DISCUSSION PAPER SERIES

IZA DP No. 8758

Robust Confidence Intervals for Average Treatment Effects under Limited Overlap

Christoph Rothe

January 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Robust Confidence Intervals for Average Treatment Effects under Limited Overlap

Christoph Rothe
Columbia University
and IZA

Discussion Paper No. 8758
January 2015

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8758
January 2015

ABSTRACT

Robust Confidence Intervals for Average Treatment Effects under Limited Overlap*

Estimators of average treatment effects under unconfounded treatment assignment are known to become rather imprecise if there is limited overlap in the covariate distributions between the treatment groups. But such limited overlap can also have a detrimental effect on inference, and lead for example to highly distorted confidence intervals. This paper shows that this is because the coverage error of traditional confidence intervals is not so much driven by the total sample size, but by the number of observations in the areas of limited overlap. At least some of these "local sample sizes" are often very small in applications, up to the point where distributional approximation derived from the Central Limit Theorem become unreliable. Building on this observation, the paper proposes two new robust confidence intervals that are extensions of classical approaches to small sample inference. It shows that these approaches are easy to implement, and have superior theoretical and practical properties relative to standard methods in empirically relevant settings. They should thus be useful for practitioners.

JEL Classification: C12, C14, C25, C31

Keywords: average treatment effect, causality, overlap, propensity score, treatment effect heterogeneity, unconfoundedness

Corresponding author:

Christoph Rothe
Department of Economics
Columbia University
420 W 118th St.
New York, NY 10027
USA
E-mail: cr2690@columbia.edu

* I would like to thank Shakeeb Khan, Ulrich M¨uller, Bernard Salanie, and seminar audiences at Duke, Columbia and the 2014 Greater NY Metropolitan Area Colloquium at Princeton for their helpful comments.

1. Introduction

There are many empirical studies in economics whose goal it is to assess the effect of a binary treatment, such as the participation in an active labor market program, on some outcome of interest. The main empirical challenge in such studies is that differences in the outcomes of treated and non-treated units may not only be caused by the treatment, but can also be due to selection effects. Following the seminal work of Rubin (1974) and Rosenbaum and Rubin (1983), one important strand of the program evaluation literature addresses this issue by imposing the assumption that the treatment is unconfounded. This means that the selection into the treatment is modeled as being independent of the potential outcomes if certain observable covariates are held constant. A large number of estimators of average treatment effects that exploit this structure have been proposed in the literature, and these procedures have become increasingly popular in applications. See, among others, Hahn (1998), Heckman, Ichimura, and Todd (1998), Hirano, Imbens, and Ridder (2003), Abadie and Imbens (2006), Imbens, Newey, and Ridder (2007) or Chen, Hong, and Tarozzi (2008); and Imbens (2004) or Imbens and Wooldridge (2009) for comprehensive surveys.

A common concern for empirical practice is that these estimators can become rather imprecise if there are regions of the covariate space with only few observations in either the treatment or the non-treatment group. Such areas of *limited overlap* naturally arise if the overall sample size is relatively small to begin with. However, they can also occur in very large samples if the propensity score, which is defined as the conditional probability of taking the treatment given the covariates, takes on values that are close to either 0 or 1 (relative to the sample size). Since the variance of treatment effect estimators generally depends inversely on these conditional treatment and non-treatment probabilities, it can potentially be very large in this case. Moreover, Khan and Tamer (2010) show that if propensity scores can become arbitrarily close to 0 or 1, nonparametric estimators of average treatment effects can exhibit irregular behavior, and might converge at rates slower than the usual parametric one; see

also Khan and Nekipelov (2013) and Chaudhuri and Hill (2014). Appropriate overlap is thus important for obtaining precise point estimates of average treatment effects, and this fact seems to be widely appreciated by practitioners (see Imbens, 2004, Section 5.C, for example).

A more subtle issue that has received relatively little attention in the literature is that limited overlap also has a detrimental effect on inference. If propensity scores are close to 0 or 1, average treatment effects are only *weakly identified*, in the sense that the data generating process is close to one in which point identification fails. Weak identification of the parameter of interest is known to cause problems for inference in many econometric models, such as instrumental variables regressions with weak instruments (e.g. Staiger and Stock, 1997). For the treatment effect models with unconfoundedness, similar problems occur. For example, Kahn and Tamer's (2010) results imply that nonparametric estimators of treatment effects may no longer be $\sqrt{n}$-consistent and asymptotically normal if propensity scores are arbitrarily close to 0 or 1, and thus the justification for the usual 95% confidence interval of the form "point estimate$\pm$1.96$\times$standard error" breaks down. By extension, one should also be concerned about the accuracy of such a confidence interval in applications where the propensity score is bounded away from 0 and 1, but only by a relatively small constant. In a simulation study reported in this paper, we demonstrate that the *actual* coverage probability of such a confidence interval can indeed be substantially below the nominal level of 95%, making estimates seem more precise than they are. It is important to note that this phenomenon cannot be explained by the fact that standard errors are generally larger under limited overlap, as this by itself only affects the length but not the coverage probability of the confidence interval. Roughly speaking, under limited overlap standard confidence intervals tend to be "too short" even though they are typically rather wide to begin with.

This paper explores the channels through which limited overlap affects the accuracy of inference, and provides some practical solutions to the challenges created by this issue. We

begin by considering a "small" nonparametric model in which the covariates have known finite support. This benchmark setup has the advantage that many commonly used estimation strategies are numerically identical here. Our first main contribution is then to show that the order of the coverage error of a standard confidence interval is not driven by the total 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 of such a confidence interval can be substantial. Inference on average treatment effects under limited overlap is therefore essentially a problem of *locally small sample sizes*, even if the overall sample size is large. The issue is thus conceptually quite different from the problems for inference caused by weak identification in other econometric models, such weak IV models. To the best of our knowledge, this paper is the first to formally make this point.

Moving from a description towards a solution of the problem, we consider the construction of robust confidence intervals that maintain approximately correct coverage probability by adapting automatically to the degree of overlap in a data-driven way. Given our previous analysis, the usefulness of traditional first order *large sample* arguments for addressing this issue seems limited at best. We therefore do not pursue an approach based on a drifting sequence of propensity scores, for example. Instead, we propose to extend classical methods that were specifically designed for *small sample* inference to our setting. This is the second main contribution of this paper. We exploit the fact that with discrete covariates the estimators of average treatment effects take the form of a linear combination of independent sample means. Inference on treatment effects can thus be thought of as a generalized version of the Behrens-Fisher problem (Behrens, 1928; Fisher, 1935), which has a long tradition in the statistics literature.

We consider two different ways to construct robust confidence intervals for average treatment effects. Both are formally derived under the assumption that the data are distributed as a scale mixture of normals. While this class contains a wide range of continuous, unimodal

and symmetric distributions, the condition is clearly restrictive and somewhat unusual in the context of nonparametric treatment effect inference. We treat it as an auxiliary assumption for the construction of our robust confidence intervals, in the sense that these procedures will have good properties if the condition is either literally or approximately satisfied, and are at least not going to be invalid in a classical sense if this assumption is violated. Without some restrictions of the data distribution 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, as first-order asymptotic approximations are going to be unreliable.

Our first approach to constructing a robust confidence intervals can be interpreted as bounding the distribution of the studentized estimator by a "squeezed" version of a t distribution with degrees of freedom equal to the number of observations in the smallest covariate-treatment cell minus one. This then leads to a conservative confidence interval that is valid for any finite sample size irrespective of the degree of overlap (under our distributional assumption). This type of inference is similar in spirit to that in Ibragimov and Müller (2013), although the problem considered by them is quite different.

