Econometric Theory, **41**, 2025, 1129–1162.
doi:10.1017/S0266466624000161

TESTING LIMITED OVERLAP

XINWEI MA
University of California, San Diego

YUYA SASAKI
Vanderbilt University

YULONG WANG
Syracuse University

Extreme propensity scores arise in observational studies when treated and control units have very different characteristics. This is commonly referred to as *limited overlap*. In this paper, we propose a formal statistical test that 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.

1. INTRODUCTION

Treatment take-up is usually heterogeneous in observational studies, as the decision to participate in a program may depend on the unit's demographic or socioeconomic characteristics. The presence of treatment heterogeneity renders standard procedures such as mean comparison invalid for estimating causal effects. Following the seminal work by Rubin (1974, 1997), Rosenbaum and Rubin (1983), and Rosenbaum (1989), there is extensive literature on causal effect estimation using inverse probability weighting, where each unit is inversely weighted by its probability of receiving treatment (a.k.a. the propensity score). A key assumption for estimating the average treatment effect is the *strong overlap*, which requires that the propensity score 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).

We would like to express our sincere gratitude to Peter C. B. Phillips and three anonymous reviewers for their invaluable comments and insights, which greatly enhanced the quality of this paper. Sasaki thanks Brian and Charlotte Grove, Chair for research support. Address correspondence to Yuya Sasaki, Department of Economics, Vanderbilt University, 415 Calhoun Hall, Nashville, TN 37240, USA; e-mail: yuya.sasaki@vanderbilt.edu.

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

A common empirical strategy is to restrict the target population by trimming observations in the region of poor overlap, so that the subpopulation obtained after trimming only exhibits a mild level of treatment heterogeneity and the overlap assumption is satisfied (Crump et al., 2009; Chaudhuri and Hill, 2014; Ma and Wang, 2020; Sasaki and Ura, 2022).

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 (Cattaneo, Jansson, and Ma, 2020, 2024). It is also a common practice to report summary statistics (of key covariates) separately for the treatment group and the control group as an illustration of treatment heterogeneity. To the best of our 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 the 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 et al., 2018; Farrell, Liang, and Misra, 2021) and (ii) test whether a subpopulation obtained after trimming satisfies the overlap condition for estimating subpopulation average treatment effects.

2. OVERVIEW

We first present an overview of the proposed statistical test. Assume that there is a random sample of size n consisting of (X_i, D_i), $i = 1, 2, \ldots, n$, 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, $\mathbb{P}[D_i = 1|X_i] = e(X_i) =: e_i$. If treated and control units have quite different characteristics, the propensity score may take extreme values (i.e., being close to zero or one).

To test, for example, the presence of small propensity scores, our procedure concerns the following competing hypotheses:

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

One distinguishing feature of our test is that rejecting the null indicates either no or a very mild degree of limited overlap.[1] 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.

[1] In the literature, limited overlap refers to the scenario where the propensity score can be arbitrarily close to 0 or 1 (in contrast, strong overlap requires the propensity score to be bounded away from 0 and 1). However, limited overlap does not necessarily lead to an infinite expectation of the inverse propensity score, as long as the tail is thin enough. Thus, the alternative hypothesis in (2.1) also allows for a mild degree of limited overlap.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

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

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

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

Testing the hypotheses in (2.1) boils down to investigating the tail properties of the propensity score. Specifically, suppose that the distribution of the propensity score has a regularly varying tail at zero with tail index ξ.[2] In this case, the hypotheses in (2.1) can be rewritten as[3]

$$\mathsf{H}_0 : \xi \in [1, \bar{\xi}\,] \quad \text{against} \quad \mathsf{H}_1 : \xi \in (0, 1)\,. \tag{2.2}$$

To test H_0, we employ the following self-normalized statistic:

$$\mathbf{T} = \frac{1}{e_{(1)}^{-1} - e_{(k)}^{-1}}\Big(e_{(1)}^{-1} - e_{(k)}^{-1},\ e_{(2)}^{-1} - e_{(k)}^{-1},\ e_{(3)}^{-1} - e_{(k)}^{-1},\ \ldots,\ e_{(k-1)}^{-1} - e_{(k)}^{-1},\ 0\Big)', \tag{2.3}$$

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

$$\text{reject } \mathsf{H}_0 \quad \text{if} \quad \Big(\int_0^1 f_\xi(\mathbf{T})\,\mathrm{d}W_1(\xi)\Big)\Big/\Big(\int_1^{\bar{\xi}} f_\xi(\mathbf{T})\,\mathrm{d}W_0(\xi)\Big) > \mathrm{cv}_\alpha, \tag{2.4}$$

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). Alternatively, it is possible to construct the test statistic by taking the

[2]Loosely speaking, this requires that the distribution of the propensity score admits a "polynomial tail": $\mathbb{P}[e_i \le t] \propto t^{1/\xi}$ as $t \to 0$ (up to additional slowly varying factors). Accordingly, the density $f_{e^{-1}}$ of e_i^{-1} satisfies that $f_{e^{-1}}(t) \propto t^{-1-1/\xi}$ as $t \to \infty$ and hence $\mathbb{E}[1/e_i^r] = \infty$ for $r \ge 1/\xi$. See Assumption 1 in Section 3 for a formal definition.

[3]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. See Section 3 for more detailed discussions.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

maximum over $\xi \in (0, 1)$ (the alternative space) and $\xi \geq 1$ (the null space) in the numerator and the denominator, respectively. Numerical evidence (not reported here) suggests that this strategy yields similar size and power properties across our simulation designs.

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 (3.1) the limiting density $f_\xi(\cdot)$ for any fixed k, and hence the generalized likelihood ratio approach in (2.4) 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 a fixed k has been similarly explored in Müller and Wang (2017) for inference about extreme quantiles and tail conditional expectations,[4] and in Sasaki and Wang (2023) for testing the bounded moment conditions in extremum estimation and generalized method of moments.

To operationalize the generalized likelihood approach in (2.3) and (2.4), one usually needs to estimate the propensity score in the 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: $\max_{1 \leq i \leq n} |e_i/\hat{e}_i - 1| \to 0$ in probability (Assumption 2 below). Given that this is a nonstandard assumption (i.e., it is not implied by extremum estimation results, such as those in 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 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.

[4]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.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Assumption 1. The distribution of e_i satisfies

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

for some tail index ξ.

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

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

where $F_{e(X)^{-1}}(\cdot)$ denotes the distribution function of the inverse propensity score. Then, it follows that $\mathbb{E}[1/e(X_i)]$ is finite/infinite if ξ is below/above 1, and hence the hypotheses in (2.1) can be rewritten as (2.2) under Assumption 1. The boundary case with $\xi = 1$ corresponds to the Cauchy distribution whose density decays at the polynomial rate $t^{-1-1/\xi}$ with $\xi = 1$. See the discussion in Footnote 2. 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. In addition, it allows for slowly varying components in the tail (see equations (8.5) and (8.6) in Feller, 1991 for additional discussions).

We also note that Assumption 1 will relate to tail restrictions on the covariates through functional form specifications on the propensity score model. Consider the standard logit model for example. A regularly varying tail with $\xi > 0$ for the propensity score is equivalent to an exponential-type tail for some index $X_i'\beta_0$. When X_i has a thinner tail, such as sub-Gaussian, or when X_i has a bounded support, simulation evidence suggests that our test rejects the null hypothesis with nontrivial power, which is informative and provides statistical evidence suggesting either no or very mild degree of limited overlap.

To test the hypotheses in (2.2), 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 an 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, $e_{(1)} \le e_{(2)} \le \cdots \le e_{(k)}$. By the extreme value theory (e.g., de Haan and Ferreira, 2007, Chap. 1), Assumption 1 implies that there exist sequences of constants a_n and b_n such that[5]

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

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

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

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

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

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 (2.2). 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 (2.3). 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, Chap. 6). In other words, the test statistic as a function of $\mathbf{T}$ remains unchanged if the data are shifted and multiplied by any nonzero constant. Such invariance property is desirable for our purposes as tail features should preserve no matter how the data are linearly transformed.

