Robust Confidence Intervals for Average Treatment Effects under Limited Overlap

Christoph Rothe*

Abstract

Limited overlap between the covariate distributions of groups with different treatment assignments does not only make estimates of average treatment effects rather imprecise, but can also lead to substantially distorted confidence intervals. This paper argues that this is because the coverage error of traditional confidence intervals is driven by the number of observations in the areas of limited overlap. Some of these "local sample sizes" can be very small in applications, up to the point that distributional approximation derived from classical asymptotic theory become unreliable. Building on this observation, this paper constructs confidence intervals based on classical approaches to small sample inference. The approach is easy to implement, and has superior theoretical and practical properties relative to standard methods in empirically relevant settings.

JEL Classification: C12, C14, C25, C31

Keywords: *Average treatment effect; Causality; Overlap; Propensity score; Treatment effect heterogeneity; Unconfoundedness*

*First version: December 3, 2014. This version: October 31, 2016. Christoph Rothe, Department of Economics, Columbia University, 420 W 118th St., New York, NY 10027, Email: cr2690@columbia.edu. Website: http://www.christophrothe.net. I would like to thank the co-editor, the referees, Shakeeb Khan, Ulrich Müller, Miikka Rokkanen, Bernard Salanie, and seminar audiences at Columbia, Duke, Syracuse and the 2014 Greater NY Metropolitan Area Colloquium for their helpful comments.

1. Introduction

Empirical economic studies that involve estimating average treatment effects (ATEs) under the assumption of unconfounded assignment (Rosenbaum and Rubin, 1983) often face the problem of having only few observations in either the treatment or the non-treatment group in some regions of the covariate space. Even if the overall sample size is large, such areas of limited overlap can occur naturally if the propensity score takes on values close to either 0 or 1. Limited overlap has an adverse effect on the precision of many ATE estimators, whose asymptotic variance increases sharply as propensity scores accumulate closer to the boundaries of the unit interval. Moreover, nonparametric estimators of ATEs might converge at slower-than-usual rates if the propensity score can be arbitrarily close to 0 or 1 (Khan and Tamer, 2010). Appropriate overlap is thus important for obtaining precise ATE estimates, and this fact is widely appreciated by practitioners (e.g. Imbens, 2004).

A more subtle issue, which has received less attention in the literature, is that limited overlap also has a detrimental effect on inference. For example, the result in Khan and Tamer (2010) implies that in the absence of strong overlap the usual 95% confidence interval (CI) of the form "point estimate$\pm$1.96$\times$standard error" may no longer be valid. This in turn raises concerns about the accuracy of such a CI in applications where the propensity score is bounded away from 0 and 1, but only by a relatively small constant. Indeed, simulation results reported in this paper show that in finite samples the actual coverage probability of such a CI can be substantially below its nominal level, making estimates seem more precise than they are.

This paper explores the channels through which limited overlap affects the accuracy of standard methods for inference, and provides a practical approach to address the issue. To convey the main points, we consider a simple setup in which the covariates have known finite support. This benchmark model has the advantage that most estimation strategies commonly used in empirical practice deliver numerically identical results here. In this frame-

work, we show that the coverage error of a standard CI is not driven by the overall sample size, but by the numbers of observations in the smallest covariate-treatment cells. Since under limited overlap some of these numbers are only modest, the coverage error can be substantial. Inference on ATEs is thus hampered under limited overlap because of what one might call "locally" small samples.

Given this result, we propose a robust CI based on classical methods for small sample inference. Since with discrete covariates the natural ATE estimate is a linear combination of independent sample means, inference in this setup can be thought of as a generalization of the Behrens-Fisher problem (Behrens, 1928; Fisher, 1935), and be conducted using tools developed for that context. Our proposed CI, which builds on Banerjee (1960) and Mickey and Brown (1966), is based on a critical-value that adjusts in a data-driven way to the degree of overlap. This approach leads to finite-sample valid inference under any degree of overlap if the outcome data are normally distributed, has similarly good properties if the normality assumption is at least approximately satisfied, and does not perform worse (in a classic asymptotic sense) than standard methods if normality is clearly violated. We work with normality since without some restriction of this type it would seem impossible to obtain meaningful theoretical statements about the distribution of (studentized) average outcomes in covariate-treatment cells with very few observations.

In empirical practice, concerns about limited overlap are often addressed by estimating the ATE only for a subpopulation obtained by trimming units with propensity scores close the boundaries of the unit interval from the data (Crump et al., 2009). These redefined ATEs can be estimated with greater precision than the full-population ATE, and there are no concerns about the validity of standard CIs in this context. On the other hand, if treatment effects are heterogeneous, their average might be very different in the trimmed population relative to the original one. Trimming can therefore introduce a substantial bias in settings where the entire population is of policy relevance. Since observations are sparse

in the trimmed areas by construction, it is difficult to determine the magnitude of this bias from the data.[1] Our robust CI should be seen as a complement to trimming, and not as a replacement. Reporting point estimates and CIs for both a trimmed and the original population in empirical applications with limited overlap offers a more nuanced view of the informational content of the data than either procedure by itself.

Limited overlap can in principle also be addressed by imposing parametric restrictions that allow extrapolation from regions of the covariate space with many observations to regions of limited overlap where data are sparse. However, estimates based on such restrictions tend to be highly sensitive to even minor changes of the model. The validity of any parametric model in an area of limited overlap is also difficult to asses due to the small number of observations in those regions. The empirical setting would thus have to strongly imply a particular functional form for parametric extrapolation to be credible (e.g. Imbens and Rubin, 2015, Chapter 14).

2. Setup

We consider the standard program evaluation setup where interest is in the causal effect of a binary treatment on a scalar outcome. Let D be a treatment indicator such that $D = 1$ if a unit receives the treatment, and $D = 0$ otherwise. Define $Y(1)$ and $Y(0)$ as the potential outcome of the unit with and without receiving the treatment, respectively. The realized outcome is $Y = Y(D)$, and X is a vector of covariates. The data are an independent and identically distributed sample $\{(Y_i, D_i, X_i)\}_{i=1}^n$ from the distribution of (Y, D, X). The population average treatment effect (PATE) and sample average treatment effect (SATE)

[1] A similar comment applies to methods using a "vanishing" trimming approach based on an asymptotic experiment in which an ever smaller proportion of observations is trimmed as the same size increases (e.g. Khan and Tamer, 2010; Chaudhuri and Hill, 2014; Yang, 2014). Similarly to fixed trimming, such methods face a bias/variance-type trade-off which due to the special structure of treatment effect models is generally very challenging to resolve in finite samples.

are given by

$$\tau_P = \mathbb{E}(Y(1) - Y(0)) \quad \text{and} \quad \tau_S = \frac{1}{n}\sum_{i=1}^{n} \tau(X_i),$$

respectively, where $\tau(x) = \mathbb{E}(Y(1) - Y(0)|X = x)$ is the conditional average treatment effect (CATE).[2] We also write $\mu_d(x) = \mathbb{E}(Y|D = d, X = x)$ and $\sigma_d^2(x) = \mathrm{Var}(Y|D = d, X = x)$. Following Imbens (2000), we refer to $p_d(x) = P(D = d|X = x)$ as the *generalized* propensity score, and write $p(x) = p_1(x)$ for the "ordinary" propensity score. Throughout the paper, we maintain the *ignorability condition* of Rosenbaum and Rubin (1983), which asserts that the treatment status is independent of the potential outcomes given the covariates, and that the distribution of the covariates has the same support among the treated and the untreated.