Our second approach is to approximate the distribution of the studentized estimate by a t distribution with data-dependent degrees of freedom determined by the Welch-Satterthwaite formula (Welch, 1938, 1947; Satterthwaite, 1946). For the case of two cells, this approximation has long been known to be very accurate even for relatively small group sizes (e.g. Wang, 1971; Lee and Gurland, 1975; Best and Rayner, 1987), and our simulations suggest that this also extends to settings with a larger number of cells. We also show that this approach formally leads to a higher-order asymptotic correction in the coverage error of the confidence interval, although in practice it seems to work better than this result alone would suggest.

Our proposed confidence intervals are easy to implement as they both take the familiar

form "point estimate±critical value×standard error", where only the critical value differs relative to the usual construction. No modifications of the treatment effect estimator or the estimate of its variance are required, and no additional tuning parameters need to be chosen. The critical values are adaptive in the sense that they are larger for data sets where overlap is more severely limited, thus providing an accurate reflection of sampling uncertainty in such settings. At the nominal 95% level, for example, our robust confidence intervals can potentially be up to six and a half times longer than the traditional one that uses the critical value 1.96, although in empirical applications an increase in length by 10%-30% seems to be more typical. Under strong overlap, our robust intervals are virtually identical to the traditional one if the overall sample size is large.

The third main contribution of this paper is to show how to extend our methods to "large" nonparametric models with continuously distributed covariates. Here the main idea is that the techniques we developed for the discrete case can be applied with very little modification if treatment effects are estimated by first partitioning the covariate space into a finite number of cells, and then fitting a constant or some higher-order polynomial within each element of the partition by least squares. Such an approach is often referred to as *partitioning regression*. See Györfi, Krzyzak, Kohler, and Walk (2002) for a textbook treatment, and Cattaneo and Farrell (2011, 2013) for some recent applications in econometrics. In this case, our approach yields robust confidence intervals for the sum of the true treatment effect and the bias resulting from the "piecewise constant" or "piecewise polynomial" approximation. This bias can be made negligible for practical purposes by choosing the partition and the order of the local polynomial appropriately.

In a simulation study, we show that our construction leads to confidence intervals with good finite-sample coverage properties under limited overlap, even if the auxillary assumption about the data distribution is substantially violated. We also apply our methods to the National Supported Work (NSW) demonstration data, analyzed originally by LaLonde

(1986). There we show that for a partition chosen by a modern machine learning algorithm our methods suggest confidence intervals that are up to about 15% wider than the standard one, which shows the practical relevance of our correction.

In empirical practice, concerns about limited overlap are commonly addressed by redefining the population of interest, i.e. by estimating the average treatment effect only for that part of the population with propensity scores that are well-separated from the boundaries of the unit interval; see for example Crump, Hotz, Imbens, and Mitnik (2009). Our robust confidence intervals should be seen as a complement to such an approach, and not as a replacement. Trimming observations with very low or very high propensity scores has the advantage that the resulting redefined average treatment effect parameter can typically be estimated with greater precision, and there are no concerns about the validity of standard confidence intervals 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. If the entire population is of policy relevance, trimming therefore introduces a bias. Since observations are sparse in the trimmed areas by construction, the magnitude of this bias is difficult to determine from the data.[1] In an empirical application with limited overlap, it would therefore seem reasonable to present point estimates and confidence intervals for both a trimmed and the original population, thus offering readers a more nuanced view of the informational content of the data.

The remainder of this paper is structured as follows. In Section 2, we introduce the basic setup. In Section 3, we show the detrimental effect of limited overlap on the performance of standard methods for inference in a setting with discrete covariates. In Section 4, we propose two new robust confidence intervals. In Section 5, we extend our approach to settings with

[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.

continuously distributed covariates. Section 6 contains the result of a simulation study and an empirical illustration using the well known LaLonde data. Finally, Section 7 concludes. All proofs are collected in the appendix.

2. The Basic Setup

2.1. Model. We are interested in determining the causal effect of a binary treatment on some economic outcome. Let D be a treatment indicator that takes the value 1 if the treatment is received, and 0 otherwise. Define $Y(1)$ as the potential outcome of an individual if the treatment is imposed exogenously, and $Y(0)$ as the corresponding potential outcome in the absence of the treatment. The realized outcome is then given by $Y \equiv Y(D)$. Also, let X be a vector of covariates measured prior to the treatment. The analyst observes n realizations of (Y, D, X), where one should think of n as a relatively large integer. We make the following assumption about the sampling scheme.

Assumption 1 (Sampling)**.** *The data* $\{(Y_i, D_i, X_i) : i \leq n\}$ *are an independent and identically distributed sample from the distribution of the random vector* (Y, D, X)*.*

There are several parameters that can be used to summarize the distribution of individual level causal effects $Y(1)-Y(0)$ in this context. We primarily focus on the population average treatment effect (PATE) and sample average treatment effect (SATE), which are given by

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

respectively. Here $\tau(x) \equiv \mathbb{E}(Y(1) - Y(0)|X = x)$ is the conditional average treatment effect (CATE) given X.[2] However, the analysis in this paper can easily be extended to other common estimands, such as the population and sample average treatment effect on

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

the treated. See Imbens (2004) for a discussion of these and other related estimands. In the following, we use the notation that $\mu_d(x) \equiv \mathbb{E}(Y|D=d, X=x)$ and $\sigma_d^2(x) \equiv \text{Var}(Y|D=d, X=x)$. We refer to $p_d(x) \equiv P(D=d|X=x)$ as the *generalized propensity score* (GPS), and write $p(x) \equiv p_1(x)$ for the "ordinary" propensity score.

Throughout the paper, we maintain the *ignorability condition* of Rosenbaum and Rubin (1983), which asserts that conditional on the covariates, the treatment indicator is independent of the potential outcomes, and that the distribution of the covariates has the same support among the treated and the untreated. These conditions are strong and arguably not realistic in certain empirical settings; but see Imbens (2004) for a discussion of their merit in those cases. They can be stated formally as follows:

Assumption 2 (Unconfoundedness)**.** $(Y(1), Y(0)) \perp D|X$.

Assumption 3 (Overlap)**.** $0 < p(X) < 1$ *with probability 1.*

Under Assumptions 2–3 the conditional average treatment effect $\tau(x)$ is identified from the joint distribution of (Y, D, X) over the entire support of X through the relationship $\tau(x) = \mu_1(x) - \mu_0(x)$. The population and sample average treatment effects can then be identified by averaging $\tau(x)$ over the population and sampling distribution of X, respectively:

$$\tau_P = \mathbb{E}(\tau(X)) \quad \text{and} \quad \tau_S = \frac{1}{n}\sum_{i=1}^{n} \tau(X_i). \tag{2.1}$$

See e.g. Imbens (2004) for an overview of other representations of average treatment effects in terms of the distribution of observable quantities, such as inverse probability weighting.

2.2. Estimation, Inference, and the Overlap Condition. Estimators of the PATE that are semiparametrically efficient under Assumptions 1–3 and 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 3 is strengthened to:

$$\epsilon < p(X) < 1 - \epsilon \text{ with probability 1 for some } \epsilon > 0. \tag{2.2}$$

Condition (2.2) is often referred to as *strong overlap* in the literature. Khan and Tamer (2010) show that without this condition the semiparametric efficiency bound for estimating τ_P or τ_S is not finite, and thus no $\sqrt{n}$-consistent and asymptotically normal ($\sqrt{n}$-CAN) semiparametric estimator of these parameters exists. This has important implications for empirical practice, because it does not only imply that standard estimators might have poor finite sample properties, but potentially also a failure of commonly used methods for inference that build on these estimators. For example, if (2.2) does not hold, the actual coverage probability of a standard confidence interval of the form "point estimate$\pm$1.96$\times$standard error" can differ substantially from its 95% nominal level even if the available sample is very large, because the formal justification of such confidence intervals is precisely a "$\sqrt{n}$-CAN"-type result.