It follows from the continuous mapping theorem that

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

whose density function becomes

$$f_\xi(\mathbf{t}) = \Gamma(k) \int_0^\infty u^{k-2} \exp\left(-(1+1/\xi)\left(\sum_{i=2}^{k-1} \log(1+\xi t_i u) + \log(1+\xi u)\right)\right) \mathrm{d}u, \tag{3.1}$$

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 (3.1), the hypotheses in (2.1) and (2.2) can be tested by the generalized likelihood ratio statistic in (2.4). 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. It affects the rejection probability with data generated from different values of ξ under the alternative hypothesis. Typical choices in the existing literature are the exponential weight and the uniform weight (e.g., Andrews and Ploberger, 1994). We employ the uniform weight in our simulation study and the empirical applications, which can be easily modified. See Figures 1 and 2 and the discussion thereof on the power of our test.

In contrast, the weight, $W_0(\cdot)$, cannot be arbitrarily chosen and instead needs to be defined so as to guarantee size control. Since the null space is compact and

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

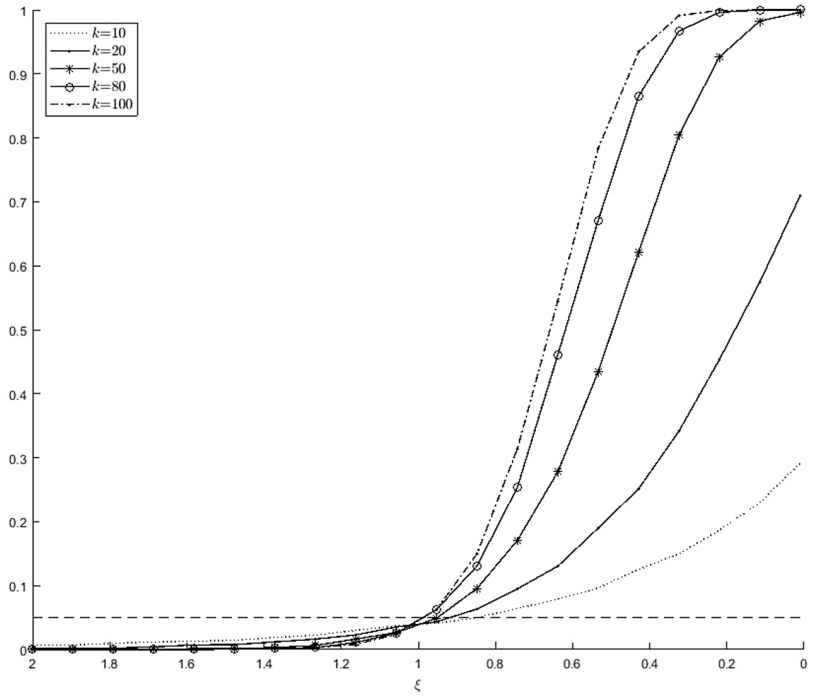


FIGURE 1. Asymptotic power curves of the test in (2.4). The curves are constructed numerically based on 10,000 simulation draws from (3.1) with $\xi \in (0, 2]$.

the density is continuous, we can naively set W_0 to be a point mass at any $\xi \in \mathsf{H}_0$ and adjust the critical value accordingly. A more sophisticated method is to use the least favorable distribution. In particular, given W_1 and k, W_0 can be understood as the least favorable distribution (e.g., Lehmann and Romano, 2005, Chap. 3), which are defined on the null space $\left[1, \bar{\xi}\right]$ to maintain the asymptotic size control that

$$\lim_{n\to\infty} \mathbb{P}[\text{Reject } \mathsf{H}_0] \leq \alpha \tag{3.2}$$

for any fixed $\xi \in \left[1, \bar{\xi}\right]$. However, the least favorable distribution does not necessarily exist. We instead resort to Elliott, Müller, and Watson (2015) who propose an *approximate* least favorable distribution (ALFD) that always exists, and provide a numerical algorithm to construct ALFD. We adopt their algorithm to our setup and determine $W_0(\cdot)$ and the critical value cv_α. See also Müller and Wang (2017). The output of this numerical algorithm is defined as W_0, which turns out to be the point mass allocated solely on $\xi = 1$ in our simulations, and the critical value, which is the 95% quantile of the simulated test statistics. Across various values of

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

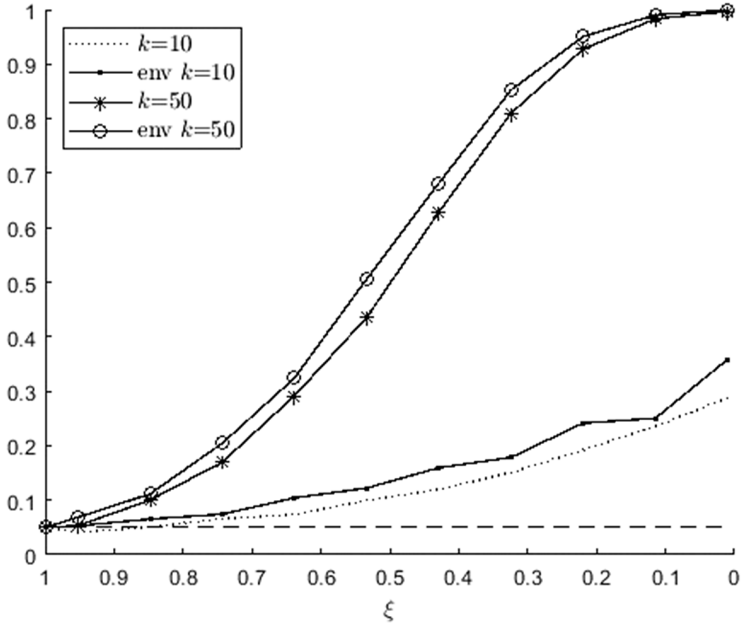


FIGURE 2. Asymptotic power curves and envelopes of the test in (2.4). The curves are constructed numerically based on 10,000 simulation draws from (3.1) with $\xi \in (0, 1]$. The lines labelled with "env" depict the rejection probabilities of the infeasible likelihood ratio test based on (2.4) that allocates all weights of W_0 on $\xi = 1$ and all weights of W_1 on one particular value of ξ in the alternative space.

k and α, $W_0(\cdot)$ and cv_α only need to be computed once. We provide the MATLAB algorithms and tabulate the simulated values of cv_α in Table C1.

Figure 1 presents the theoretical asymptotic (as $n \to \infty$ for fixed k) power curves of the test (2.4) based on (3.1) for each of $k = 10$, 20, 50, 80, and 100. Since there is no analytic expression for the theoretical power, we numerically construct these curves.[6] 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. We also observe that the power of the test increases with k.

Figure 2 compares the asymptotic power of the test (2.4) with the power envelope. To be more clear, the lines labeled with "env" depict the rejection probabilities of the infeasible likelihood ratio test based on (2.4) that allocates all weights of W_0 to $\xi = 1$ and all weights of W_1 to one particular value of ξ in the alternative space. We present the results with $k = 10$ and 50. The power of our test

[6]The numerical construction is based on 10,000 random draws simulated from (3.1) with $\xi \in [0.01, 2]$ for each k.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

is quite close to the envelope (the difference is less than 0.1), suggesting that the choice of W_1 (and $\bar{\xi}$) does not make a substantial difference.

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 predetermined k as long as n is sufficiently large. We recommend practitioners to report results with a variety of k as sensitivity analysis, as we will do in Section 7. Additional discussions and a rule-of-thumb choice are given in Section 5 to conserve space.

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 (2.3). Fortunately, $\hat{\mathbf{T}}$ will have the same asymptotic distribution as $\mathbf{T}$ under the following high-level uniform consistency assumption.

Assumption 2.