Assumption 1. *(i)* $(Y(1), Y(0)) \perp D|X$ *and (ii)* $0 < p(X) < 1$ *with probability 1.*

Under this assumption, the CATE is identified as $\tau(x) = \mu_1(x) - \mu_0(x)$, and the PATE and SATE are identified as averages of $\tau(x)$ over the population and sampling distribution of X, respectively. Estimators of the PATE that are semiparametrically efficient under certain additional regularity conditions have been proposed for example by Hahn (1998), Hirano et al. (2003) and Imbens et al. (2007). These estimators are also appropriate and efficient for the SATE (Imbens, 2004). In addition to smoothness conditions on functions such as $\mu_d(x)$ or $p(x)$, the regularity conditions required by these estimators include that Assumption 1(ii) is strengthened to a *strong overlap* condition:

$$\epsilon < p(X) < 1 - \epsilon \text{ with probability 1 for some } \epsilon > 0. \tag{2.1}$$

Khan and Tamer (2010) show that without (2.1) the semiparametric efficiency bound for estimating τ_P or τ_S may not be finite, and thus no regular $\sqrt{n}$-consistent and asymptotically normal estimator might exist. We informally refer to a setting where (2.1) only holds for

[2] Our terminology follows that of Crump et al. (2009). The terms *conditional* and *sample average treatment effect* are sometimes used differently in the literature; see Imbens (2004) for example.

some very small $\epsilon > 0$ as having *limited overlap*.

3. Discrete Covariates

To show how exactly limited overlap affects the coverage error of standard CIs, and how this issue can be addressed, it is instructive to consider a simple setup where X has finite support $\mathcal{X} = \{x_1, \dots, x_J\}$, and the SATE is the parameter of interest.

3.1. Limited Overlap and Standard Inference. Write $f(x) = P(X = x)$, let $\mathcal{M}_d(x) = \{i : D_i = d, X_i = x\}$ be the set of indices of those observations with treatment status $D_i = d$ and covariates $X_i = x$, let $N_d(x) = \#\mathcal{M}_d(x)$ be the cardinality of this set, and put $N(x) = N_1(x) + N_0(x)$. We refer to $N_d(x)$ and $n_d(x) = \mathbb{E}(N_d(x))$ as the *realized* and *expected local sample size* at (d, x) in the following. Writing

$$\widehat{\mu}_d(x) = \frac{1}{N_d(x)} \sum_{i \in \mathcal{M}_d(x)} Y_i, \quad \widehat{f}(x) = \frac{N(x)}{n}, \quad \widehat{p}_d(x) = \frac{N_d(x)}{N(x)}, \quad \text{and} \quad \widehat{p}(x) = \widehat{p}_1(x),$$

the natural estimator[3] of the SATE (and the PATE) is then given by

$$\widehat{\tau} = \sum_{j=1}^{J} \widehat{f}(x_j)\widehat{\tau}(x_j) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}(X_i), \quad \text{where} \quad \widehat{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x).$$

The asymptotic variance $\omega_S^2 = \sum_{d,j} f(x_j)\sigma_d^2(x_j)/p_d(x_j)$ of $\widehat{\tau}$ as an estimator of the SATE can be estimated consistently by

$$\widehat{\omega}_S^2 = \sum_{d,j} \frac{\widehat{\sigma}_d^2(x_j)}{\widehat{p}_d(x_j)} \cdot \widehat{f}(x_j), \quad \text{where} \quad \widehat{\sigma}_d^2(x) = \frac{1}{N_d(x) - 1} \sum_{i \in \mathcal{M}_d(x)} (Y_i - \widehat{\mu}_d(x))^2.$$

This estimator is numerically well-defined as long as $\min_{d,x} N_d(x) \geq 2$, and all our analysis in the following is to be understood conditional on this. The resulting asymptotic normality of the studentized estimator $T_{S,n} = \sqrt{n}(\widehat{\tau} - \tau_S)/\widehat{\omega}_S$ then motivates the usual two-sided CI

[3]Note that with discrete covariates $\widehat{\tau}$ is numerically identical to other popular estimators based on sample analogues of alternative representations of ATEs. For example, our estimator also has an "inverse probability weighting" representation $\widehat{\tau} = n^{-1} \sum_{i=1}^{n} Y_i(D_i - \widehat{p}(X_i)) \cdot (\widehat{p}(X_i)(1 - \widehat{p}(X_i)))^{-1}$, as in Hirano et al. (2003). Working with discrete covariates thus shows that complications from limited overlap are not specific to one estimation strategy.

for τ_S with nominal level $1-\alpha$:

$$\mathcal{I}_{S,1} = \left(\widehat{\tau} - z_\alpha \cdot \widehat{\omega}_S/\sqrt{n}, \widehat{\tau} + z_\alpha \cdot \widehat{\omega}_S/\sqrt{n}\right),$$

where $z_\alpha = \Phi^{-1}(1-\alpha/2)$. The next proposition studies the coverage properties of $\mathcal{I}_{S,1}$.

Proposition 1. *(i) Under regularity conditions (Hall and Martin, 1988), it holds that*

$$P(\tau_S \in \mathcal{I}_{S,1}) = 1 - \alpha + n^{-1}\phi(z_\alpha)q_2(z_\alpha, f, p) + O\left(n^{-2}\right),$$

where $\phi(\cdot)$ is the standard normal density, and $q_2(z_\alpha, f, p)$ is a polynomial in z_α that is given explicitly in the appendix.

(ii) For sequences $f^{(n)}(x)$ of covariate densities and $p_d^{(n)}(x)$ of generalized propensity scores such that $nf^{(n)}(x)p_d^{(n)}(x) \to \infty$ as $n \to \infty$ for all (d,x), it holds that

$$n^{-1}\phi(z_\alpha)q_2(z_\alpha, f^{(n)}, p^{(n)}) = O(n_{d^*}(x^*)^{-1}),$$

where (d^*, x^*) is the point at which the ratio $p_d^{(n)}(x)/f^{(n)}(x)$ takes its smallest value; that is, (d^*, x^*) is such that $p_{d^*}(x^*)/f(x^*) = \liminf_{n\to\infty}\left(\min_{d,x} p_d^{(n)}(x)/f^{(n)}(x)\right)$.

The proposition shows that while the coverage error of $\mathcal{I}_{S,1}$ is formally of the order $O(n^{-1})$, it is effectively more similar to that of a CI computed from a sample whose size is equal to the expected local sample size in the covariate-treatment cell where the ratio of the generalized propensity score and the covariate density takes its smallest value. Under limited overlap, this local sample size can be small itself. The coverage error of $\mathcal{I}_{S,1}$ can therefore be substantial even when n is very large.

3.2. Robust Confidence Intervals. Since ATE inference under limited overlap has essential properties of a small sample problem, the use of large sample approximations to address the issue does not seem promising. Instead, we propose to adapt classical small sample methods to our setting. To motivate the approach, note that without covariates the

studentized estimator $T_{S,n}$ defined above is the test statistic of a two-sample t-test. Conditional on the number of treated and untreated individuals, inference on τ_S then reduces to the Behrens-Fisher problem of conducting inference on the difference of the means of two populations with unknown and potentially different variances.

Our setting is a generalized version of the Behrens-Fisher problem, since conditional on $\mathbf{M}_n = \{(X_i, D_i)\}_{i=1}^n$ the statistic $T_{S,n}$ is the studentized version of a linear combination of $2J$ independent sample means, each calculated from $N_d(x)$ realizations of a random variable with mean $(-1)^{1-d} \cdot \widehat{f}(x)\mu_d(x)$ and variance $\widehat{f}(x)^2\sigma_d^2(x)$. We can thus apply techniques from a longstanding literature in statistics that has studied solutions to Behrens-Fisher-type problems with small group sizes. Instead of relying on first-order asymptotic theory, this literature exploits assumptions about the distribution of the data. We consider the following assumption with the same purpose in mind.