By extension, one would also be concerned that standard inference could be unreliable if (2.2) only holds for some $\epsilon > 0$ that is very small relative to the sample size (in some appropriate sense). We will be particularly concerned with this case in our paper, and will informally refer to such a setting where the generalized propensity score takes on values that are "close" to but bounded away from 0 as having *limited overlap*. While $\sqrt{n}$-CAN estimators formally exist in such settings, one would expect methods for inference justified by this property to perform rather poorly.

2.3. A Simple Estimator. Our aim in this paper is to provide some further insights into why exactly limited overlap causes problems for inference, and to derive simple confidence intervals that have good coverage properties in finite samples if (2.2) holds for some $\epsilon > 0$

that is very close to zero. To do this, we begin with adopting the assumption that the covariates X have known finite support, and denote the corresponding probability density function by $f(x)$.

Assumption 4 (Finite Support)**.** *The distribution of X has finite support $\mathcal{X} = \{x_1, \ldots, x_J\}$ and probability density function $f(x) = P(X = x)$.*

Assumption 4 is a modeling device that will simplify the following theoretical arguments. The condition is not overly restrictive as any continuous distribution can be arbitrarily well approximated by a discrete one with J large enough.[3] Our main motivation for using this setup is that with discrete covariates most popular estimators of average treatment effects, including those proposed by Hahn (1998), Hirano et al. (2003) and Imbens et al. (2007), are all numerically identical. This shows that the complications caused by limited overlap are not specific to a particular estimation strategy. Finally, our proposed solution for improving inference under limited overlap, which we present in Section 4, will be motivated by a setting with discrete covariates, although we also discuss an extension of our method to settings with continuously distributed covariates in Section 5.

Next, we introduce some additional notation. For $d \in \{0, 1\}$ and $x \in \mathcal{X}$, 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 will refer to $N_d(x)$ as the *realized local sample size* at (d, x) in the following. Another quantity that will be of central importance is the *expected local sample size* at (d, x), defined as

$$n_d(x) \equiv \mathbb{E}(N_d(x)).$$

This is the number of observations we expect to observe in any given covariate-treatment

[3]That is, if X contains some continuously distributed components, we can form cells over their support, discretize the data, and re-define X accordingly. While such a discretization results in a bias in the estimator $\widehat{\tau}$, this bias can be made small by choosing a suitably fine partition of the support. We discuss this issue more formally in Section 5 below.

cell. Note that in our setup we have that

$$n_d(x) = nf(x)p_d(x).$$

With this notation, the natural estimators of the density function $f(x)$ and the generalized propensity score $p_d(x)$ are

$$\widehat{f}(x) = \frac{N(x)}{n} \quad \text{and} \quad \widehat{p}_d(x) = \frac{N_d(x)}{N(x)},$$

respectively, and we write $\widehat{p}(x) = \widehat{p}_1(x)$ for the estimate of the usual propensity score. We also define estimators of the conditional expectation $\mu_d(x)$ and the conditional average treatment effect $\tau(x)$ as

$$\widehat{\mu}_d(x) = \frac{1}{N_d(x)} \sum_{i \in \mathcal{M}_d(x)} Y_i \quad \text{and} \quad \widehat{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x).$$

The natural estimator of both the PATE and the SATE is then given by

$$\widehat{\tau} = \sum_{j=1}^{J} \widehat{f}(x_j)\widehat{\tau}(x_j) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}(X_i).$$

Note that while this estimator is expressed as a sample analogue of the moment condition (2.1) here, with discrete covariates this estimator is actually numerically identical to other popular estimators based on sample analogues of alternative representations of average treatment effects. For example, our estimator could also be written in "inverse probability weighting" form as $\widehat{\tau} = n^{-1} \sum_{i=1}^{n} Y_i(D_i - \widehat{p}(X_i)) \cdot (\widehat{p}(X_i)(1 - \widehat{p}(X_i)))^{-1}$, as in Hirano et al. (2003). We use the expression given above merely for notational convenience.

3. The Impact of Limited Overlap

Conventional estimators of average treatment effects can have large variances under limited overlap, and can thus be rather imprecise in finite samples (e.g. Imbens, 2004; Crump et al.,

2009). While large variances are of course undesirable from an empirical point of view, their presence alone does not cause the usual methods for inference to break down. Generally speaking, even if the variance of some parameter estimate is large, a confidence interval constructed by inverting the decision of the corresponding t-test should still have approximately correct coverage probability; it will just be rather wide. We now show that the situation is different for treatment effect estimation under limited overlap. In particular, we argue that low values of the generalized propensity score have a strongly detrimental effect on the coverage error of standard confidence intervals.

To understand the nature of the problems for inference caused by limited overlap, consider the task of deriving a confidence interval for the SATE τ_S. Under our Assumptions 1–4, it would seem that this can formally be done in the usual way, as the estimator $\widehat{\tau}$ has standard asymptotic properties. In particular, as $n \to \infty$, we have that

$$\sqrt{n}(\widehat{\tau} - \tau_S) \xrightarrow{d} \mathcal{N}\left(0, \omega_S^2\right) \quad \text{with} \quad \omega_S^2 \equiv \mathbb{E}\left(\frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)}\right).$$

In our setup with discrete covariates, an equivalent expression for the asymptotic variance ω_S^2 is given by

$$\omega_S^2 = \sum_{d,j} \frac{f(x_j)}{p_d(x_j)} \cdot \sigma_d^2(x_j).$$

This representation shows that low generalized propensity scores will drive up the value of ω_S^2 if they occur in areas where the covariate density is (relatively) high. The asymptotic variance ω_S^2 can then be estimated consistently by

$$\widehat{\omega}_S^2 = \sum_{d,j} \frac{\widehat{f}(x_j)}{\widehat{p}_d(x_j)} \widehat{\sigma}_d^2(x_j),$$

where

$$\widehat{\sigma}_d^2(x) = \frac{1}{N_d(x) - 1} \sum_{i \in \mathcal{M}_d(x)} (Y_i - \widehat{\mu}_d(x))^2$$

is the natural estimator of $\sigma_d^2(x)$. 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 that event taking place. We then find that as $n \to \infty$ the studentized version of our estimator is asymptotically standard normal; that is

$$T_{S,n} \equiv \frac{\sqrt{n}(\widehat{\tau} - \tau_S)}{\widehat{\omega}_S} \xrightarrow{d} \mathcal{N}(0,1). \tag{3.1}$$

A result like (3.1) would then commonly used to justify a Gaussian approximation to the sampling distribution of $T_{S,n}$, that is $P(T_{S,n} \leq c) \approx \Phi(c)$, which in turn justifies the usual two-sided confidence interval for τ_S with nominal level $1 - \alpha$:

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

where $z_\alpha = \Phi^{-1}(1 - \alpha/2)$. The next proposition studies the coverage properties of this confidence interval.

Proposition 1. *Suppose that Assumptions 1–4 hold, and 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* $(d, x) \in \{0, 1\} \times \mathcal{X}$. *Under regularity conditions (Hall and Martin, 1988; Hall, 1992), it holds that*

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

where ϕ denotes the standard normal density function,

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

and $\omega_S^2 = \sum_{d,j} f(x_j)\sigma_d^2(x_j)/p_d(x_j)$ is as defined above.