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

Assumption 2 requires that the estimation error of the propensity score is asymptotically dominated in magnitude by the true large order statistics of $1/e_i$. We are unaware of generic high-level conditions that can be employed to verify this assumption, and we therefore discuss two commonly used propensity score models, the Logit and the Probit. Encouragingly, as we will show in Section 4, Assumption 2 will hold under mild conditions (such as finite moments of the covariates) for Logit and Probit propensity score models.

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 that Assumptions 1 and 2 are satisfied. Then (3.2) holds for any fixed k.*

We close this section by discussing some features of the test in (2.4). 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 $k_n/n \to 0$. As a result, our test controls size 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).

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

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 $\mathsf{H}_0 : \mathbb{E}[1/(e_i(1-e_i))] = \infty$ versus $\mathsf{H}_1 : \mathbb{E}[1/(e_i(1-e_i))] < \infty$. 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 A.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 et al., 2008, Chap. 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 that the propensity score is more reliably 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, Lem. 1)

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

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 $\bar{\varsigma}$ denotes the upper bound of the parameter space. As before, rejection of the null hypothesis will imply either no or a 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. For completeness, we conduct such an analysis in Appendix D.

4. ESTIMATION OF THE PROPENSITY SCORE

This section justifies Assumption 2 in several commonly used models for 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.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

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

$$\hat{\beta} = \operatorname*{argmax}_{\beta} \sum_i D_i \ln\left(G(X_i'\beta)\right) + (1 - D_i)\ln\left(1 - G(X_i'\beta)\right), \tag{4.2}$$

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 (say, $\|\hat{\beta} - \beta_0\| = o_p(1)$), such a result generally 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 primitive sufficient conditions 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 the remarks below that it holds in both Logit and Probit models under very mild moment conditions on the covariates.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

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

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

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 thus consider the Logit and Probit models.

Remark 1. Assume that 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, Thm. 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

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

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

with δ_{0j} and δ_j denoting the jth element of δ_0 and δ, respectively. With this representation, the propensity score model can be summarized by the sieve parameter $\psi_0 = (\beta_0', \delta_0')'$. However, this representation is overparameterized and ψ cannot be identified without further restrictions, such as the two-quantile restrictions

$$H_0(u_1) = u_1 \qquad \text{and} \qquad H_0(u_2) = u_2$$

for $u_1, u_2 \in (0,1)$ with $u_1 \neq u_2$ (cf. Bierens, 2014, Sect. 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 the estimation of ψ_0, we consider the sieve space

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

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 the convenience of writing those tailored conditions, we introduce some notation. Let ∂_{ψ_j} denote the partial derivative with respect to the jth coordinate ψ_j of ψ. For any $j, \ell \in \mathbb{N}$, let

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

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Let $K_{j,\ell}(\psi_0)$ and $L_{j,\ell}(\psi_0)$ 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. We remark that these definitions implicitly require $j \geq \ell$. 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) $\psi_{0,n} = \Xi_{\ell_n}\psi_0 \in \text{int}(\Psi)$ 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 (iv) specifies regularity conditions of the unknown function H_0. Part (v) 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 (iv) 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.*

Finally, we remark that other non/semiparametric propensity score estimators can be considered for our test on limited overlap, provided that Assumption 2 is satisfied. The primary challenge lies in satisfying the uniform consistency assumption, which necessitates precise estimation of the propensity scores, even in the tail region. Our preliminary theoretical work (not reported here to conserve space) suggests that the sieve Logit model could serve as a viable alternative. The logistic link function in this model helps constrain the tail behavior of the estimated propensity scores. On the contrary, it would be considerably more challenging to theoretically justify the use of nonparametric estimators based on local smoothing, as they may exhibit volatile behavior in tail regions where observations are scarce.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

5. CHOICE OF K

The optimal choice of k has been a challenging question in the literature on extreme value theory. It is more difficult in our setup as the propensity scores are estimated. In particular, 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 (cf. Müller and Wang, 2017). (As an analogy, the optimal choice of bandwidth in kernel regressions typically requires higher-order derivatives either by assumption or additional estimation.)

The choice of k faces a bias-variance trade-off. On the one hand, a large k means that we treat a larger proportion of data as stemming from the tail, which leads to more bias. On the other hand, the estimated tail index may not be very precise if only a few extreme values are used (i.e., when k is small). Such trade-off typically incurs a delicate choice of k in finite samples, especially when the sample size is only moderate. Our test is built on the fixed-k framework, which essentially chooses a small k that is even negligible as $n \to \infty$. Such a choice eliminates the bias but also rules out consistent estimation of ξ. We, therefore, treat ξ as a nuisance parameter and design our test to be size-controlling for all values of ξ of empirical relevance. In practice, we would recommend practitioners 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 propensity score distribution (i.e., small propensity scores) as an illustration. The right tail of the propensity score can be studied by using $1/(1-e(X_i))$.

Researchers have developed several data-driven choices of k for estimating the tail index ξ (e.g., Drees and Kaufmann, 1998; Danielsson et al., 2001; Gomes and Oliveira, 2001; Guillou and Hall, 2001). See also Scarrott and MacDonald (2012) for a review of more methods. We follow Guillou and Hall (2001) to use the following algorithm.

Again denote $Y_i = 1/e(X_i)$. Given a random sample $\{Y_1, Y_2, \ldots, Y_n\}$, we first sort them in descending order and denote the order statistics by $Y_{(1)} \geq Y_{(2)} \geq \cdots \geq Y_{(n)}$. Define $Z_i = i \log(Y_{(i)}/Y_{(i+1)})$ for $i = 1, \ldots, n-1$. For each $k = 1, \ldots, n-1$, construct

$$\mathcal{T}_k \equiv \left(\sum_{i=1}^{k} w_i^2 \right)^{-1/2} \hat{\xi}_k^{-1} U_k,$$

where $w_i = \operatorname{sgn}(k-2i+1)\,|k-2i+1|$, $U_k \equiv \sum_{i=1}^{k} w_i Z_i$, and $\hat{\xi}_k$ denotes the Hill estimator

$$\hat{\xi}_k = \frac{1}{k} \sum_{i=1}^{k} \left(\log Y_{(i)} - \log Y_{(k+1)} \right).$$

When the distribution of Y_i (the inverse propensity score) is well approximated by the Pareto distribution, $\mathcal{T}_k$ should have its mean close to zero and variance close

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

to one. Accordingly, we can minimize the following criteria based on a moving average of $\mathcal{T}_k^2$:

$$\mathcal{C}_k = \left((2\lfloor k/2 \rfloor + 1)^{-1} \sum_{j=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \mathcal{T}_{k+j}^2 \right)^{1/2}.$$

Denote $\tilde{\mathcal{C}}_k = \min_{t \geq k} \mathcal{C}_t$. The optimal value k^* of k is

$$k^* = \min_{1 \leq k \leq \bar{n}} \{k : \tilde{\mathcal{C}}_k > 1.5\}, \tag{5.1}$$

where we set $\bar{n} = \lfloor 0.05n \rfloor$ so that we treat at most top 5% extreme values as stemming from the Pareto tail.

Note that this choice of k relies on additional assumptions that are stronger than our Assumption 1. Therefore, the optimal choice in (5.1) serves only as a heuristic starting point, and we still recommend using a range of k for robustness checks. We implement such procedures in our empirical applications.

We close this section by summarizing the empirical strategy in the following steps.

Step 1 Choose a propensity score model, such as the Logit or the semiparametric method. This should be the same specification used by the practitioner for subsequent analyses, say, treatment effect estimation.

Step 2 Sort the estimated propensity scores and use the above data-driven method to select the initial k.

Step 3 Choose a range of k around the initial choice and for each value of k in the range, apply the proposed test as in (2.4).