Assumption 2. *$Y|(D, X) = (d, x) \sim N(\mu_d(x), \sigma_d^2(x))$ for all $(d, x) \in \{0, 1\} \times \mathcal{X}$.*

Assumption 2 is clearly restrictive; but without imposing some additional structure it would seem impossible to conduct valid inference in the presence of small groups.[4] Our proposed robust CI for the SATE is given by

$$\mathcal{I}_{S,2} = \left(\widehat{\tau} - c_\alpha(\delta_{\min})\rho_\alpha \cdot \widehat{\omega}_S/\sqrt{n}, \widehat{\tau} + c_\alpha(\delta_{\min})\rho_\alpha \cdot \widehat{\omega}_S/\sqrt{n}\right),$$

where $c_\alpha(\delta) = F_t^{-1}(1 - \alpha/2, \delta)$, $F_t(\cdot, \delta)$ denotes the CDF of Student's t-distribution with δ degrees of freedom, $\delta_{dj} = N_d(x_j) - 1$, $\delta_{\min} = \min_{d,j} \delta_{dj}$, and

$$\rho_\alpha = \left(\frac{\sum_{d,j}(c_\alpha(\delta_{dj})/c_\alpha(\delta_{\min}))^2 \cdot \widehat{f}(x_j)^2\widehat{\sigma}_d^2(x_j)/N_d(x_j)}{\sum_{d,j} \widehat{f}(x_j)^2\widehat{\sigma}_d^2(x_j)/N_d(x_j)}\right)^{1/2}.$$

[4]One can think of Assumption 2 as an "asymptotically irrelevant parameterization", as results obtained without this condition via asymptotic arguments do not change if this assumption holds. This is the case for the asymptotic normality of $T_{S,n}$ or Proposition 1, for example. Since the distribution of $Y|D, X$ is symmetric under Assumption 2, the summands in the definition of $q_2(t)$ in Proposition 1(i) that involve $\gamma_d(x)$ vanish, but the order of the coverage error and the statement of Proposition 1(ii) remain the same in this case.

The following proposition shows that under Assumption 2 the CI $\mathcal{I}_{S,2}$ does not under-cover the parameter of interest in finite samples for all values of the covariate density and the generalized propensity score, and is thus robust to weak overlap. It also shows that if Assumption 2 does not hold $\mathcal{I}_{S,2}$ has the same first-order asymptotic coverage error as $\mathcal{I}_{S,1}$, and is thus equally valid from a traditional large sample point of view.

Proposition 2. *(i) Under Assumptions 1–2, we have that* $P(\tau_S \in \mathcal{I}_{S,2}) \geq 1-\alpha$. *(ii) Under Assumption 1 and the regularity conditions of Proposition 1, we have that* $P(\tau_S \in \mathcal{I}_{S,2}) = P(\tau_S \in \mathcal{I}_{S,1}) + O(n^{-2})$.

The inequality in part (i) is sharp in the sense that $\inf_{\{\sigma_d^2(x_j):d=0,1;j=1,\ldots,J\}} P(\tau_S \in \mathcal{I}_{S,2}) = 1-\alpha$. The CI $\mathcal{I}_{S,2}$ thus implicitly inverts the decision of a two-sided hypothesis test with size α. This test is not similar, and in finite samples the coverage probability of $\mathcal{I}_{S,2}$ thus generally exceeds $1-\alpha$. However, our numerical results below suggest that the procedure only leads to minimal over-coverage in realistic settings.[5,6]

CIs of the form of $\mathcal{I}_{S,2}$ go back to at least Banerjee (1960); see also Hayter (2014) for a more recent reference. To understand their construction, note that $\mathcal{I}_{S,2}$ is not based on the usual t-statistic $T_{S,n}$. Instead, we begin by considering the class of test statistics of the form

$$T_{S,n}(h) = \frac{\sqrt{n}(\widehat{\tau} - \tau_S)}{\widehat{\omega}_S(h)}, \quad \widehat{\omega}_S^2(h) = \sum_{d,j} h_{dj} \cdot \frac{\widehat{\sigma}_d^2(x_j)}{\widehat{p}_d(x_j)} \cdot \widehat{f}(x_j)$$

indexed by the vector $h = (h_{01}, \ldots, h_{0J}, h_{11}, \ldots, h_{1J})' \in \mathbb{R}_+^{2J}$. This class comprises the statistic $T_{S,n}$ by setting $h = (1, \ldots, 1)'$. From an extension of the argument in Mickey and

[5]The work of Linnik (1966, 1968) and Salaevskii (1963) has shown that there are no exactly similar tests for the Behrens-Fisher problem that have desirable properties. A procedure that has correct size and only leads to minimal over-coverage even when cells contain as few as two observations thus appears to be very reasonable for this setting.

[6]In view of Ibragimov and Müller (2016), we conjecture that the above result continues to hold if Assumptions 2 is weakened to allow the conditional distribution of the outcome variable to follow a scale mixture of normals; but a formal proof of this statement is beyond the scope of this paper.

Brown (1966), it follows that for every $u > 0$ and every vector h we have

$$P(T_{S,n}(h) \leq u | \mathbf{M}_n) \geq \min_{d,j} F_t(u h_{dj}^{1/2}, \delta_{dj}).$$

This lower bound on the CDF of $T_{S,n}(h)$ translates directly into a bound on its quantiles, which in turn motivates CIs with nominal level $1-\alpha$ of the form

$$\left(\widehat{\tau} - \max_{d,j} \frac{c_\alpha(\delta_{dj})}{h_{dj}^{1/2}} \cdot \frac{\widehat{\omega}_S(h)}{\sqrt{n}}, \widehat{\tau} + \max_{d,j} \frac{c_\alpha(\delta_{dj})}{h_{dj}^{1/2}} \cdot \frac{\widehat{\omega}_S(h)}{\sqrt{n}} \right). \tag{3.1}$$

One can show that setting $h_{dj}^{1/2} \propto c_\alpha(\delta_{dj})$ for all (d,j) minimizes the length of this interval. This choice of h then yields $\mathcal{I}_{S,2}$ as the shortest, and in this sense "optimal", CI within the class of intervals of the form (3.1).

The critical value $c_\alpha(\delta_{\min})\rho_\alpha$ used in the construction of $\mathcal{I}_{S,2}$ adapts automatically to the degree of overlap. Some algebra that $c_\alpha(\delta_{\min}) \geq c_\alpha(\delta_{\min})\rho_\alpha \geq c_\alpha(n-2J)$, and that these relationships can potentially hold with equality. The CI $\mathcal{I}_{S,2}$ is thus always wider than $\mathcal{I}_{S,1}$; and if the realized size of some local sample is small, the difference in length can be substantial. For example, if $\delta_{\min} = 1$, which is the smallest value for which the CI is numerically well-defined, and $\widehat{f}(x_j)^2 \widehat{\sigma}_d^2(x_j)/N_d(x_j) \approx 0$ for all (d,j) except for that cell corresponding to $\delta_{\min}$, then $c_\alpha(\delta_{\min})\rho_\alpha \approx c_\alpha(1) \approx 6.48 \cdot z_\alpha$ for $\alpha = .05$. On the other hand, $\mathcal{I}_{S,1}$ and $\mathcal{I}_{S,2}$ are very similar if $\delta_{\min} \geq 50$ or so, since at conventional significance levels the quantiles of the standard normal distribution do not differ much from those of a t distribution with at least 50 degrees of freedom.

4. General Covariates

In many empirical applications the covariates are continuously distributed, or have discrete support that is sufficiently rich that there are less than two observations in some of the cells. In such cases some aggregation or smoothing is needed to estimate treatment effects. Among the many different empirical strategies that are available for this purpose, the one

that combines most naturally with our approach to building robust CIs for the SATE is subclassification on the propensity score (Cochran, 1968; Imbens and Rubin, 2015). We first estimate the propensity score by some method deemed suitable for the respective context, then choose a partition of $[0,1]$, and finally treat= an indicator for the cell containing a unit's estimated propensity score in the same way we treated a discrete covariate in Section 3. Partitioning and propensity score estimation introduce a bias that can be reduced by adjusting for covariates among units whose propensity scores fall within the same cell.