Proposition 1 follows from a standard Edgeworth expansion of the distribution of $T_{S,n}$ (Hall and Martin, 1988; Hall, 1992). Formally, the coverage error of $\mathcal{I}_{S,1}$ is of the order n^{-1}, which is the order we generally expect for confidence intervals of this type based on a regular parametric estimator (Hall, 1992). However, such an interpretation of the result can be misleading in finite samples of any size, as both the covariate density $f(x)$ and the generalized propensity score $p_d(x)$ strongly affect the constant associated with this rate. For any fixed sample size n, there exist data generating processes for which this constant, and thus the coverage error, can be very large. The following result shows that it is therefore better to think of the accuracy of $\mathcal{I}_{S,1}$ as not being driven by the *total* sample size n, but but the *expected local* sample sizes.

Proposition 2. *Recall that $n_d(x) = nf(x)p_d(x)$, and consider a sequence of covariate densities $f(x)$ and generalized propensity scores $p_d(x)$ such that $\min_{d,x} n_d(x) \to \infty$ as $n \to \infty$. Then it holds that*

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

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

Proposition 2 derives an approximation to the leading term $n^{-1}\phi(z_\alpha)q_2(z_\alpha)$ of the Edgeworth expansion in Proposition 1 that allows for the possibility that at least some values of

the generalized propensity score are close to 0. It shows that in practice the accuracy of the interval $\mathcal{I}_{S,1}$ is effectively similar to that of a confidence interval computed from a sample of the *expected local sample size* $n_{d^*}(x^*)$ in the covariate-treatment cell where the ratio of the generalized propensity score and the covariate density takes its *smallest* value, instead of size n. Under limited overlap, where $p_{d^*}(x^*)$ is potentially small, the local sample size $n_{d^*}(x^*) = nf(x^*)p_{d^*}(x^*)$ can easily be of an order of magnitude at which asymptotic approximations based on the Central Limit Theorem and Slutsky's Theorem are deemed unreliable. As a consequence, the probability that τ_S is contained in $\mathcal{I}_{S,1}$ can deviate substantially from the nominal level $1-\alpha$ even if the overall sample size n is very large. This is an important practical impediment for valid inference under limited overlap.

The proposition also shows that low generalized propensity scores are not problematic for inference by themselves, but only if they occur in areas where the covariate density is (relatively) high. This is because inference is based on a *density weighted* average of the sample means in each covariate- treatment cell. Even if some local sample size is small, the resulting uncertainty is dampened if the corresponding density weight is small as well. This mirrors the structure of the asymptotic variance discussed above. A mere inspection of the generalized propensity score alone does therefore in general not conclusively indicate whether standard confidence intervals are likely to be misleading; one would have to study the covariate density as well to make this determination.

A result analogous to Propositions 1–2 could also be derived for confidence intervals for the PATE, but we omit the details in the interest of brevity. To sketch the argument, note that $\widehat{\tau} - \tau_P = (\widehat{\tau} - \tau_S) + (\tau_S - \tau_P)$. and that the two terms in this decomposition are asymptotically independent. Moreover, the first term is the one we studied above, and the second term $\tau_S - \tau_P = n^{-1}\sum_{i=1}^{n}\tau(X_i) - \mathbb{E}(\tau(X))$ is simply a sample average of n random variables with mean zero and finite variance that does not depend on the propensity score. This term is therefore unproblematic, as its distribution can be well approximated

by a Gaussian one irrespective of the degree of overlap. Taken together, the accuracy of a Gaussian approximation to the sampling distribution of a studentized version of $\widehat{\tau} - \tau_P$ will be driven by the accuracy of such an approximation to the studentized version of $\widehat{\tau} - \tau_S$, and this result carries over to the corresponding confidence intervals.

4. Robust Confidence Intervals under Limited Overlap

The result of the previous section shows that inference on average treatment effects under limited overlap is essentially a *small sample* problem, even if the overall sample size n is large. For this reason, traditional arguments based on first order *large sample* approximations seem not very promising for addressing this issue. In this section, we therefore argue in favor of alternative approaches to constructing confidence intervals, which are based on extending classical methods specifically devised for small sample inference to our setting.

4.1. Robust Confidence Intervals for the SATE. As in the previous section, we begin by studying inference on the SATE. Robust confidence intervals for the PATE can be derived similarly, as discussed in Section 4.2.

4.1.1. Preliminaries. To motivate our approach, consider the simple case in which the covariates X are absent from the model, and the data are thus generated from a randomized experiment. In this case, the statistic $T_{S,n}$ defined in (3.1) is analogous to the test statistic of a standard two-sample t-test. Indeed, conditional on the number of treated and untreated individuals, inference on τ_S reduces to the Behrens-Fisher problem (Behrens, 1928; Fisher, 1935), i.e. the problem of conducting inference on the difference of the means of two populations with unknown and potentially different variances. Our setting with covariates can be thought of as a generalized version of the Behrens-Fisher problem, since conditional on the

set $M = \{(X_i, D_i), i \leq n\}$ of treatment indicators and covariates,[4] 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 $\widehat{f}(x)^2\sigma_d^2(x)$. The advantage of taking this point of view is that there is a longstanding literature in statistics that has studied solutions to Behrens-Fisher-type problems with relatively small group sizes. Instead of relying on first-order asymptotic theory, this literature exploits assumptions about the distribution of the data. Our aim is to extend some of these approaches to the context of treatment effect estimation under limited overlap. To this end, we introduce the following auxiliary assumption.

Assumption 5 (Data Distribution)**.** $Y(d) = \mu_d(X) + \sigma_d(X) \cdot \varepsilon_d(X) \cdot \eta_d(X)$*, where* $\varepsilon \equiv \{\varepsilon_d(x) : (d,x) \in \{0,1\} \times \mathcal{X}\}$ *is a collection of standard normal random variables,* $\eta \equiv \{\eta_d(x) : (d,x) \in \{0,1\} \times \mathcal{X}\}$ *is a collection of positive random variables with unit variance, and the components of* ε *and* η *are all independent of the data and of each other.*

Assumption 5 states that $Y|D, X$ is distributed as a scale mixture of normals,[5] which is clearly restrictive. Still, this assumption covers a wide class of continuous, unimodal and symmetric distributions on the real line, which includes the normal distribution, discrete mixtures and contaminated normals, the Student t family, the Logistic distribution, and the double-exponential distribution, among many others. We will use this condition to construct confidence intervals for average treatment effects that are robust to limited overlap, in the sense that they have good properties if Assumption 5 is either literally or at least approximately satisfied. While derived under a distributional assumption, these confidence intervals are not going to be invalid if this assumption is violated, in the sense that they will at least not be worse than the traditional confidence interval $\mathcal{I}_{S,1}$ in such settings. One

[4]Note that the set $\{N_d(x) : (d,x) \in \{0,1\} \times \mathcal{X}\}$ of realized local sample sizes would be a sufficient statistic for M in the following context.

[5]The distribution of a generic random variable $Z = A \cdot B$ is referred to as a scale mixture of normals if A follows a standard normal distribution, B is a strictly positive random variable, and A and B are stochastically independent.

can therefore think of Assumption 5 as an *asymptotically irrelevant parametrization*, in the sense that results obtained without this condition via standard asymptotic arguments do not change if this assumption holds.[6]

4.1.2. A Conservative Approach. Our first approach is to construct a confidence interval for the SATE that is guaranteed to meet the specified confidence level in any finite sample under Assumption 5. The price one has to pay for this desirable property is that the resulting interval will generally be conservative.