6. SIMULATIONS

This section evaluates the finite-sample performance of the test (2.4) with both the Logit and the semiparametric propensity score estimators. 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 Logit 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

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

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

					Frequency of rejecting H_0				
ξ	δ_0	n	$\|\hat{\beta} - \beta_0\|^2$	$\max_i \left\|\frac{e_i}{\hat{e}_i} - 1\right\|$	$k = 5$	25	50	75	100
2.0	Design 1	2,000	0.042	7.973	0.022	0.002	0.001	0.000	0.000
		4,000	0.021	2.756	0.014	0.003	0.000	0.000	0.000
		8,000	0.010	1.557	0.017	0.001	0.000	0.000	0.000
	2	2,000	0.031	5.539	0.022	0.002	0.001	0.000	0.000
		4,000	0.015	2.025	0.015	0.003	0.000	0.000	0.000
		8,000	0.007	1.321	0.017	0.001	0.000	0.000	0.000
	3	2,000	0.032	5.247	0.022	0.002	0.001	0.000	0.000
		4,000	0.015	1.856	0.016	0.003	0.000	0.000	0.000
		8,000	0.008	1.197	0.017	0.001	0.000	0.000	0.000
1.0	Design 1	2,000	0.059	1.651	0.043	0.055	0.066	0.077	0.090
		4,000	0.034	1.173	0.039	0.044	0.059	0.069	0.073
		8,000	0.018	0.810	0.042	0.048	0.043	0.054	0.055
	2	2,000	0.042	1.450	0.044	0.057	0.062	0.078	0.088
		4,000	0.021	1.070	0.041	0.045	0.058	0.068	0.069
		8,000	0.011	0.781	0.042	0.048	0.046	0.053	0.055
	3	2,000	0.042	1.392	0.045	0.058	0.064	0.083	0.090
		4,000	0.021	1.020	0.041	0.045	0.062	0.071	0.073
		8,000	0.011	0.740	0.042	0.050	0.046	0.056	0.058
0.67	Design 1	2,000	0.083	1.243	0.059	0.156	0.281	0.397	0.496
		4,000	0.052	0.937	0.055	0.145	0.266	0.408	0.494
		8,000	0.030	0.699	0.058	0.146	0.257	0.397	0.507
	2	2,000	0.058	1.124	0.055	0.162	0.289	0.429	0.510
		4,000	0.030	0.841	0.061	0.157	0.282	0.426	0.523
		8,000	0.016	0.644	0.061	0.153	0.260	0.423	0.540
	3	2,000	0.057	1.108	0.056	0.168	0.293	0.435	0.519
		4,000	0.030	0.821	0.061	0.161	0.286	0.433	0.535
		8,000	0.016	0.621	0.061	0.155	0.267	0.429	0.550

ξ across simulations. Notice that $\mathbb{P}\left[X_i'\beta_0 < -x\right] = 0.5\exp(-x/\xi)$ under this data generating design, and thus $\xi \geq 1$ (resp. < 1) implies the null (resp. alternative) hypothesis.

Table 1 collects simulation results. Displayed are the simulated average of $\|\hat{\beta} - \beta_0\|$, simulated average of $\max_{1\leq i\leq n}\left|e_i/\hat{e}_i - 1\right|$, 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.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Furthermore, the simulated average of $\max_{1\le i\le n}\left|e_i/\hat{e}_i-1\right|$ 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.

In the following, we summarize the findings in our simulation study regarding the rejection frequency of the test (2.4). 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 a 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 procedure performs very well both in terms of size control and statistical power.

7. APPLICATIONS

To illustrate the empirical applicability of our method, we revisit two datasets that 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 a 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 the 1970s with the aim of providing work experience to economically disadvantaged workers lacking job skills. Since LaLonde (1986), this dataset 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 ($D_i=0$). 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 detailed information on the 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 3.

To conduct a formal diagnosis, we implement our proposed test (2.4) with various choices of k to both the left tail and the right tail of the distribution of the estimated propensity score. We implement the algorithm in Section 5 for the

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

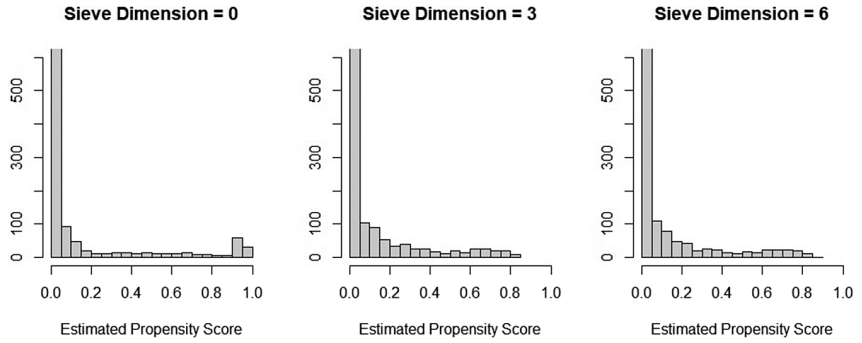


FIGURE 3. Histograms of estimated propensity scores in the NSW illustration. The case $\delta = 0$ corresponds to our baseline parametric estimate with a Logit specification.

TABLE 2. P-values of the fixed-k test in the NSW illustration. Left tail: testing if $\mathbb{E}[1/e(X_i)] = \infty$; Right tail: testing if $\mathbb{E}[1/(1-e(X_i))] = \infty$. 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	0.96	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

initial choice of k. The selected k ranges from 27 to 133, depending on δ. For robustness, we report the p-values for k ranging from 25 to 150. In particular, the left tail entails testing if $\mathbb{E}[1/e(X_i)] = \infty$, and the right tail entails testing if $\mathbb{E}[1/(1-e(X_i))] = \infty$. 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, 2022)[7] or change the parameter of interest (say, employ a trimming strategy or instead estimate the treatment effect on the treated).

Our second application examines the Right Heart Catheterization (RHC) dataset studied by Connors et al. (1996) and subsequently by Hirano and Imbens (2001)

[7]These methods produce debiased estimates for treatment effects with valid standard errors under their assumptions, which we do not further discuss in this paper.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

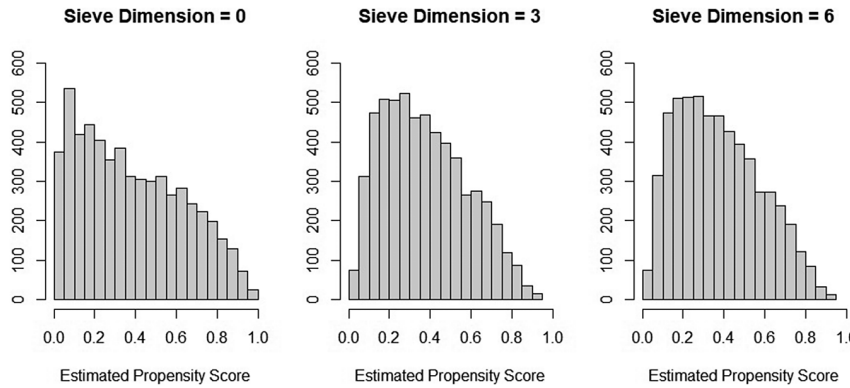


FIGURE 4. Histograms of estimated propensity scores in the RHC illustration. The case $\delta = 0$ corresponds to our baseline parametric estimate with a Logit specification.

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 spans the entire interval $(0, 1)$. Crump et al. (2009) thus propose to trim observations near the tails when conducting inverse propensity score weighting.

Our data consist 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 4 collects the histograms of the estimated propensity scores.

We implement our proposed test (2.4) with various choices of k. The choice based on the algorithm in Section 5 yields k ranging from 123 to 286. The p-values are reported in Table 3. Since the results with large k are all zeros, we present the results up to 150 as in Table 2. Except for the case with $k = 50$ and $\delta = 3$ and 6, the test always rejects the null hypothesis of heavy tails, and therefore we expect that the root-n asymptotic Gaussian approximation of the classical average treatment effect estimator is reliable for this application.

8. CONCLUDING REMARKS

We proposed a formal statistical test that can help assess the degree of limited overlap and hence treatment heterogeneity. This test takes the largest/smallest estimated propensity scores as input, and controls size over a large range of underlying propensity score tail distributions. Rejecting the null hypothesis indicates either

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

TABLE 3. *P*-values of the fixed-k test in the RHC 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

no or a 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.