To describe the procedure formally, let be $\widehat{p}(x)$ be an estimate of the propensity score, choose constants $\{\pi_j\}_{j=0}^J$ satisfying $0 = \pi_0 < \pi_1 < \ldots < \pi_J = 1$, and put $S_j(x) = \mathbb{I}(\pi_{j-1} \leq \widehat{p}(x) < \pi_j)$ for $j = 1, \ldots, J-1$ and $S_J(x) = \mathbb{I}(\pi_{J-1} \leq \widehat{p}(x) \leq \pi_J)$. For any $x \in \mathbb{R}^{\dim(X)}$ and $K \in \mathbb{N}$, let $R^K(x)$ be a column vector containing all polynomials in x up to order $K-1$. We then write $R_j(x) = S_j(x)R^K(x)$, and define[7]

$$\widehat{\mu}_d(x) = \sum_{j=1}^{J} R_j(x)'\widehat{\beta}_{dj}, \text{ where } \widehat{\beta}_{dj} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} \mathbb{I}(D_i = d) \left(Y_i - R_j(X_i)'\beta\right)^2.$$

The natural estimate of the SATE is $\widehat{\tau} = n^{-1}\sum_{i=1}^n(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i))$; and following arguments along the lines of those in Section 3.2, we obtain the robust CI

$$\bar{\mathcal{I}}_{S,2} = \left(\widehat{\tau} - c_\alpha(\bar{\delta}_{\min})\bar{\rho}_\alpha \cdot \widehat{\omega}_S/\sqrt{n}, \widehat{\tau} + c_\alpha(\bar{\delta}_{\min})\bar{\rho}_\alpha \cdot \widehat{\omega}_S/\sqrt{n}\right)$$

where

$$\begin{aligned}
\widehat{\omega}_S &= \sum_{d,j} \widehat{L}_j'\widehat{Q}_{dj}^{-1}\widehat{L}_j\widehat{\sigma}_{dj}^2, \quad \widehat{\sigma}_{dj}^2 = \frac{1}{N_{dj} - K}\sum_{i=1}^n S_j(X_i)\mathbb{I}(D_i = d)(Y_i - \widehat{\mu}_d(X_i))^2, \\
\widehat{L}_j &= \frac{1}{n}\sum_{i=1}^n R_j(X_i), \quad \widehat{Q}_{dj} = \frac{1}{N_{dj}}\sum_{i=1}^n \mathbb{I}(D_i = d)R_j(X_i)R_j(X_i)', \\
\bar{\rho}_\alpha &= \left(\frac{\sum_{d,j}(c_\alpha(\bar{\delta}_{dj})/c_\alpha(\bar{\delta}_{\min}))^2 \cdot \widehat{L}_j'\widehat{Q}_{dj}^{-1}\widehat{L}_j\widehat{\sigma}_{dj}^2/N_{dj}}{\sum_{d,j}\widehat{L}_j'\widehat{Q}_{dj}^{-1}\widehat{L}_j\widehat{\sigma}_{dj}^2/N_{dj}}\right)^{1/2},
\end{aligned}$$

[7]The "argmin" operator in the following equation is to be understood such that it returns the solution with the smallest Euclidean length in case the set of minimizers of the corresponding least squares problem is not unique.

with $\bar{\delta}_{\min} = \min_{d,j} N_{dj} - K$, $\bar{\delta}_{dj} = N_{dj} - K$, and $N_{dj} = \sum_{i=1}^{n} S_j(X_i)\mathbb{I}(D_i = d)$ the number of observations with treatment status d in the jth cell of propensity score values. Now write $\bar{\tau}_S = \mathbb{E}(\widehat{\tau}|\mathbf{M}_n)$, $\bar{\mu}_d(x) = \mathbb{E}(\widehat{\mu}_d(x)|\mathbf{M}_n)$, let $\{\bar{\sigma}^2_{d,j}, d = 0, 1; j = 1, \ldots, J\}$ be some positive constants, and let $j(x)$ be such that $S_{j(x)}(x) = 1$.

Corollary 1. *Suppose Assumption 1 holds. (i) If* $Y_i|\mathbf{M}_n \overset{\text{iid}}{\sim} \mathcal{N}(\bar{\mu}_{D_i}(X_i), \bar{\sigma}^2_{D_i,j(X_i)})$*, then* $P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,2}) \geq 1 - \alpha$*. (ii) Under the regularity conditions of Proposition 1, we have that* $P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,2}) = P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,1}) + O(n^{-2})$*, where* $\bar{\mathcal{I}}_{S,1} = (\widehat{\tau} \pm z_\alpha \cdot \widehat{\omega}_S/\sqrt{n})$ *is the standard CI.*

The parameter $\bar{\tau}_S$ is the sum of the SATE and the bias resulting from propensity score estimation and the fact that propensity scores are only approximately constant on the chosen partition of the unit interval. The corollary therefore justifies the use of $\bar{\mathcal{I}}_{S,2}$ as an approximate CI for τ_S if the bias is deemed negligible relative to sampling uncertainty, and the data generating process for the outcome (conditional on $\mathbf{M}_n$) is sufficiently well approximated by a piecewise linear model with normal, homoskedastic errors over sections of the covariate space defined by the values of the estimated propensity score. Note that $\bar{\mathcal{I}}_{S,2}$ adapts to the choice of tuning parameters, as for example it generally becomes wider if a finer partition of the unit interval or a higher-order polynomial approximation within cells is used.[8] Also note that the precise nature of the estimator of the propensity score does not affect our results since inference on the SATE is conditional on $\mathbf{M}_n$, and thus the estimate is effectively a non-random quantity.

5. Simulations

This section reports results from a simple Monte Carlo study. To ensure that the SATE remains constant across simulation runs, we hold $\mathbf{M}_n = \{(D_i, X_i)\}_{i=1}^n$ constant in each repetition, and only simulate new values of the outcome variables. Specifically, we put

[8]The choice of tuning parameters affects the properties of nonparametric two-stage estimators and corresponding methods for inference in general, not just in treatment effect settings. See Robins and Ritov (1997), Cattaneo et al. (2013), or Rothe and Firpo (2016) for some recently studies in this area.

$n = 2000$, $\mathcal{X} = \{1, 2, \ldots, 10\}$, and construct $\mathbf{M}_n$ such that $\widehat{f}(x) = 0.1$ for all $x \in \mathcal{X}$ and $\widehat{p}(x) = 0.5$ for $x \in \mathcal{X} \backslash \{10\}$. We then consider various scenarios where $\widehat{p}(10)$ ranges over the set $\{0.5, 0.25, \ldots, 0.015, 0.01\}$. Our simulations thus include settings with good, moderate and extremely limited overlap. We also put $\mu_1(x) = x^{6/5}$, $\mu_0(x) = 1$, $\sigma_1^2(x) = 1 + 3^{x-9}$, and $\sigma_0^2(x) = 1$ for all $x \in \mathcal{X}$. We generate outcomes as $Y_i = \mu_{D_i}(X_i) + \sigma_{D_i}(X_i) \cdot \varepsilon_{D_i}(X_i)$, where $\varepsilon_d(x) \sim \mathcal{N}(0, 1)$.[9] In addition to the standard CI $\mathcal{I}_{S,1}$ and our robust CI $\mathcal{I}_{S,2}$, we also consider three further CIs: $\mathcal{I}_{S,3}$ is based on using the linear specification $\mu_d(x) = \beta_{0d} + \beta_{1d}x$ instead of a nonparametric specification for the outcome function;[10] $\mathcal{I}_{S,4}$ is constructed by approximating the distribution of T_n via the weighted bootstrap;[11] and $\mathcal{I}_{S,5}$ is an infeasible version of $\mathcal{I}_{S,1}$ that uses the true quantiles of the distribution of T_n (which are known in a simulation context) as critical values. The performance of the latter CI serves as a bound on what can potentially be achieved by feasible methods.