Let $c_\alpha(\delta) = F_t^{-1}(1-\alpha/2, \delta)$, where $F_t(\cdot, \delta)$ denotes the CDF of Student's t-distribution with δ degrees of freedom, and put $\delta_{dj} = N_d(x_j) - 1$ and $\delta_{\min} = \min_{d,j} \delta_{dj}$ for notational simplicity. The studentized statistic $T_{S,n}$ would seem like a natural starting point for the construction of a confidence interval, but it will be beneficial to begin with considering the larger class of test statistics of the form

$$T_{S,n}(h) = \frac{\sqrt{n}(\widehat{\tau} - \tau_S)}{\widehat{\omega}_S(h)},$$

where

$$\widehat{\omega}_S^2(h) = \sum_{d,j} h_{dj} \cdot \frac{\widehat{f}(x_j)}{p_d(x_j)} \cdot \widehat{\sigma}_d^2(x_j)$$

and $h = \{h_{dj} : d = 0, 1; j = 1, \ldots, J\}$ is a vector of $2J$ positive constants. Our statistic $T_{S,n}$ is obtained by setting $h \equiv 1$. From an extension of the argument in Mickey and Brown (1966) similar to that in Hayter (2014), it follows that for every $u > 0$ and every vector h

$$P(T_{S,n}(h) \leq u | M, \eta) \geq \max_{d,j} F_t(u h_{dj}^{1/2}, \delta_{dj});$$

see the appendix. This lower bound on the CDF of $T_{S,n}(h)$ translates directly into a bound

[6]This would be the case for the normality result in (3.1) or Propositions 1–2, for example. Note that since the distribution of $Y|D, X$ is symmetric under Assumption 5, the summands in the definition of $q_2(t)$ in Proposition 1 that involve $\gamma_d(x)$ will vanish, but the order of the coverage error and the statement of Proposition 2 remain the same in this case.

on its quantiles, which in turn motivates conservative confidence intervals with nominal level $1-\alpha$ of the form

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

It is easily verified that the length of such a confidence interval is minimized by putting $h_{dj}^{1/2} \propto c_\alpha(\delta_{dj})$ for all (d, j). Denoting such a choice of h by h^*, we obtain the "optimal" conservative confidence interval within this class as

$$\mathcal{I}_{S,2} = \left(\widehat{\tau} - \frac{\widehat{\omega}_S(h^*)}{\sqrt{n}}, \widehat{\tau} + \frac{\widehat{\omega}_S(h^*)}{\sqrt{n}}\right).$$

This confidence interval can be expressed in the more familiar form "point estimate±critical value×standard error" as

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

where

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

For numerical purposes, this confidence interval can heuristically be interpreted as being derived from a (conservative) approximation to the distribution of the usual t-statistic $T_{S,n}$; namely that $P(T_{S,n} \leq u|M, \eta) \approx F_t(u/\rho_{\alpha(u)}, \delta_{\min})$, where $\alpha(u)$ solves $u/c_\alpha(\delta_{\min}) = \rho_a$ in a. This view will be helpful when extending this approach to confidence intervals for the PATE in the following section.

In contrast to $\mathcal{I}_{S,1}$, the new interval adapts automatically to the severity of the issue of limited overlap. One can show that $c_\alpha(\delta_{\min})\rho_\alpha \geq c_\alpha(n-2J) > z_\alpha$, and hence by construction the new interval $\mathcal{I}_{S,2}$ is always wider than $\mathcal{I}_{S,1}$. If the generalized propensity score takes on values close to zero relative to the overall sample size, and thus the realized size of some local samples is likely to be small, the difference in length can be substantial. In the extreme case

where $\delta_{\min} = 1$, which is the smallest value for which our confidence intervals are numerically well-defined, the new interval could be up to six and a half times wider than the original one for $\alpha = .05$. This is because $c_\alpha(\delta_{\min})\rho_\alpha \leq c_\alpha(\delta_{\min})$, with equality if $\delta_{dj} = \delta_{\min}$ for all (d, j), and $c_\alpha(1)/z_\alpha = 6.48$ for $\alpha = .05$. On the other hand, if the propensity score, the covariate density and the overall sample size are such that $\delta_{\min}$ is larger than about 50 with high probability, the difference between $\mathcal{I}_{S,1}$ and $\mathcal{I}_{S,2}$ is not going to be of much practical relevance. This is because at conventional significance levels the quantiles of the standard normal distribution do not differ much from those of a t distribution with more than 50 degrees of freedom.

The next proposition formally shows that under Assumption 5 the interval $\mathcal{I}_{S,2}$ does not under-cover the parameter of interest in finite samples, and is asymptotically valid in a traditional sense if Assumption 5 does not hold.

Proposition 3. *(i) Under Assumptions 1–5, we have that*

$$P(\tau_S \in \mathcal{I}_{S,2}) \geq 1 - \alpha.$$

(ii) Under Assumptions 1–4 and the regularity conditions of Proposition 1, we have that

$$P(\tau_S \in \mathcal{I}_{S,2}) = P(\tau_S \in \mathcal{I}_{S,1}) + O(n^{-2}).$$

Proposition 3(i) is a finite sample result that holds for all values of the covariate density and the generalized propensity score, and is thus robust to weak overlap. Note that the bound on the coverage probability is sharp, in the sense that it holds with equality if the variance of the group with the smallest local sample size tends to infinity. Proposition 3(ii) shows that if Assumption 5 does not hold the new interval 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.

We remark that confidence intervals of the form of $\mathcal{I}_{S,2}$ are not new in principle, but go back to at least Banerjee (1960); see also Hayter (2014) for a more recent reference. Also

note that our confidence interval is potentially much shorter than the one resulting from the bounds in Mickey and Brown (1966), which would correspond to the case that $\rho_\alpha = 1$.

4.1.3. A Welch-Approximation Approach. The confidence interval $\mathcal{I}_{S,2}$ is generally conservative, and can potentially have coverage probability much larger than $1-\alpha$. We therefore also consider an alternative approach in which $P(T_{S,n} \leq c|M,\eta)$ is approximated by a t distribution with data-dependent degrees of freedom $\delta_* \in (\delta_{\min}, n-2J)$, which are given by

$$\begin{aligned}\delta_* &\equiv \left(\sum_{d,j} \frac{\widehat{f}(x_j)^2\widehat{\sigma}_d^2(x_j)}{N_d(x_j)}\right)^2 \Bigg/ \left(\sum_{d,j} \frac{\widehat{f}(x_j)^4\widehat{\sigma}_d^4(x_j)}{\delta_{dj}N_d(x_j)^2}\right). \\ &= \left(\sum_{d,j} \widehat{f}(x_j)\frac{\widehat{\sigma}_d^2(x_j)}{\widehat{p}_d(x_j)}\right)^2 \Bigg/ \left(\sum_{d,j} \frac{\widehat{f}(x_j)^2\widehat{\sigma}_d^4(x_j)}{\delta_{dj}\widehat{p}_d(x_j)^2}\right).\end{aligned}$$

This so-called Welch-Satterthwaite approximation, due to Welch (1938, 1947) and Satterthwaite (1946), has a long history in statistics. When applied to the standard two-sample t-statistic, it leads to Welch's two-sample t-test, which is implemented in all standard statistical software packages. This test is known to have a number of desirable properties. First, it is approximately similar[7] with only minor deviations from its nominal level when the smallest group has as few as four observations (e.g. Wang, 1971; Lee and Gurland, 1975; Best and Rayner, 1987). Second, it is asymptotically uniformly most powerful against one-sided alternatives in the class of all translation invariant tests (Pfanzagl, 1974). Third, it is robust to moderate departures from the distributional assumptions about the data (Scheffé, 1970). This all suggests that the following confidence interval for τ_S resulting from the approximation that $P(T_{S,n} \leq u|M,\eta) \approx F_t(u,\delta_*)$ should have analogously attractive properties:

$$\mathcal{I}_{S,3} = \left(\widehat{\tau} - c_\alpha(\delta_*) \times \frac{\widehat{\omega}_S}{\sqrt{n}}, \widehat{\tau} + c_\alpha(\delta_*) \times \frac{\widehat{\omega}_S}{\sqrt{n}}\right).$$

[7]The work of Linnik (1966, 1968) and Salaevskii (1963) has shown that exactly similar test for the Behrens-Fisher problem necessarily have highly undesirable properties, and thus the literature has since focused on approximate solutions.