A. 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 $\mathsf{H}_0 : \mathbb{E}[1/(e_i(1-e_i))] = \infty$ versus $\mathsf{H}_1 : \mathbb{E}[1/(e_i(1-e_i))] < \infty$. 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 A.1. *The following equivalence holds:*

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

A proof of this lemma is found in the following appendix. 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

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

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 et al., 2008, Chap. 8). Therefore, we can construct the likelihood ratio statistic in a similar fashion to (2.4), where the integrals are taken with respect to both ξ_1 and ξ_2.

B. APPENDIX B: Proofs

B.1. Proof of Lemma A.1

For the "only if" part, note that

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

$$\mathbb{E}\Big[\frac{1}{e_i(1-e_i)}\Big] = \mathbb{E}\Big[\frac{1}{e_i(1-e_i)}\mathbb{1}(e_i \le 0.5)\Big] + \mathbb{E}\Big[\frac{1}{e_i(1-e_i)}\mathbb{1}(e_i > 0.5)\Big].$$

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

$$\infty = \mathbb{E}\Big[\frac{1}{e_i(1-e_i)}\mathbb{1}(e_i \le 0.5)\Big] \le 2\mathbb{E}\Big[\frac{1}{e_i}\mathbb{1}(e_i \le 0.5)\Big] \le 2\mathbb{E}\Big[\frac{1}{e_i}\Big].$$

Therefore, $\mathbb{E}[1/e_i]$ is ∞.

For the "if" part, note that

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

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

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

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

$$\max_i |\eta_i| = \max_i \Big|\hat{Y}_i - Y_i\Big| \le Y_{(1)} \max_i \Big|\frac{\hat{Y}_i}{Y_i} - 1\Big| = o_p(a_n).$$

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

Then by the continuous mapping theorem, we have

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

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

The rest of the proof follows from the construction of the weights W_0 and cv_α. See Appendix C ahead. In particular, denote our test (2.4) 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 $\hat{\beta}$. Letting c_1 be some constant, we further decompose the above based on two events

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

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.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

B.4. Omitted Details of Remark 1

Part 1. As the population moment condition is uniquely maximized at β_0, we know that, for example, $\mathbb{E}[D_i \ln\big(G(X_i'\beta_0)\big)] > -\infty$. To show that $\hat{\beta}$ is consistent for β_0, it suffices to provide an integrable envelop function (Lemma 2.4 in Newey and McFadden, 1994). Take an arbitrary β in the parameter space $\mathcal{B}$, then $G(X_i'\beta) \geq e^{X_i'\beta}$ up to a multiplicative constant. This implies, again up to some constants that do not depend on β, $\mathbb{E}[D_i \ln\big(G(X_i'\beta)\big)] \geq \mathbb{E}[X_i'\beta] \geq \mathbb{E}[-\|X_i\|]$. This last expectation will be finite since the parameter space is compact and that X_i has a finite $2+\epsilon$ moment. The argument used to bound $\mathbb{E}[(1-D_i)\ln\big(1-(G(X_i'\beta_0)\big)\big)]$ is completely analogous.

Next, we consider the score

$$s_i(\beta) = \Big[\frac{D_i}{G(X_i'\beta)} - \frac{1-D_i}{1-G(X_i'\beta)}\Big] G^{(1)}(X_i'\beta) X_i.$$

To bound its second moment (at the true parameter β_0), it suffices to consider, for example,

$$\mathbb{E}\Big[\frac{D_i}{G(X_i'\beta_0)^2} G^{(1)}(X_i'\beta_0)^2 \|X_i\|^2\Big] = \mathbb{E}\Big[\Big(1-G(X_i'\beta_0)\Big)\|X_i\|^2\Big],$$

which is finite given our assumption on X_i. To characterize the $\sqrt{n}$-consistency and asymptotic normality of $\hat{\beta}$, what remains is to show the convergence to the Hessian matrix. To make it precise, let

$$H_i(\beta) = \Bigg[D_i \frac{G^{(2)}(X_i'\beta) - G(X_i'\beta)\Big(1-G(X_i'\beta)\Big)^2}{G(X_i'\beta)} - (1-D_i)\frac{G^{(2)}(X_i'\beta) + G(X_i'\beta)^2\Big(1-G(X_i'\beta)\Big)}{1-G(X_i'\beta)}\Bigg] X_i X_i',$$

and we would like to show $\mathbb{E}[\sup_{\|\beta-\beta_0\|\leq \frac{\epsilon}{\sqrt{n}}} \|H_i(\beta)\|] < \infty$ for all $\epsilon > 0$. This is particularly easy due to the Logit model we consider. To be precise, one has $|G^{(2)}(X_i'\beta)| \leq G(X_i'\beta)$ and $|G^{(2)}(X_i'\beta)| \leq 1 - G(X_i'\beta)$. As a result, $\|H_i(\beta)\| \leq \|X_i\|^2$ up to a multiplicative constant which does not depend on β. This concludes our proof that $\|\hat{\beta} - \beta_0\| = O_p(1/\sqrt{n})$.

Part 2. Now we prove Assumption 3(ii) with an explicit order c_n. To start, note that

$$\left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| \leq \frac{e^{X_i'\beta_0}}{1+e^{X_i'\beta_0}} \frac{1}{e^{X_i'\beta}} (1+\|X_i\|) \leq e^{X_i'(\beta_0-\beta)}(1+\|X_i\|),$$

which implies that

$$\sup_{\|\beta-\beta_0\|\leq \frac{c}{\sqrt{n}}} \left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| \leq (1+\|X_i\|) e^{\frac{c}{\sqrt{n}}(1+\|X_i\|)}.$$

As a result,

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

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

As long as $\|X_i\|$ has a finite $2+\epsilon$ moment for some $\epsilon > 0$, it will be true that $\max_{1\le i\le n}\|X_i\| = O_p(n^{\frac{1}{2+\epsilon}}) = o_p(\sqrt{n})$, and hence

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

which means we can set $c_n = n^{\frac{1}{2+\epsilon}-\frac{1}{2}} = n^{-\frac{\epsilon}{2(2+\epsilon)}} = o(1)$.

B.5. Omitted Details of Remark 2

Part 1. To show the consistency of $\hat{\beta}$, we again need to find an integrable envelop function. Take an arbitrary β in the parameter space $\mathcal{B}$, then we consider the expectation

$$\mathbb{E}[D_i\ln\big(G(X_i'\beta)\big)] = \mathbb{E}[D_i\ln\big(G(X_i'\beta)\big)\mathbb{1}_{(X_i'\beta\le -c)}] + \mathbb{E}[D_i\ln\big(G(X_i'\beta)\big)\mathbb{1}_{(X_i'\beta> -c)}],$$

for some c positive and large enough. Note that the second term is bounded from below by a finite constant which does not explicitly depend on the choice of β. As for the first term, we further employ a well-known bound on the normal tail probability (Proposition 2.1.2 in Vershynin, 2018)

$$\mathbb{E}[D_i\ln\big(G(X_i'\beta)\big)\mathbb{1}_{(X_i'\beta\le -c)}] \ge -\mathbb{E}\Big[D_i\ln(2X_i'\beta)\mathbb{1}_{(X_i'\beta\ge c)}\Big] - \mathbb{E}\Big[D_i|X_i'\beta|^2\mathbb{1}_{(X_i'\beta\ge c)}\Big],$$

where $g(\cdot) = G^{(1)}(\cdot)$ is the standard normal density function. The calculation for the other term, $\mathbb{E}[(1-D_i)\ln\big(1-G(X_i'\beta)\big)]$, is analogous. From the above, it is clear that one can find an integrable envelop function as the parameter space is compact.