The left and right panel of Figure 1 show the finite sample coverage probabilities and corresponding average lengths, respectively, of the various CIs for the SATE as a function of $\widehat{p}(10)$. By construction, the infeasible CI $\mathcal{I}_{S,5}$ has exact coverage for all levels of overlap, and its average length serves as a benchmark for the other procedures. The coverage rate of the standard CI $\mathcal{I}_{S,1}$ is close to the nominal level for $\widehat{p}(10) \geq 0.05$, but heavily deteriorates for smaller values of $\widehat{p}(10)$, eventually deviating from the nominal level by about 17 percentage points. As suggested by its construction, the coverage probability of our robust CI $\mathcal{I}_{S,2}$

[9] To investigate the robustness of $\mathcal{I}_{S,2}$ against deviation from Assumption 2, we also ran simulations where the distribution of $\varepsilon_d(x)$ is a mixture of a standard normal and a standard exponential distribution centered at zero. That is, $\varepsilon_d(x) \sim \lambda \cdot \mathcal{N}(0, 1) + (1 - \lambda) \cdot (\text{Exp}(1) - 1)$, with $\lambda \in [0, 1]$ a mixture weight. Results with $\lambda = .5$ were virtually identical to those reported below, and are thus omitted.

[10] The unknown parameters are estimated by OLS, and a standard heteroskedasticity-robust variance estimate is used to construct the t-statistic of the corresponding treatment effect estimate. Note that this model is mildly misspecified among treated units.

[11] Bootstrap versions of the t-statistic T_n are created by assigning random weights $(\omega_1, \ldots \omega_n)$ to the observations, where $\omega_i = w_i - N_d(x)^{-1} \sum_{i \in \mathcal{M}_d(x)} w_i$ for $i \in \mathcal{M}_d(x)$ and the w_i are i.i.d. standard exponential. With this type of bootstrap every observation receives a positive weight, and we obtain positive within-cell sample variances in every bootstrap data set. If we were to use a bootstrap based on independent sampling of realized outcomes within covariate-treatment cells, this would often result in bootstrap data where the outcome variable only takes on a single value in some cells, and thus the t-statistic is not well-defined.

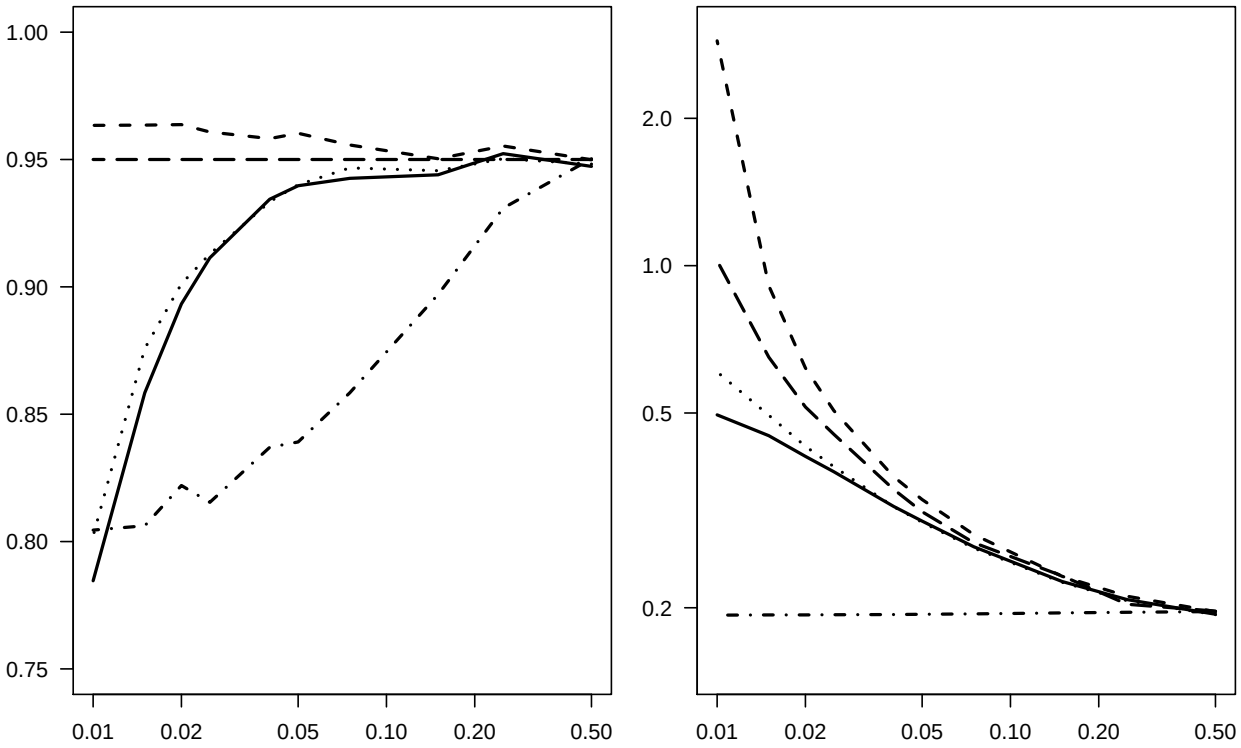

Figure 1: Empirical coverage probabilities (left panel) and average length (right panel) of $\mathcal{I}_{S,1}$ (standard; solid line), $\mathcal{I}_{S,2}$ (robust; short-dashed line), $\mathcal{I}_{S,3}$ (parametric; dot-dashed line), $\mathcal{I}_{S,4}$ (bootstrap; dotted line), and $\mathcal{I}_{S,5}$ (infeasible benchmark; long-dashed line) for values of $\widehat{p}(10)$ between 0.01 and 0.5 (or, equivalently, values of realized local sample size $N_1(10)$ between 2 and 100). Note that the horizontal axis of both plots is on a logarithmic scale.

is above the nominal level for all values of the propensity score. However, the deviations are rather minor, and do not exceed 1.3 percentage points even for the smallest value of $\widehat{p}(10)$. The average length of $\mathcal{I}_{S,2}$ is also very similar to that of the infeasible CI $\mathcal{I}_{S,5}$ for $\widehat{p}(10) \geq 0.05$, which implies that the added robustness comes at hardly any meaningful loss of power. The CI $\mathcal{I}_{S,3}$ based on a moderately misspecified regression function has poor properties for $\widehat{p}(10) \leq 0.25$, as low values of the propensity score amplify the misspecification bias. This shows that addressing limited overlap by imposing parametric restrictions will only work if this specification is very close to being correct. The bootstrap CI $\mathcal{I}_{S,4}$ has properties that only marginally improve upon those of $\mathcal{I}_{S,1}$. The superior higher-order properties of the

bootstrap under strong overlap thus have no impact on finite-sample performance in settings with limited overlap.

6. Extensions

6.1. Population Treatment Effects. The idea behind the construction of $\mathcal{I}_{S,2}$ can be extended to the PATE, which is arguably a more commonly used parameter in applications. As an estimator of τ_P, the asymptotic variance of $\widehat{\tau}$ is given by $\omega^2 = \omega_S^2 + \omega_P^2$, where $\omega_P^2 = \mathbb{E}((\tau(X) - \tau_P)^2)$. This parameter can be estimated by $\widehat{\omega}^2 = \widehat{\omega}_S^2 + \widehat{\omega}_P^2$, where $\widehat{\omega}_P^2 = \sum_j \widehat{f}(x_j)(\widehat{\tau}(x_j) - \widehat{\tau})^2$. The statistic $T_n = \sqrt{n}(\widehat{\tau} - \tau_P)/\widehat{\omega}$ can then be decomposed as

$$T_n = \frac{\widehat{\omega}_S}{\widehat{\omega}} \cdot T_{S,n} + \frac{\widehat{\omega}_P}{\widehat{\omega}} \cdot T_{P,n}, \quad \text{where} \quad T_{P,n} = \frac{\sqrt{n}(\tau_S - \tau_P)}{\widehat{\omega}_P}$$