As $\mathcal{I}_{S,2}$, the length of $\mathcal{I}_{S,3}$ does not only depend on the vector of realized local sample sizes, but also on the corresponding empirical variances. In particular, if the term $\widehat{f}(x)\widehat{\sigma}_d^2(x)/\widehat{p}_d(x)$ is very large at some point (d,x) relative its values elsewhere, then δ_* will be approximately equal to δ_{dj}, the realized local sample size at this point (minus one). In the extreme case that $\delta_* = \delta_{\min}$, the intervals can again be up to about six and a half times wider than the conventional interval $\mathcal{I}_{S,1}$ when $\alpha = .05$. If the propensity score, the covariate density and the overall sample size are such that $\delta_{\min}$ is larger than about 50 with high probability, the difference between $\mathcal{I}_{S,1}$, $\mathcal{I}_{S,2}$ and $\mathcal{I}_{S,3}$ is again not going to be of much practical relevance.

The extensive existing simulation evidence on the Welch-Satterthwaite approximation for the case of two groups suggests that under our Assumptions 1–5 one should find in finite samples that

$$P(\tau_S \in \mathcal{I}_{S,3} | \min_{d,x} N_d(x) \geq 4) \approx 1 - \alpha \tag{4.1}$$

with very high accuracy for conventional significance levels $\alpha \in (0.01, 0.1)$. This is confirmed by our own simulation experiments reported in Section 6. These simulations also show that the approximation is robust to certain reasonable departures from Assumption 5. More formally, we show that using the Welch-Satterthwaite approximation instead of a standard normal critical value leads to a higher-order correction in the asymptotic coverage error of the corresponding confidence interval if Assumption 5 holds, and does not affect the asymptotic coverage error otherwise. To simplify the exposition, we only state this result for the special case that Assumption 5 holds with $Y|D,X$ being normally distributed; but an analogous result holds with $Y|D,X$ following a scale mixture of normals.

Proposition 4. *(i) Suppose that Assumptions 1–4 hold, and that Assumption 5 holds with* $\eta_d(x) \equiv 1$. *Then we have that*

$$P(\tau_S \in \mathcal{I}_{S,3}) = 1 - \alpha + n^{-2}\phi(z_\alpha)\widetilde{q}_2(z_\alpha) + O\left(n^{-3}\right),$$

where

$$\widetilde{q}_2(t) = \frac{3t + 5t^3 + t^5}{3} \cdot \left(\frac{1}{\omega_S^3} \sum_{j,d} \frac{f(x_j)^2 \sigma_d^6(x_j)}{p_d(x_j)^5} - \frac{1}{\omega_S^4} \left(\sum_{j,d} \frac{f(x_j)\sigma_d^4(x_j)}{p_d(x_j)^3} \right)^2 \right)$$

and $\omega_S^2 = \sum_{j,d} f(x_j)\sigma_d^2(x_j)/p_d(x_j)$ *as defined above.*

(ii) Under Assumptions 1–4 and the regularity conditions of Proposition 1, we have that

$$P(\tau_S \in \mathcal{I}_{S,3}) = P(\tau_S \in \mathcal{I}_{S,1}) + O(n^{-2}).$$

Proposition 4(ii) shows that if Assumption 5 fails the new interval has again 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 4(i) implies that if Assumption 5 holds, the coverage error of $\mathcal{I}_{S,3}$ is formally of the order n^{-2}, which is better than the rate of n^{-1} we obtained for $\mathcal{I}_{S,1}$ in Proposition 1. Under limited overlap, the effective order of accuracy of $\mathcal{I}_{S,3}$ is again much smaller, but we nevertheless find a substantial improvement over $\mathcal{I}_{S,1}$. This is formally shown in the following proposition.

Proposition 5. *Recall that* $n_d(x) = nf(x)p_d(x)$*, and consider a sequence of covariate densities* $f(x)$ *and generalized propensity scores* $p_d(x)$ *such that* $n \cdot \min_{d,x} n_d(x) \to \infty$ *as* $n \to \infty$*. Then it holds that*

$$n^{-2}\phi(z_\alpha)\widetilde{q}_2(z_\alpha) = O(n_{d^*}(x^*)^{-2}),$$

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

Proposition 5 derives an approximation to the leading term $n^{-1}\phi(z_\alpha)\widetilde{q}_2(z_\alpha)$ of the Edgeworth expansion in Proposition 4 that allows for the possibility that at least some values of the generalized propensity score are close to 0. It shows that the coverage error of $\mathcal{I}_{S,3}$ is effectively similar to that of a confidence interval computed from a sample of size $n_{d^*}(x^*)^2$ instead of size n. This result should be contrasted with Proposition 2, which showed that

the coverage error of the traditional interval $\mathcal{I}_{S,1}$ effectively behaved as if a sample of size $n_{d^*}(x^*)$ was used. The Welch-Satterthwaite approximation thus improves the accuracy of the confidence interval by an order of magnitude.[8]

4.2. Robust Confidence Intervals for the PATE. In this subsection, we show how the idea behind the construction of the new confidence intervals $\mathcal{I}_{S,2}$ and $\mathcal{I}_{S,3}$ for the SATE can be extended to obtain robust confidence intervals for the PATE, which is arguably a more commonly used parameter in applications. To begin with, note that $\widehat{\tau}$ is also an appropriate estimator of τ_P, and that when viewed as such it has standard asymptotic properties under our Assumptions 1–4. In particular, we have that

$$\sqrt{n}(\widehat{\tau} - \tau_P) \xrightarrow{d} \mathcal{N}\left(0, \omega^2\right) \quad \text{with} \quad \omega^2 \equiv \mathbb{E}\left(\frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)} + (\tau(X) - \tau_P)^2\right),$$

as $n \to \infty$, and that the asymptotic variance ω^2 can be consistently estimated by

$$\widehat{\omega}^2 = \widehat{\omega}_S^2 + \widehat{\omega}_P^2, \quad \text{where} \quad \widehat{\omega}_P^2 = \sum_j \widehat{f}(x_j)(\widehat{\tau}(x_j) - \widehat{\tau})^2$$

and $\widehat{\omega}_S^2$ is as defined above. The studentized version of our estimator is again asymptotically standard normal as $n \to \infty$, that is

$$T_n \equiv \frac{\sqrt{n}(\widehat{\tau} - \tau_P)}{\widehat{\omega}} \xrightarrow{d} \mathcal{N}(0, 1),$$

which leads to the usual two-sided confidence interval for τ_P with nominal level $1-\alpha$, namely

$$\mathcal{I}_{P,1} = \left(\widehat{\tau} - z_\alpha \times \frac{\widehat{\omega}}{\sqrt{n}}, \widehat{\tau} + z_\alpha \times \frac{\widehat{\omega}}{\sqrt{n}}\right),$$

Following the argument at the end of Section 3, one can show that $\mathcal{I}_{P,1}$ has poor coverage properties for any finite sample size under limited overlap, with effective coverage error of

[8]We remark that Beran (1988) showed that in a two-sample setting with Gaussian data the higher order improvements achieved by the Welch-Satterthwaite approximation are asymptotically similar to those achieved by the parametric bootstrap.

the order $n_{d^*}(x^*)^{-1}$, where (d^*, x^*) is as defined in Proposition 2.