Next, we consider the score

$$s_i(\beta) = \Big[\frac{D_i}{G(X_i'\beta)} - \frac{1-D_i}{1-G(X_i'\beta)}\Big]g(X_i'\beta)X_i.$$

To bound its second moment (at the true parameter β_0), it suffices to consider, for example,

$$\mathbb{E}\Big[\frac{D_i}{G(X_i'\beta_0)^2}g(X_i'\beta_0)^2\|X_i\|^2\Big]$$
$$= \mathbb{E}\Big[\frac{1}{G(X_i'\beta_0)}g(X_i'\beta_0)^2\|X_i\|^2\mathbb{1}_{(X_i'\beta_0\le -c)}\Big] + \mathbb{E}\Big[\frac{1}{G(X_i'\beta_0)}g(X_i'\beta_0)^2\|X_i\|^2\mathbb{1}_{(X_i'\beta_0> -c)}\Big].$$

Up to a multiplicative constant, the second term is finite provided that X_i has a finite variance. The first term can be bounded by

$$\mathbb{E}\Big[\frac{1}{G(X_i'\beta_0)}g(X_i'\beta_0)^2\|X_i\|^2\mathbb{1}_{(X_i'\beta_0\le -c)}\Big] \le \mathbb{E}\Big[\frac{2|X_i'\beta_0|}{g(X_i'\beta_0)}g(X_i'\beta_0)^2\|X_i\|^2\mathbb{1}_{(X_i'\beta_0\le -c)}\Big],$$

which is finite if X_i has a finite third moment.

To characterize the $\sqrt{n}$-consistency and asymptotic normality of $\hat{\beta}$, what remains is to show the convergence to the Hessian matrix. To be precise, let

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

$$H_i(\beta) = \Bigg[D_i \frac{G(X_i'\beta)G^{(2)}(X_i'\beta) - G^{(1)}(X_i'\beta)^2}{G(X_i'\beta)^2} - (1-D_i)\frac{\Big(1-G(X_i'\beta)\Big)G^{(2)}(X_i'\beta) + G^{(1)}(X_i'\beta)^2}{\Big(1-G(X_i'\beta)\Big)^2}\Bigg] X_i X_i',$$

and we would like to show $\mathbb{E}[\sup_{\|\beta-\beta_0\|\le\frac{\epsilon}{\sqrt{n}}} \|H_i(\beta)\|] < \infty$ for all $\epsilon > 0$. Consider, for example, the first half of the expectation:

$$\mathbb{E}\Big[\sup_{\|\beta-\beta_0\|\le\frac{\epsilon}{\sqrt{n}}} D_i \Big| \frac{G(X_i'\beta)G^{(2)}(X_i'\beta) - G^{(1)}(X_i'\beta)^2}{G(X_i'\beta)^2} \Big| |X_i|^2 \Big].$$

Again, it suffices to restrict to the case where $X_i'\beta$ is sufficiently small. Then we can bound the denominator by $G(X_i'\beta) \ge g(X_i'\beta)/(2|X_i'\beta|)$. As a result, finiteness of the previous follows from a finite fourth moment of X_i. This concludes our proof of the $\sqrt{n}$-consistency and asymptotic normality of $\hat{\beta}$.

Part 2. Now we prove Assumption 3(ii) with an explicit order c_n. To start, observe

$$\left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| \le \mathbb{1}_{(X_i'\beta \ge -2)} G(-2)^2 G(X_i'\beta_0) g(X_i'\beta)(1+\|X_i\|) + \mathbb{1}_{(X_i'\beta<-2)} \frac{g(X_i'\beta_0)}{g(X_i'\beta)} \left(\frac{|X_i'\beta|^3}{|X_i'\beta|^2 - 1} \right)^2 (1+\|X_i\|).$$

The first term above is easily bounded by

$$\mathbb{1}_{(X_i'\beta\ge-2)} G(-2)^2 G(X_i'\beta_0) g(X_i'\beta)(1+\|X_i\|) \le G(-2)^2 g(0)(1+\|X_i\|),$$

and the second term can be bounded by

$$\mathbb{1}_{(X_i'\beta<-2)} \frac{g(X_i'\beta_0)}{g(X_i'\beta)} \left(\frac{|X_i'\beta|^3}{|X_i'\beta|^2-1} \right)^2 (1+\|X_i\|) \le \frac{16}{9} e^{\frac{1}{2}(1+\|X_i\|)^2 \|\beta-\beta_0\| \|\beta+\beta_0\|} (1+\|X_i\|)^3.$$

Now assume $\|X_i\|$ has a finite $6+\epsilon$ moment for some $\epsilon > 0$. Then $\max_{1\le i\le n} \|X_i\| = O_p(n^{\frac{1}{6+\epsilon}})$, which implies that

$$\max_{1\le i\le n} \sup_{\|\beta-\beta_0\|\le\frac{c}{\sqrt{n}}} \left\| \frac{G(X_i'\beta_0)}{G(X_i'\beta)^2} \frac{\partial G(X_i'\beta)}{\partial \beta} \right\| = O_p(n^{\frac{3}{6+\epsilon}}).$$

Hence, we can set $c_n = n^{\frac{3}{6+\epsilon} - \frac{1}{2}} = n^{-\frac{\epsilon}{2(6+\epsilon)}}$.

B.6. Proof of Theorem 3

To prove Theorem 3, we first establish some auxiliary lemmas.

LEMMA B.1. *If Assumption 4(i–iv) is satisfied, then* $\|\hat{\psi} - \psi_0\|_1 = o_p(1)$ *and* $\|\hat{\psi} - \psi_0\|_2 = o_p(1)$.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Proof. The claim in this lemma follows from Bierens (2014, Thm. 4.2) with the metric d. We thus check his Assumptions 4.1 and 4.2 with d. Assumption 4.1(a) follows from our Assumption 4(i). Assumption 4.1(b) follows from our definitions of Ψ and d. Assumption 4.1(c) follows from our definition of g. Assumption 4.1(d) follows from Lemma 4.1 of Bierens (2014). Assumption 4.1(e) holds trivially. Assumption 4.1(f) follows from Lemma 2.1 of Bierens (2014) under our Assumption 4(ii–iv). Assumption 4.1(g) and (h) follows from our definitions of Ψ and Ψ_n. Assumption 4.1(i) holds with $\psi_n = (\beta_0', \Xi_{\ell_n - p}\delta_0')'$—see Section 4.2.2 of Bierens (2014). Assumption 4.1(j) follows from our Assumption 4(iv) and that G is a logistic link. Assumption 4.2 follows from our definitions of Ψ and Ψ_n. □

LEMMA B.2. *If Assumption 4 is satisfied, then* $\|\hat{\beta} - \beta_0\|_2 = O_p(n^{-1/2})$.

Proof. The claim in this lemma follows from Bierens (2014, Thm. 6.1) along with our definition of the metric $d_{(2)}$. We thus check his Assumptions 6.1–6.8. Assumptions 6.1(a) and 6.2 follow from the same argument as in the proof of Lemma B.1 under our Assumption 4(i–iv). Assumption 6.1(b)–(d) follows from our Assumption 4(iv) and (v). Assumptions 6.3–6.5 follow by Lemma 7.1 of Bierens (2014) from our definitions of Ψ and Ψ_n under our Assumption 4(ii–v). Assumption 6.6 follows from our Assumption 4(v). Assumptions 6.7 and 6.8 follow from our Assumption 4(v). □

LEMMA B.3. *Assume Assumption 4 holds. Let* c_n *be a vanishing sequence satisfying*

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

for all $c > 0$. *Then*

$$\max_{1\le i\le n} \left| \frac{G(X_i'\beta_0)}{G(X_i'\hat{\beta})} - 1 \right| = O_p(c_n).$$

Proof. This lemma can be established using the same proof strategy as in Theorem 2. □

LEMMA B.4. *Suppose Assumption 4 holds. If, in addition, that H is continuously differentiable with uniformly bounded derivatives, then*

$$\max_{1\le i\le n} \left| \frac{H(G(X_i'\hat{\beta});\delta_0) - H(G(X_i'\beta_0);\delta_0)}{G(X_i'\beta_0)} \right| = o_p(1).$$

Proof. The result follows from Lemma B.3 and the mean value expansion under Assumption 4(iv) implying that H is continuously differentiable with uniformly bounded derivatives. □