and $T_{S,n}$ is as defined above. Under our assumptions $T_{S,n}$ and $T_{P,n}$ are asymptotically independent. Since $\tau_S - \tau_P = n^{-1}\sum_{i=1}^n \tau(X_i) - \mathbb{E}(\tau(X))$ does not involve the propensity score, the CLT approximation $P(T_{P,n} \leq u) \approx \Phi(u)$ should be accurate in large samples irrespective of the degree of overlap. In Section 3.2 we also showed that under Assumption 2 the finite sample distribution of $T_{S,n}$ given $\mathbf{M}_n$ can be approximated as $P(T_{S,n} \leq u|\mathbf{M}_n) \approx F_t(u/\rho_{\alpha(u)}, \delta_{\min})$, where $\alpha(u)$ is such that $\rho_{\alpha(u)} = u/c_\alpha(\delta_{\min})$. We can thus approximate the distribution of T_n by a (data-dependent) weighted mixture of $F_t(u/\rho_{\alpha(u)}, \delta_{\min})$ with a standard normal CDF. Specifically, for positive constants ω_1, ω_2, δ, and ρ we define the distribution function

$$G(u; \omega_1, \omega_2, \delta, \rho) = P\left(\frac{\omega_1 U_C(\delta, \rho) + \omega_2 V}{(\omega_1^2 + \omega_2^2)^{1/2}} \leq u\right),$$

where $U(\delta, \rho)$ and V are independent random variables such that $P(U(\delta, \rho) \leq u) = F_t(u/\rho, \delta)$ and $P(V \leq u) = \Phi(u)$. This CDF is difficult to tabulate, but it can easily be computed numerically or by simulation. Writing $g_\alpha(\delta, \rho) = G^{-1}(1 - \alpha/2; \widehat{\omega}_S, \widehat{\omega}_P, \delta, \rho)$ for $\alpha \in (0, .5)$, an

extension of $\mathcal{I}_{S,2}$ to inference on τ_P is given by

$$\mathcal{I}_{P,2} = \left(\widehat{\tau} - g_\alpha(\delta_{\min}, \rho_\alpha) \cdot \widehat{\omega}/\sqrt{n}, \widehat{\tau} + g_\alpha(\delta_{\min}, \rho_\alpha) \cdot \widehat{\omega}/\sqrt{n}\right).$$

This CI can be shown to be robust to limited overlap in a similar sense as $\mathcal{I}_{S,2}$ when the overall sample size is large. We omit a formal result in the interest of brevity.

6.2. Treatment Effects on the Treated. Our approach can easily be extended to the cases of the population and sample ATE on the treated (PATT and SATT, respectively). These alternative causal parameters are given, respectively, by

$$\tau_{P,T} = \mathbb{E}(Y(1) - Y(0)|D = 1) \quad \text{and} \quad \tau_{S,T} = \frac{1}{N_1} \sum_{i \in \mathcal{M}_1} \tau(X_i),$$

with $\mathcal{M}_1 = \{i : D_i = 1\}$ the set of the indices of those units that receive the treatment. Identification is achieved under a weaker version of Assumption 1 which only requires that (i) $Y(0) \perp D|X$ and (ii) $p(X) < 1$ with probability 1. Let $N_1 = \#\mathcal{M}_1$ denote the number of treated units, and put $\widehat{f}_1(x) = N_1(x)/N_1$, and $\widehat{\mu}_1 = N_1^{-1} \sum_{i \in \mathcal{M}_1}^n Y_i$. The natural estimator of both the PATT and the SATT is

$$\widehat{\tau}_T = \widehat{\mu}_1 - \sum_{j=1}^{J} \widehat{\mu}_0(x_j) \widehat{f}_1(x_j).$$

Conditional on $\mathbf{M}_n = \{(D_i, X_i)\}_{i=1}^n$, $\widehat{\tau}_T$ is a linear combination of $1 + J$ independent sample means. Since its structure is thus analogous to that of $\widehat{\tau}$, we can employ the same idea for constructing a robust CI. As an estimator of the SATT, the asymptotic variance of $\widehat{\theta}_T$ is given by $\omega_{S,T}^2 = \sigma_1^2/p_1 + \sum_j f_1(x_j)^2 \sigma_0^2(x_j)/p_0(x_j)$, where $\sigma_1^2 = \text{Var}(Y|D = 1)$ and $p_1 = P(D = 1)$. Now let $\widehat{\sigma}_1^2 = (N_1 - 1)^{-1} \sum_{i \in \mathcal{M}_1}^n (Y_i - \widehat{\mu}_1)^2$, $\widehat{p}_1 = N_1/n$, and define

$$\mathcal{I}_{S,2,T} = \left(\widehat{\tau}_T - c_\alpha(\delta_{\min})\rho_\alpha \cdot \widehat{\omega}_{S,T}/\sqrt{n}, \widehat{\tau}_T + c_\alpha(\delta_{\min})\rho_\alpha \cdot \widehat{\omega}_{S,T}/\sqrt{n}\right),$$

where $\delta_{\min} = \min\{\delta_1, \delta_{01}, \ldots, \delta_{0J}\}$, $\widehat{\omega}^2_{S,T}$ is the sample analogue of $\omega^2_{S,T}$,

$$\rho_\alpha = \left(\frac{(c_\alpha(\delta_1)/c_\alpha(\delta_{\min}))^2 \cdot \widehat{\sigma}_1^2/N_1 + \sum_j (c_\alpha(\delta_{0j})/c_\alpha(\delta_{\min}))^2 \cdot \widehat{f}_1(x_j)^2 \widehat{\sigma}_0^2(x_j)/N_0(x_j)}{\widehat{\sigma}_1^2/N_1 + \sum_j \widehat{f}_1(x_j)^2 \widehat{\sigma}_0^2(x_j)/N_0(x_j)} \right)^{1/2},$$

and $\delta_1 = N_1 - 1$. It then follows from arguments analogous to those used for $\mathcal{I}_{S,2}$ that under Assumption 2 we have $P(\tau_{S,T} \in \mathcal{I}_{S,2,T}) \geq 1 - \alpha$ in finite samples of any size. Other robustness properties carry over analogously as well.

7. Empirical Illustration

To illustrate the methods proposed in this paper, we reanalyze observational data from a well-known study by Connors et al. (1996) on the impact of right heart catheterization (RHC) on patient mortality. RHC is a diagnostic procedure used for critically ill patients, in which a thin tube is inserted into the right side of the heart to monitor its function. This information is then used by critical care physicians to determine the further course of treatment. The data used by Connors et al. (1996) contain information on 5735 patients. For each individual we observe the treatment status, where treatment is defined as RHC being applied within 24 hours of admission, the outcome, which is an indicator for survival at 30 days, and 50 covariates considered by a panel of experts to be related to the decision to perform the RHC. See Connors et al. (1996) for summary statistics and a more detailed description of the data. Using a propensity score matching approach, they reached the controversial conclusions that RHC causes a substantial increase in patient mortality.

For our analysis, we follow Hirano and Imbens (2001) and Crump et al. (2009) and first estimate the propensity score using a logistic model that includes all the covariates. Figure 2 shows the distribution of estimated propensity scores by treatment group. In both groups, the support of the estimated propensity scores is nearly the entire unit interval, and inference is thus potentially affected by limited overlap. Next, we partition the unit interval as (0,.05], (.05,.1],. . . ,(.95,1] into 20 cells, and discretize the estimated propensity score such that it