To motivate alternative confidence intervals similar to those we proposed for the SATE, note that the statistic T_n can be decomposed as

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

and $T_{S,n}$ is as defined above. Under our assumptions, it holds $(\widehat{\omega}_S, \widehat{\omega}_P, \widehat{\omega}) \xrightarrow{p} (\omega_S, \omega_P, \omega)$, and that $T_{S,n}$ and $T_{P,n}$ are asymptotically independent. Moreover, it is easily seen that $T_{P,n} \xrightarrow{d} \mathcal{N}(0,1)$, and given the discussion at the end of Section 3, we expect the approximation that $P(T_{P,n} \leq u) \approx \Phi(u)$ to be reasonably accurate in large samples irrespective of the values of the generalized propensity score. While it also formally holds that $T_{S,n} \xrightarrow{d} \mathcal{N}(0,1)$, we have seen in Section 3 that the approximation that $P(T_{S,n} \leq u) \approx \Phi(u)$ is not reliable under limited overlap. However, we have seen that under Assumption 5 the finite sample distribution of $T_{S,n}$ given M, η can be conservatively approximated by a "squeezed" t distribution with $\delta_{\min}$ degrees of freedom, or alternatively through the Welch approach by a t distribution with δ_* degrees of freedom with very good accuracy (at least if $N_d(x) \geq 4$). We therefore consider approximating the distribution of T_n by a (data-dependent) weighted mixture of one of these two distributions with a standard normal. Specifically, for positive constants ω_1, ω_2, δ, and ρ we define the distribution functions

$$G_C(u; \omega_1, \omega_2, \delta, \rho) \equiv P\left(\frac{\omega_1 U_C(\delta, \rho) + \omega_2 V}{(\omega_1^2 + \omega_2^2)^{1/2}} \leq u\right) \text{ and}$$
$$G_W(u; \omega_1, \omega_2, \delta,) \equiv P\left(\frac{\omega_1 U_W(\delta) + \omega_2 V}{(\omega_1^2 + \omega_2^2)^{1/2}} \leq u\right),$$

where $U_C(\delta, \rho), U_W(\delta)$ and V are independent random variables such that $P(U_C(\delta, \rho) \leq u) = F_t(u/\rho, \delta)$, $P(U_W(\delta) \leq u) = F_t(u, \delta)$, and $P(V \leq u) = \Phi(u)$. Given the number of arguments, these distribution functions are difficult to tabulate, but they can easily be

computed numerically or by simulation methods. Now let

$$g_{C,\alpha}(\delta, \rho) = G_C^{-1}(1 - \alpha/2; \widehat{\omega}_S, \widehat{\omega}_P, \delta, \rho) \text{ and}$$
$$g_{W,\alpha}(\delta) = G_W^{-1}(1 - \alpha/2; \widehat{\omega}_S, \widehat{\omega}_P, \delta)$$

be the corresponding $(1 - \alpha/2)$–quantiles for $\alpha \in (0, 0.5)$. Then an extension of our conservative confidence interval $\mathcal{I}_{S,2}$ to inference on τ_P is given by

$$\mathcal{I}_{P,2} = \left(\widehat{\tau} - g_{C,\alpha}(\delta_{\min}, \rho_\alpha) \times \frac{\widehat{\omega}}{\sqrt{n}}, \widehat{\tau} + g_{C,\alpha}(\delta_{\min}, \rho_\alpha) \times \frac{\widehat{\omega}}{\sqrt{n}} \right);$$

and an extension of our Welch-type confidence interval $\mathcal{I}_{S,3}$ to inference on τ_P is given by

$$\mathcal{I}_{P,3} = \left(\widehat{\tau} - g_{W,\alpha}(\delta_*) \times \frac{\widehat{\omega}}{\sqrt{n}}, \widehat{\tau} + g_{W,\alpha}(\delta_*) \times \frac{\widehat{\omega}}{\sqrt{n}} \right).$$

The theoretical properties of these intervals are analogous to those of $\mathcal{I}_{S,2}$ and $\mathcal{I}_{S,3}$, respectively. In particular, both can be shown to be robust to limited overlap in a similar sense. We omit a formal result in the interest of brevity.

5. Extensions to Continuously Distributed Covariates

We have introduced the assumption that the covariates X have known finite support as a modeling device that substantially simplified the theoretical arguments. We now describe a more formal way of dealing with continuously distributed covariates.

5.1. Overview and Main Ideas. If the covariates X are continuously distributed, one simple way to implement an estimator of the SATE or the PATE is discretize them and proceed as described above. That is, one could partition the support of X into J disjoint cells, recode the covariates such they take the value j if the original realization is within the jth cell, and then use the estimator we described in Section 2.3. Following Cochran (1968), such an estimation strategy is often referred to as *subclassification*. The discretization

involved in this procedure generally introduces a bias, but if the partition is not too coarse this quantity should be small. A way to further reduce the bias is to fit a more complex local model within each cell, such as a higher-order polynomial in the covariates rather than just a constant. Such an approach is often referred to as *partitioning regression*. See Györfi et al. (2002) for a textbook treatment, and Cattaneo and Farrell (2011, 2013) for some recent applications in econometrics.

Our main idea is that the techniques developed in Section 4 can be applied with very little modification to estimators of average treatment effects based on partitioning regression. We will consider an auxiliary setup that treats the local models within each cell as correctly specified linear regressions with error terms of a particular structure, and uses classical results for finite sample inference in linear regression models to construct confidence intervals for average treatment effects. We then show that these new intervals are robust to limited overlap in the sense that they have good coverage properties if the auxiliary setup is at least approximately correct, and are as good as standard approaches from a traditional large sample point of view.

5.2. Partitioning Regression. Suppose that the covariates X are continuously distributed with compact support $\mathcal{X} \subset \mathbb{R}^s$. Then a simple way to estimate the function $\mu_d(x)$ is to partition $\mathcal{X}$ into J_d disjoint cells, approximate the function by a polynomial of order K_{dj} within the jth cell, and estimate the corresponding coefficients by ordinary least squares. The partition and the order of the approximating polynomials can be different for $d \in \{0, 1\}$, and our empirical application below studies such a case.