LEMMA B.5. *Suppose Assumption 4 holds. Then*

$$\sup_{u\in(0,1)} \left| \frac{H(u;\hat{\delta}) - H(u;\delta_0)}{u} \right| = o_p(1).$$

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Proof. First, note that the conclusion of Lemma B.1 implies

$$\|\hat{\delta} - \delta_0\|_1 = o_p(1), \tag{B.2}$$

$$\|\hat{\delta} - \delta_0\|_2 = o_p(1). \tag{B.3}$$

Furthermore, by the definition of the parameter space Ψ, we have

$$\sum_{j=1}^{\infty} \delta_{0j}^2 < \infty. \tag{B.4}$$

By the triangle inequality and Cauchy–Schwartz inequality,

$$\Big|\sum_{j=1}^{\infty}(\hat{\delta}_j^2 - \delta_{0j}^2)\Big| \le \sum_{j=1}^{\infty}(\hat{\delta}_j - \delta_{0j})^2 + 2\Big(\sum_{j=1}^{\infty}(\hat{\delta}_j - \delta_{0j})^2\Big)^{1/2}\Big(\sum_{j=1}^{\infty}\delta_{0j}^2\Big)^{1/2} = o_p(1)$$

and

$$\begin{aligned}\Big|\sum_{j=2}^{\infty}\sum_{m=1}^{j-1}(\hat{\delta}_j\hat{\delta}_m - \delta_{0j}\delta_{0m})\Big| &= \Big(\sum_{j=2}^{\infty}(\hat{\delta}_j - \delta_{0j})^2\Big)^{1/2}\Big(\sum_{m=1}^{j-1}(\hat{\delta}_m - \delta_{0m})^2\Big)^{1/2} \\ &+ \Big(\sum_{j=2}^{\infty}\delta_{0j}^2\Big)^{1/2}\Big(\sum_{m=1}^{j-1}(\hat{\delta}_m - \delta_{0m})^2\Big)^{1/2} + \Big(\sum_{j=2}^{\infty}(\hat{\delta}_j - \delta_{0j})^2\Big)^{1/2}\Big(\sum_{m=1}^{j-1}\delta_{0m}^2\Big)^{1/2} = o_p(1)\end{aligned}$$

by (B.3) and (B.4). Therefore,

$$\begin{aligned}&\sum_{j=1}^{\infty}\left(\frac{\hat{\delta}_j}{1+\sum_{i=1}^{\infty}\hat{\delta}_i^2} - \frac{\delta_{0j}}{1+\sum_{i=1}^{\infty}\delta_{0i}^2}\right) \\ &\quad= \frac{(1+\sum_{i=1}^{\infty}\delta_{0i}^2)\sum_{j=1}^{\infty}(\hat{\delta}_j - \delta_{0j}) - \sum_{j=1}^{\infty}\delta_{0j}\sum_{i=1}^{\infty}(\hat{\delta}_i^2 - \delta_{0i}^2)}{(1+\sum_{i=1}^{\infty}\delta_{0i}^2)(1+\sum_{i=1}^{\infty}\delta_{0i}^2 + \sum_{i=1}^{\infty}(\hat{\delta}_i^2 - \delta_{0i}^2))} = o_p(1), \\ &\sum_{j=1}^{\infty}\left(\frac{\hat{\delta}_j^2}{1+\sum_{i=1}^{\infty}\hat{\delta}_i^2} - \frac{\delta_{0j}^2}{1+\sum_{i=1}^{\infty}\delta_{0i}^2}\right) \\ &\quad= \frac{(1+\sum_{i=1}^{\infty}\delta_{0i}^2)\sum_{j=1}^{\infty}(\hat{\delta}_j^2 - \delta_{0j}^2) - \sum_{j=1}^{\infty}\delta_{0j}^2\sum_{i=1}^{\infty}(\hat{\delta}_i^2 - \delta_{0i}^2)}{(1+\sum_{i=1}^{\infty}\delta_{0i}^2)(1+\sum_{i=1}^{\infty}\delta_{0i}^2 + \sum_{i=1}^{\infty}(\hat{\delta}_i^2 - \delta_{0i}^2))} = o_p(1), \qquad \text{and}\end{aligned}$$

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

$$\sum_{j=2}^{\infty}\sum_{m=1}^{j-1}\left(\frac{\hat{\delta}_j\hat{\delta}_m}{1+\sum_{i=1}^{\infty}\hat{\delta}_i^2}-\frac{\delta_{0j}\delta_{0m}}{1+\sum_{i=1}^{\infty}\delta_{0i}^2}\right)$$
$$=\frac{(1+\sum_{i=1}^{\infty}\delta_{0i}^2)\sum_{j=2}^{\infty}\sum_{m=1}^{j-1}(\hat{\delta}_j\hat{\delta}_m-\delta_{0j}\delta_{0m})-\sum_{j=2}^{\infty}\sum_{m=1}^{j-1}\delta_{0j}\delta_{0m}\sum_{i=1}^{\infty}(\hat{\delta}_i^2-\delta_{0i}^2)}{(1+\sum_{i=1}^{\infty}\delta_{0i}^2)(1+\sum_{i=1}^{\infty}\delta_{0i}^2+\sum_{i=1}^{\infty}(\hat{\delta}_i^2-\delta_{0i}^2))}$$
$$=o_p(1)$$

by (B.2)–(B.4). Note also that

$$\left|\frac{\sin(j\pi u)}{j\pi u}\right|\le 1$$

holds for any $j\in\mathbb{Z}$ and $u\in(0,1)$. These equations together imply

$$\sup_{u\in(0,1)}\left|\frac{\Upsilon(u;\hat{\delta})-\Upsilon(u;\delta_0)}{u\left(1+\sum_{i=1}^{\infty}\delta_{0i}^2\right)}\right|=o_p(1).$$

From the definition of $H(u;\delta)$, this completes a proof of the lemma. □

With these lemmas, we are now ready to establish Theorem 3.

Proof of Theorem 3. It suffices to prove that

$$\max_{1\le i\le n}\left|\frac{H(G(X_i'\beta_0);\delta_0)}{H(G(X_i'\hat{\beta});\hat{\delta})}-1\right|=o_p(1)$$

holds under Assumption 4. First, note that we have

$$\max_{1\le i\le n}\left|\frac{G(X_i'\hat{\beta})-G(X_i'\beta_0)}{G(X_i'\beta_0)}\right|=o_p(1),$$
$$\max_{1\le i\le n}\left|\frac{H(G(X_i'\hat{\beta});\delta_0)-H(G(X_i'\beta_0);\delta_0)}{G(X_i'\beta_0)}\right|=o_p(1),\qquad\text{and}$$
$$\max_{1\le i\le n}\left|\frac{H(G(X_i'\hat{\beta});\hat{\delta})-H(G(X_i'\hat{\beta});\delta_0)}{G(X_i'\hat{\beta})}\right|=o_p(1)$$

by Lemmas B.3–B.5, respectively. Together with these uniform convergences, the equality

$$\frac{H(G(X_i'\beta_0);\delta_0)}{H(G(X_i'\hat{\beta});\hat{\delta})}-1$$
$$=-\left[\frac{H(G(X'\hat{\beta});\hat{\delta})-H(G(X'\hat{\beta});\delta_0)}{G(X'\hat{\beta})}\left(1+\frac{G(X'\hat{\beta})-G(X'\beta_0)}{G(X'\beta_0)}\right)\right.$$
$$\left.+\frac{H(G(X'\hat{\beta});\delta_0)-H(G(X'\beta_0);\delta_0)}{G(X'\beta_0)}\right]\Bigg/\left[\frac{H(G(X'\beta_0);\delta_0)}{G(X'\beta_0)}\right.$$