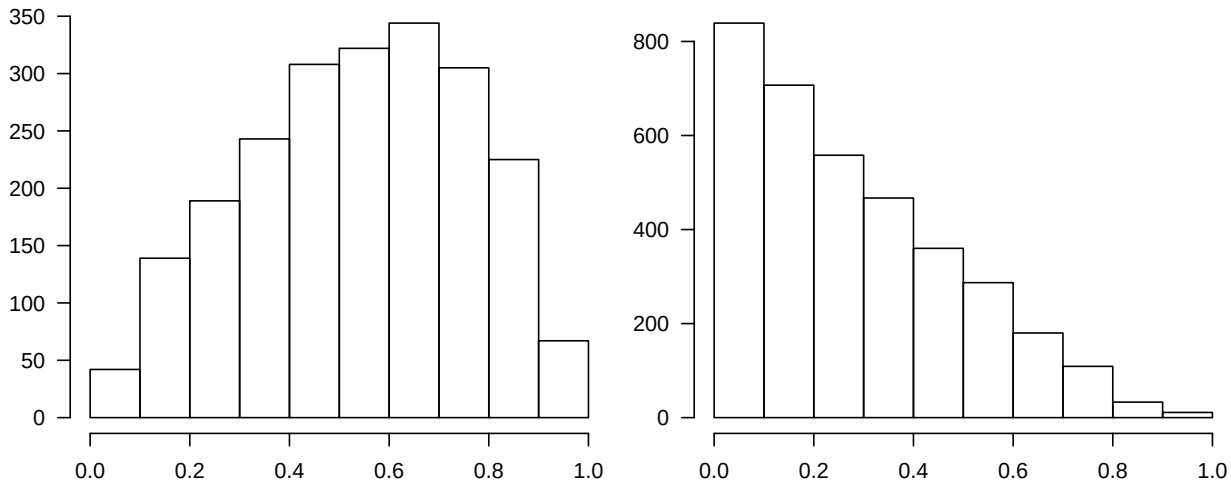

Figure 2: Histogram of the estimated propensity score among treated individuals (left panel) and untreated individuals (right panel).

takes the value j if the original estimate falls into the jth cell. We then estimate the SATE adjusting for between-group differences in the discretized propensity score. After computing the point estimate, we calculate both the classical CI $\mathcal{I}_{S,1}$ and our robust CI $\mathcal{I}_{S,2}$, with the nominal level being set to the usual 95%.[12]

Table 1 reports our empirical results. The point estimate of the SATE of RHC on patient mortality suggests an increase in the probability of death within 30 days of admission by about 4 percentage points, with a standard error of about 2.5 percentage points. When conducting inference on the SATE, our robust approach yields a critical value of 3.65, and thus $\mathcal{I}_{S,2}$ is about 1.85 times wider than the standard CI $\mathcal{I}_{S,1}$ based on the usual critical value 1.96. This discrepancy is mostly due to very small realized local sample sizes in two of the 40 "propensity score-treatment" cells resulting from our chosen partition of the unit interval. Both CIs contain the value of zero, suggesting that there is no strong evidence that

[12]We omit controlling for covariates within cells for simplicity, so the estimator corresponds to the one described in Section 4 with $K = 0$. Note that since inference on the SATE is conditional on the value of the covariates and the treatment indicator, no adjustments are necessary to account for the fact that an estimate of the propensity score is being used here. This is because the estimated propensity score is non-random given the original covariates and the treatment indicator. See Imbens (2015) for further details.

Table 1: Sample Average Treatment Effect of Right Heart Catheterization

Estimation Results:	Point Estimate	0.0398
	Standard Error	0.0252
95% Critical Value:	Standard	1.9600
	Robust	3.6451
95% Confidence Interval:	Standard	[-0.0096, 0.0893]
	Robust	[-0.0521, 0.1318]

RHC increases mortality.

8. Conclusions

Limited overlap creates a number of challenges for empirical studies that wish to conduct inference on the average effect of a treatment under the assumption of unconfounded assignment. This paper provides some new insights for why standard inference tends to be distorted under limited overlap, and proposes a new robust CI that has good theoretical and practical properties in empirically relevant settings. While formally derived in setting with discrete covariates, our empirical illustration shows how robust inference can be conducted in more general settings.

A. Proofs

A.1. Proof of Proposition 1. Put $\gamma_d(x) = \mathbb{E}((Y - \mu_d(x))^3 | D = d, X = x)$ and $\kappa_d(x) = \mathbb{E}((Y - \mu_d(x))^4 | D = d, X = x) - 3$ for all $(d, x) \in \{0, 1\} \times \mathcal{X}$. We then show that part (i) of the proposition holds with

$$\begin{aligned} q_2(t, f, p) = &\frac{t^3 - 3t}{6\omega_S^4} \cdot \sum_{d,j} \frac{f(x_j)\kappa_d(x_j)}{p_d(x_j)^3} - \frac{t^5 + 2t^3 - 3t}{9\omega_S^6} \cdot \left(\sum_{d,j} \frac{f(x_j)\gamma_d(x_j)(-1)^{1-d}}{p_d(x_j)^2} \right)^2 \\ &- \frac{t}{\omega_S^4} \cdot \sum_{(d,j) \neq (d',j')} \frac{\sigma_d^2(x_j)\sigma_{d'}^2(x_{j'})(f(x_j)p_d(x_j) + f(x_{j'})p_{d'}(x_{j'}))}{(p_d(x_j)p_{d'}(x_{j'}))^2} \\ &- \frac{(t^3 + 3t)}{2\omega_S^4} \cdot \sum_{d,j} \frac{f(x_j)\sigma_d^4(x_j)}{p_d(x_j)^3}, \end{aligned}$$

where $\omega_S^2 = \sum_{d,j} f(x_j)\sigma_d^2(x_j)/p_d(x_j)$ is as defined in the main body of the text. This follows from adapting a result of Hall and Martin (1988), who study the form of the Edgeworth

expansion of the two-sample t-statistic; see also Hall (1992). One only requires the insight that Hall and Martin's (1988) arguments remain valid if the number of samples is increased from 2 to $2J$. Denoting the distribution function of $T_{S,n}$ given $\mathbf{M}_n$ by $H_n(\cdot|\mathbf{M}_n)$, it follows from their reasoning that under the conditions of the proposition $H_n(\cdot|\mathbf{M}_n)$ satisfies the following Edgeworth expansion:

$$H_n(t|\mathbf{M}_n) = \Phi(t) + n^{-1/2}\phi(t)\widehat{q}_1(t) + n^{-1}\phi(t)\widehat{q}_2(t) + n^{-3/2}\phi(t)\widehat{q}_3(t) + O_P(n^{-2}),$$

where Φ and ϕ denote the standard normal distribution and density functions, respectively,

$$\begin{aligned}
\widehat{q}_1(t) &= \frac{2t^2+1}{6\bar{\omega}_S^3} \cdot \sum_{d,j} \frac{\widehat{f}(x_j)}{\widehat{p}_d(x_j)^2}\gamma_d(x_j), \\
\widehat{q}_2(t) &= \frac{t^3-3t}{12\bar{\omega}_S^4} \cdot \sum_{d,j} \frac{\widehat{f}(x_j)\kappa_d(x_j)}{\widehat{p}_d(x_j)^3} - \frac{t^5+2t^3-3t}{18\bar{\omega}_S^6} \cdot \left(\sum_{d,j} \frac{\widehat{f}(x_j)\gamma_d(x_j)(-1)^{1-d}}{\widehat{p}_d(x_j)^2} \right)^2 \\
&\quad - \frac{t}{2\bar{\omega}_S^4} \cdot \sum_{(d,j)\neq(d',j')} \frac{\sigma_d^2(x_j)\sigma_{d'}^2(x_{j'})(\widehat{f}(x_j)\widehat{p}_d(x_j) + \widehat{f}(x_{j'})\widehat{p}_{d'}(x_{j'}))}{(\widehat{p}_d(x_j)\widehat{p}_{d'}(x_{j'}))^2} \\
&\quad - \frac{(t^3+3t)}{4\bar{\omega}_S^4} \cdot \sum_{d,j} \frac{\widehat{f}(x_j)\sigma_d^4(x_j)}{\widehat{p}_d(x_j)^3},
\end{aligned}$$

$\bar{\omega}_S^2 = \sum_{d,j} \widehat{f}(x_j)\sigma_d^2(x_j)/\widehat{p}_d(x_j)$, and $\widehat{q}_3$ is another even function whose exact form is not important for the purpose of this argument. The conditional coverage probability of the CI $\mathcal{I}_{S,n}$ given $\mathbf{M}_n$ is given by

$$P(\tau_S \in \mathcal{I}_{S,n}|\mathbf{M}_n) = P(T_{S,n} \leq z_\alpha|\mathbf{M}_n) - P(T_{S,n} \leq -z_\alpha|\mathbf{M}_n) = H_n(z_\alpha|\mathbf{M}_n) - H_n(-z_\alpha|\mathbf{M}_n).$$

Substituting the Edgeworth expansion for $H_n(\cdot|\mathbf{M}_n)$ into this expression, we find that

$$P(\tau_S \in \mathcal{I}_{S,n}|\mathbf{M}_n) = 1 - \alpha + n^{-1}\phi(z_\alpha)\widehat{q}_2(z_\alpha) + O\left(n^{-2}\right),$$

The result of Proposition 1(i) then follows from the fact that $\mathbb{E}(\widehat{q}_2(z_\alpha)) = q_2(z_\alpha) + O(n^{-1})$, the relationship that $P(\tau_S \in \mathcal{I}_{S,n}) = \mathbb{E}(P(\tau_S \in \mathcal{I}_{S,n}|\mathbf{M}_n))$, and dominated convergence. The second part of the proposition follows from some simple algebra.

