938
25	0.685	1.186	2.216	100	-0.031	0.539	1.780
30	0.611	1.184	2.311	105	-0.063	0.540	1.793
35	0.533	1.076	2.383	110	-0.073	0.515	2.009
40	0.432	1.037	2.323	115	-0.082	0.494	1.977
45	0.377	1.004	2.340	120	-0.093	0.466	1.907
50	0.332	0.917	2.326	125	-0.108	0.454	1.791
55	0.308	0.901	2.256	130	-0.128	0.417	1.793
60	0.264	0.826	2.109	135	-0.137	0.425	1.806
65	0.205	0.785	2.077	140	-0.168	0.406	1.695
70	0.180	0.719	2.095	145	-0.183	0.371	1.713
75	0.140	0.686	2.015	150	-0.187	0.353	1.685

Table 4: Logarithm of the critical values of the test (2.3). Based on 10,000 simulations.

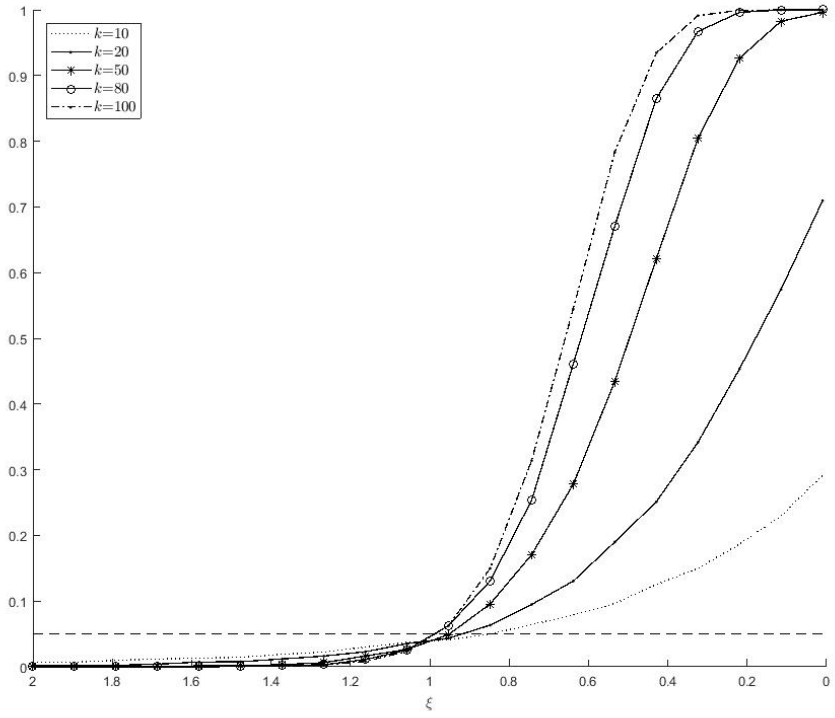

Figure 1: Asymptotic rejection probabilities of the test in (2.3), based on 10,000 simulation draws from (3.2) with $\xi \in (0, 2]$.

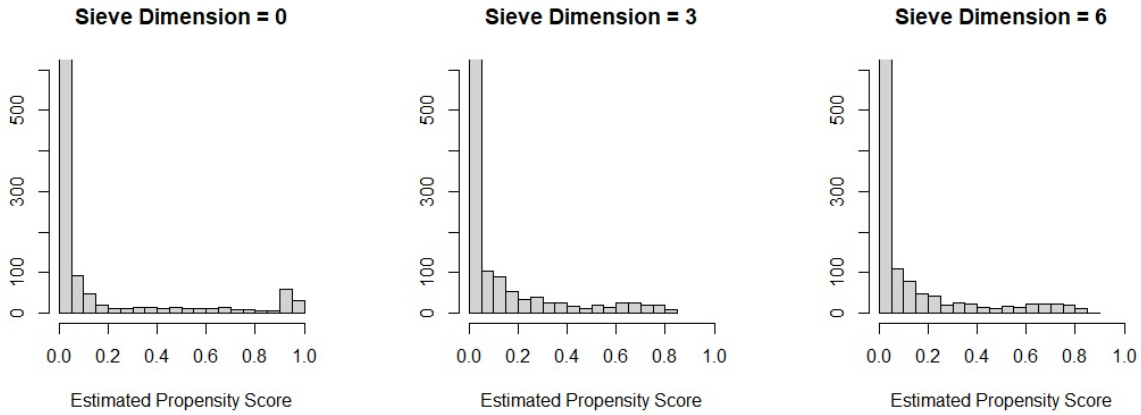

Figure 2: Histograms of estimated propensity scores in the NSW illustration. The case $\delta = 0$ corresponds to our baseline parametric estimate with a Logit specification.

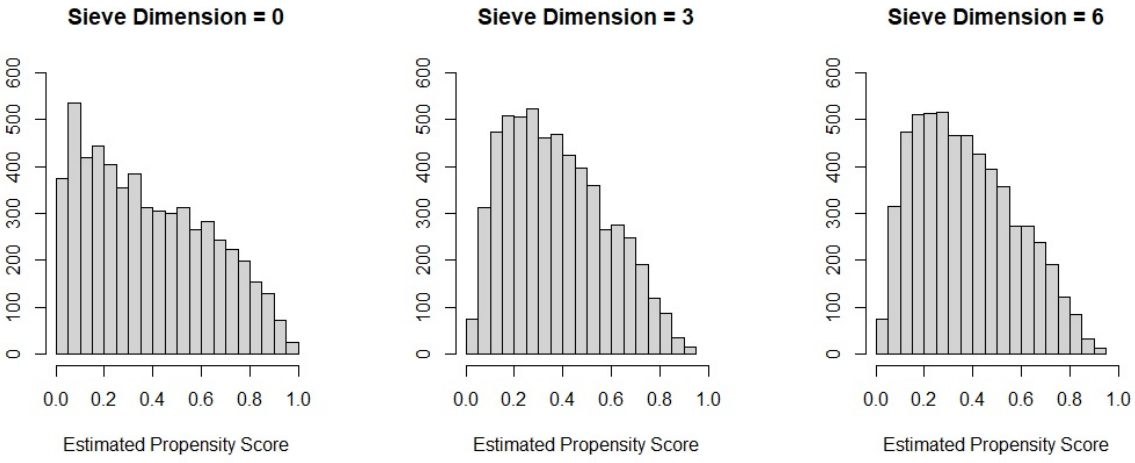

Figure 3: Histograms of estimated propensity scores in the RHC illustration. The case $\delta = 0$ corresponds to our baseline parametric estimate with a Logit specification.