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

$$+\frac{H(G(X'\hat{\beta});\hat{\delta})-H(G(X'\hat{\beta});\delta_0)}{G(X'\hat{\beta})}\left(1+\frac{G(X'\hat{\beta})-G(X'\beta_0)}{G(X'\beta_0)}\right)$$
$$+\frac{H(G(X'\hat{\beta});\delta_0)-H(G(X'\beta_0);\delta_0)}{G(X'\beta_0)}\Bigg]$$

and Assumption 4(iv) that $x \mapsto H(G(x'\beta_0);\delta_0)/G(x'\beta_0)$ is bounded away from zero prove the statement of the theorem. □

C. APPENDIX C: Computational Details

This appendix provides computational details about constructing the test (2.4). The input is $\mathbf{V}^* =: \lim_{n\to\infty}\mathbf{T}$, whose density f_ξ is in (3.1) and computed by Gaussian Quadrature. To construct the test (2.4), 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 (2022, 2023). 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 $\tilde{W}_0 = \mathrm{cv}_\alpha W_0$ to subsume the critical value. Denote $\varphi_{\tilde{W}_0}(\cdot)$ as the test (2.4) 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 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 a 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 $\mathrm{cv}_\alpha = 1$. 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 jth 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 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 well 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 on a modern PC. The most time-consuming part is the calculation of the density f_ξ by Gaussian Quadrature. After conducting

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

TABLE C1. Logarithm of the critical values of the test (2.4). Based on 10,000 simulations.

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

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 C1. 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.

D. APPENDIX D: A Consistent Estimator of ξ

Our test (2.4) is based on the fixed-k asymptotic framework, in which ξ cannot be consistently estimated. For completeness, we provide a consistent estimator of ξ and its asymptotic normality in this appendix. Since these additional results require much stronger assumptions than those imposed in the main text, we present them separately here in the appendix.

Again, let $Y_i = 1/e(X_i)$ and $\hat{Y}_i = 1/\hat{e}(X_i)$. Recall that F denotes the CDF of Y_i. We make the following assumptions.

Assumption D.1. Y_i is regularly varying at infinity with $\alpha > 0$. Moreover, F satisfies the second-order tail expansion that

$$1 - F(x) = c_1 x^{-\alpha} + c_2 x^{-\alpha+\rho\alpha}(1 + o(1)), x \to \infty$$

for some constants $c_1 > 0, c_2 \neq 0, \alpha > 0$, and $\rho < 0$.

Assumption D.2. $\sqrt{k_n} \max_{1 \le i \le n} \frac{\left|\hat{Y}_i - Y_i\right|}{1+Y_i} = o_p(1)$.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Assumption D.3. $k_n = o\left(n^{-2\rho/(1-2\rho)}\right)$.

We provide some discussions about these conditions. Assumption D.1 requires the underlying distribution of $1/e_i$ to be regularly varying, following which the consistency of Hill's estimator can be derived. To derive its asymptotic normality, we need the second-order condition, which has been widely imposed in statistics literature about estimating the tail index. See, for example, de Haan and Ferreira (2007, Chap. 2) for a review. In particular, Student-t distribution with ν degrees of freedom satisfies this condition with $\alpha = \nu$ and $\rho = -2/\nu$.

Assumption D.2 requires that the estimation error of the propensity is of a smaller order than $k_n^{-1/2}$ so that it becomes asymptotically negligible. This condition is recently proposed and studied by Girard, Stupfler, and Usseglio-Carleve (2021). Note that this condition is stronger than our previous Assumption 2. Assumption D.3 is imposed to diminish the asymptotic bias. If k_n is of the order $n^{-2\rho/(1-2\rho)}$, Hill's estimator will then have a nonzero asymptotic bias whose expression depends on the higher-order constants in Assumption D.1. Estimation of these higher-order constants, especially ρ, is very challenging and has an even slower convergence rate. Therefore, it is common to select k_n to be of a smaller order, which is close in spirit to the undersmoothing condition in kernel regressions.

Under these assumptions, we can treat $\hat{Y}_i$ as the unobserved Y_i and construct the classic Hill's estimator. In particular, for $\xi = 1/\alpha$, Hill's estimator can be written as

$$\hat{\xi} = \frac{1}{k_n}\sum_{j=1}^{k_n} \log(\hat{Y}_{(j)}) - \log(\hat{Y}_{(k_n+1)}).$$

The following theorem establishes the asymptotic normality of $\hat{\xi}$. The estimator for α is constructed as $\hat{\alpha} = 1/\hat{\xi}$, whose asymptotic normality follows from the delta method.

THEOREM D.1. *Suppose Assumptions D.1–D.3 hold. Then,*

$$\sqrt{k_n}\left(\hat{\xi} - \xi\right) \xrightarrow{d} \mathcal{N}\left(0, \xi^2\right).$$

Proof of Theorem D.1 Our Assumption D.1 is sufficient for the condition $\mathcal{C}_2(\gamma, \rho, A)$ in Girard et al. (2021), where their γ is our ξ and their ε_i corresponds to our Y_i. Our Assumption D.2 is their condition in equation (2), and our Assumption D.3 implies that their $\lambda = 0$. Then the result follows from their Corollary 2.1. □

REFERENCES

Andrews, D. W. K., & Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(2), 1383–1414.

Andrews, D. W. K., & Ploberger, W. (1995). Admissibility of the likelihood ratio test when a nuisance parameter is present only under the alternative. *Annals of Statistics*, 23(5), 1609–1629.

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

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., & Kato, K. (2018). High-dimensional econometrics and regularized GMM. arXiv:1806.01888.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Bierens, H. J. (2014). Consistency and asymptotic normality of sieve ML estimators under low-level conditions. *Econometric Theory*, 30(5), 1021–1077.

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

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531), 1449–1455.

Cattaneo, M. D., Jansson, M., & Ma, X. (2024). Local regression distribution estimators. *Journal of Econometrics*, 240(2), 105074.

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

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11), 889–897.

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

Danielsson, J., de Haan, L., Peng, L., & de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76(2), 226–248.

de Haan, L., & Ferreira, A. (2007). *Extreme value theory: An introduction*. New York: Springer.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluations of training programs. *Journal of the American Statistical Association*, 94(448), 1053–1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151–161.

Drees, H., & Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and Their Applications*, 75 (2), 149–172.

Elliott, G., Müller, U. K., & Watson, M. W. (2015). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica*, 83(2), 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., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181–213.

Feller, W. (1991). *An introduction to probability theory and its applications* (Vol. II, 2nd ed.). New York: John Wiley.

Girard, S., Stupfler, G., & Usseglio-Carleve, A. (2021). Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models. *Annals of Statistics*, 49(6), 3358–3382.

Gomes, M. I., & Oliveira, O. (2001). The bootstrap methodology in statistics of extremes—choice of the optimal sample fraction. *Extremes*, 4(4), 331–358.

Guillou, A., & Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 293–305.

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

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

Khan, S., & Tamer, E. (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(4), 604–620.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypothesis*. New York: Springer.
Ma, X., & Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532), 1851–1860.
Müller, U. K., & Wang, Y. (2017). Fixed-*k* asymptotic inference about tail properties. *Journal of the American Statistical Association*, 112(519), 1134–1143.
Newey, W. K., & McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In R. F. Engle, & D. L. McFadden (Eds.), *Handbook of econometrics* (Vol. IV, pp. 2111–2245). Elsevier.
Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408), 1024–1032.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.
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., & Ura, T. (2022). Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory*, 38(1), 66–112.
Sasaki, Y., & Wang, Y. (2022). Fixed-*k* inference for conditional extremal quantiles. *Journal of Business & Economic Statistics*, 40(2), 829–837.
Sasaki, Y., & Wang, Y. (2023). Diagnostic testing of finite moment conditions for the consistency and root-*n* asymptotic normality of the GMM and M estimators. *Journal of Business & Economic Statistics*, 41(2), 339–348.
Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10(1), 33–60.
Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators. *Journal of Econometrics*, 125(1–2), 305–353.
Vershynin, R. (2018). *High-dimensional probability*. Cambridge University Press.

https://doi.org/10.1017/S0266466624000161 Published online by Cambridge University Press