A.2. Proof of Proposition 2. To show part (i) we first prove the following auxiliary result, which is similar to a statement in Hayter (2014).

Lemma 1. *Let X be a standard normal random variable, and let $W = (a_1 W_1, \ldots, a_K W_K)'$ be a random vector with a_k a positive constant and W_k a random variable following a χ^2-distribution with s_k degrees of freedom for $k = 1, \ldots, K$, and such that X and the components of W are mutually independent. Also define the set $\Gamma = \{(\gamma_1, \ldots, \gamma_K) : \gamma_k \geq 0$ for $k = 1, \ldots, K$ and $\sum_{k=1}^K \gamma_k \leq 1\}$ with typical element γ, and let $V_\gamma = X/(W'\gamma)^{1/2}$. Then for all $\gamma \in \Gamma$ and $u > 0$ it holds that*

$$P(V_\gamma \leq u) \geq \min_{k=1,\ldots,K} F_t(u/a_k^{1/2}, s_k).$$

Proof. With Φ the CDF of the standard normal distribution and $u > 0$, the function $\Phi(ut^{1/2})$ is strictly concave in t for $t \geq 0$, as it is the combination of a strictly concave function and a strictly increasing function. Therefore it holds that

$$P(V_\gamma \leq u|W) = P(X \leq u(W'\gamma)^{1/2}|W) = \Phi(u(W'\gamma)^{1/2})$$

is a strictly concave function in γ for $\gamma \in \Gamma$ with probability one, and consequently

$$P(V_\gamma \leq u) = E(\Phi(u(W'\gamma)^{1/2}))$$

is strictly concave in γ for $\gamma \in \Gamma$. Since $P(V_\gamma \leq u)$ is also continuous in γ, and Γ is a convex compact set, the term $P(V_\gamma \leq u)$ attains a minimum in γ on the boundary of Γ. It remains to be shown that the minimum occurs for $\gamma = e_k$ for some k, where e_k denotes the K-vector whose kth entry is 1 and whose other entries are all 0. We prove this by induction. For $K = 1$ and $K = 2$ this is trivial, as the boundary of Γ only contains elements of the required form in those cases. For $K = 3$, the boundary of Γ is a triangle. If the minimum occurs on the side given by $\{(0, \gamma_2, \gamma_3) : \gamma_2, \gamma_3 \geq 0, \gamma_2 + \gamma_3 = 1\}$, it follows from the case $K = 2$ that

the minimum occurs for $\gamma = e_2$ or $\gamma = e_3$. By repeating this argument for the other sides of the triangle, it follows that the minimum must occur at $\gamma = e_k$ for some $k = 1, 2, 3$, which is what we needed to show. We then continue analogously for the cases $K = 4, 5, \ldots$, by always "going through" all $(K-1)$-dimensional "sides" of the K-dimensional simplex Γ. Since $P(V_{e_k} \leq u) = F_t(u/a_k^{1/2}, s_k)$, it then follows that $P(V_{e_k} \leq u) \geq \min_{k=1,\ldots,K} F_t(u/a_k^{1/2}, s_k)$. This completes the proof. □

The statement of part (i) of the proposition then follows from applying the Lemma to the conditional distribution of $T_{S,n}(h^*)$ given $\mathbf{M}_n$, by putting (with a slight abuse of notation)

$$X = \sqrt{n}(\widehat{\tau} - \tau_S) / \left(\sum_{d,j} c_\alpha(\delta_{dj})^2 \widehat{f}(x_j)^2 \sigma_d^2(x_j) / N_d(x_j) \right)$$

$$\gamma_k = (\widehat{f}(x_j)^2 \sigma_d^2(x_j) / N_d(x_j)) / \left(\sum_{d,j} \widehat{f}(x_j)^2 \sigma_d^2(x_j) / N_d(x_j) \right)$$

$$W_k = \widehat{\sigma}_d^2(x_j) / \sigma_d^2(x_j), \quad s_k = N_d(x_j) - 1, \quad \text{and } a_k = c_\alpha(\delta_{dj})^2,$$

and by noting that since the inequality holds conditional on $\mathbf{M}_n$ it must also hold unconditionally. Part (ii) follows from the fact that $c_\alpha(\delta) = z_\alpha + O(\delta^{-1})$, which implies that $c_\alpha(\delta_{\min}) = z_\alpha + O(n^{-1})$, and that $\rho_\alpha = 1 + O(n^{-1})$.

A.3. Proof of Corollary 1. The proof is analogous to that of Proposition 1, using standard results for homoskedastic linear models with normal errors.

References

Banerjee, S. K. (1960): "Approximate confidence interval for linear functions of means of k populations when the population variances are not equal," *Sankhya*, 22, 3.

Behrens, W. (1928): "Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen," *Landwirtschaftliche Jahrbücher*, 68.

Cattaneo, M., R. Crump, and M. Jansson (2013): "Generalized Jackknife Estimators

of Weighted Average Derivatives," *Journal of the American Statistical Association*, 108, 1243–1268.

Chaudhuri, S. and J. B. Hill (2014): "Heavy Tail Robust Estimation and Inference for Average Treatment Effects," *Working Paper*.

Cochran, W. G. (1968): "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics*, 295–313.

Connors, A. F., T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, et al. (1996): "The effectiveness of right heart catheterization in the initial care of critically III patients," *Journal of the American Medical Association*, 276, 889–897.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 1–13.

Fisher, R. (1935): "The fiducial argument in statistical inference," *Annals of Eugenics*, 6, 391–398.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*, Springer.

Hall, P. and M. Martin (1988): "On the Bootstrap and Two-Sample Problems," *Australian Journal of Statistics*, 30, 179–192.

Hayter, A. J. (2014): "Inferences on Linear Combinations of Normal Means with Unknown and Unequal Variances," *Sankhya*, 76-A, 1–23.

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

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

Ibragimov, R. and U. K. Müller (2016): "Inference with Few Heterogenous Clusters," *Review of Economics and Statistics*, 98, 83–96.

Imbens, G. (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and Statistics*, 86, 4–29.

——— (2015): "Matching methods in practice: Three examples," *Journal of Human Resources*, 50, 373–419.

Imbens, G., W. Newey, and G. Ridder (2007): "Mean-square-error calculations for average treatment effects," *Working Paper*.

Imbens, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

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

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

Linnik, Y. V. (1966): "Randomized homogeneous tests for the Behrens-Fisher problem," *Selected Translations in Mathematical Statistics and Probability*, 6, 207–217.

——— (1968): *Statistical problems with nuisance parameters*, American Mathematical Society.

Mickey, M. R. and M. B. Brown (1966): "Bounds on the distribution functions of the Behrens-Fisher statistic," *Annals of Mathematical Statistics*, 37, 639–642.

Robins, J. and Y. Ritov (1997): "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theroy for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319.

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

Rothe, C. and S. Firpo (2016): "Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically," *Working Paper*.

Salaevskii, O. (1963): "On the non-existence of regularly varying tests for the Behrens-Fisher problem," *Soviet Mathematics, Doklady*, 4, 1043–1045.

Yang, T. T. (2014): "Asymptotic Trimming and Rate Adaptive Inference for Endogenous Selection Estimates," *Working Paper*.