An estimator of $\mu_d(x)$ of this form is generally referred to as a *partitioning regression estimator*, and can formally be defined as follows. For $d \in \{0, 1\}$, let $\mathcal{A}_d = \{A_{d1}, \ldots, A_{dJ_d}\}$ be a partition of $\mathcal{X}$ into J_d disjoint cells, and put $I_{dj}(x) = \mathbb{I}(x \in A_{dj})$ and $S_{d,i} = \mathbb{I}(D_i = d)$. For any $x \in \mathbb{R}^s$ and $k \in \mathbb{N}$, let $R^k(x)$ be a column vector containing all polynomials of the

form $x^u = x_1^{u_1} \cdot \ldots \cdot x_s^{u_s}$, where $u \in \mathbb{N}^s$ is such that $\sum_{t=1}^s u_t \in \{0, \ldots, k-1\}$. For example, if $s = 1$ we have that $R^k(x) = (1, x, x^2, \ldots, x^{k-1})$. With $K_d = (K_{d1}, \ldots, K_{dJ_d})$ a vector of integers, we then write $R_{dj}(x) = I_{dj}(x) R^{K_{dj}}(x)$ for the restriction of the polynomial basis $R^{K_{dj}}(x)$ to the cell A_{dj}, and define

$$\widehat{\beta}_{dj} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} S_{d,i} \left(Y_i - R_{dj}(X_i)'\beta\right)^2,$$

where the "argmin" operator 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. This definition ensures that $\widehat{\beta}_{dj}$ is well-defined even if the "local design matrix" $(S_{d1} R_{dj}(X_1), \ldots, S_{dn} R_{dj}(X_n))'$ is not of full rank. With this notation, the partitioning regression estimator of $\mu_d(x)$ is then given by

$$\widehat{\mu}_d(x) = \sum_{j=1}^{J_d} R_{dj}(x)' \widehat{\beta}_{dj}.$$

The partitioning scheme $\mathcal{A}_d$ and the degree of the polynomial approximation K_d are user-determined tuning parameters that affect the properties of $\widehat{\mu}_d(x)$. A finer partition generally decreases its bias but increases its variance; while increasing the components of K_d decreases the bias if the underlying function is sufficiently smooth, but might increase the variance because of the larger number of local parameters that need to be fitted. In view of (2.1), a natural estimator of both the PATE and the SATE is then given by

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}(X_i), \qquad \text{with} \qquad \widehat{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x).$$

This estimator is a generalization of the one we defined in Section 2.3 to a setup with continuously distributed covariates, as it would be exactly the same if the covariates X had finite support and we set $K_1 = K_0 = 1$ and $\mathcal{A}_1 = \mathcal{A}_0 = \mathcal{X}$.

5.3. Properties under "$J \to \infty$ Asymptotics". The estimator $\widehat{\tau}$ can be interpreted as a semiparametric two-step estimator that uses a particular linear sieve estimator for the nuisance function $\mu_d(x)$ in its first stage. The asymptotic properties of such estimators have been studied in Cattaneo and Farrell (2013); and Cattaneo and Farrell (2011) apply these result to a treatment effect estimator similar to ours.

Following arguments in Cattaneo and Farrell (2011, 2013), one can show that under Assumptions 1–3, equation (2.2) and certain regularity conditions on the shape of the elements of $\mathcal{A}_1$ and $\mathcal{A}_0$, and the orders K_1, K_0 of the local polynomials the estimator $\widehat{\tau}$ is $\sqrt{n}$-CAN and semiparametrically efficient for both the SATE and the PATE if $J_1, J_0 \to \infty$ as $n \to \infty$ at an appropriate rate; that is

$$\begin{aligned}
&\sqrt{n}(\widehat{\tau} - \tau_S) \xrightarrow{d} \mathcal{N}\left(0, \omega_S^2\right) \quad \text{with} \quad \omega_S^2 \equiv \mathbb{E}\left(\frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)}\right), \text{ and} \\
&\sqrt{n}(\widehat{\tau} - \tau_P) \xrightarrow{d} \mathcal{N}\left(0, \omega^2\right) \quad \text{with} \quad \omega^2 \equiv \mathbb{E}\left(\frac{\sigma_1^2(X)}{p_1(X)} + \frac{\sigma_0^2(X)}{p_0(X)} + (\tau(X) - \tau_P)^2\right).
\end{aligned}$$

The precise nature of the conditions under which these results hold are interesting and delicate (see Cattaneo and Farrell, 2011, 2013), but they are not important for our paper and thus omitted. In the following, we will simply refer to a theoretically valid argument in which the partition becomes increasingly fine when the sample size increases as "$J \to \infty$ *asymptotics*".

The asymptotic variances in the last two equations can be estimated in a variety of ways. Since a linear regression model is fitted within each cell, one way to estimate ω_S^2 is the homoskedasticity-based estimator

$$\widehat{\omega}_S^2 = \sum_{j=1}^{J} \widehat{L}_{dj}' \widehat{\Sigma}_{dj}^{-1} \widehat{L}_{dj} \widehat{\sigma}_{dj}^2,$$

where we use the notation that

$$\widehat{L}_{dj} = \frac{1}{n}\sum_{i=1}^{n} R_{dj}(X_i), \quad \widehat{\Sigma}_{dj} = \frac{1}{n}\sum_{i=1}^{n} R_{dj}(X_i)R_{dj}(X_i)', \text{ and}$$
$$\widehat{\sigma}^2_{dj} = \frac{1}{N_{dj} - K_{dj}}\sum_{i=1}^{n} I_{dj}(X_i)S_{di}(Y_i - \widehat{\mu}_d(X_i))^2,$$

with $N_{dj} = \sum_{i=1}^{n} I_{dj}(X_i)S_{d,i}$ be the number of observations with treatment status d in the jth cell of $\mathcal{A}_d$. A simple estimator of ω^2 is then given by

$$\widehat{\omega}^2 = \widehat{\omega}^2_S + \widehat{\omega}^2_P \quad \text{where} \quad \widehat{\omega}^2_P = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\tau}(X_i) - \widehat{\tau})^2.$$

Note that if the covariates X had finite support, these estimators would be numerically identical the ones defined in Sections 3–4 if we set $K_1 = K_0 \equiv 1$ and $\mathcal{A}_1 = \mathcal{A}_0 = \mathcal{X}$. Following the arguments in Cattaneo and Farrell (2011, 2013), one can show that these estimators are consistent under "$J \to \infty$ *asymptotics*" even if the data are conditionally heteroskedastic, as long as the conditional variance function is sufficiently smooth. This is because under this rather weak regularity condition the conditional variance of $Y|D, X$ can be well approximated as constant within each (small) cell of the partition.[9] These results then motivate the usual confidence intervals for the SATE and the PATE with nominal level $1 - \alpha$ are given by

$$\bar{\mathcal{I}}_{S,1} = \left(\widehat{\tau} - z_\alpha \times \frac{\widehat{\omega}_S}{\sqrt{n}}, z_\alpha \times \frac{\widehat{\omega}_S}{\sqrt{n}}\right) \quad \text{and} \quad \bar{\mathcal{I}}_{P,1} = \left(\widehat{\tau} - z_\alpha \times \frac{\widehat{\omega}_S}{\sqrt{n}}, \widehat{\tau} + z_\alpha \times \frac{\widehat{\omega}_S}{\sqrt{n}}\right),$$

respectively; and under "$J \to \infty$ *asymptotics*" the coverage probability of these intervals formally converges to $1 - \alpha$ for every fixed data generating process satisfying the necessary regularity conditions.

[9] We could also use Eicker-White-type variance estimators here as in Cattaneo and Farrell (2011, 2013), but this would complicate the formal justification of the robust confidence intervals we develop in the following subsection, as we are not aware of any exact finite sample results for studentized statistics based on such estimators. In practice, their use might still be worthwhile.

5.4. The Impact of Limited Overlap. For reasons given above, we are concerned that under limited overlap the confidence intervals $\bar{\mathcal{I}}_{S,1}$ and $\bar{\mathcal{I}}_{P,1}$ might have poor coverage properties even when the total sample size is very large. Showing this formally is difficult under "$J \to \infty$ *asymptotics*", but some insightful results are straightforward to obtain in a setting for which the number of cells is fixed as $n \to \infty$. Under such "*fixed J asymptotics*", $\widehat{\tau}$ can be though of as an estimator of biased versions of the SATE and the PATE, say B-SATE and B-PATE, which are formally defined as

$$\bar{\tau}_S = \frac{1}{n}\sum_{i=1}^{n} \bar{\tau}(X_i) \quad \text{and} \quad \bar{\tau}_P = \mathbb{E}(\bar{\tau}(X)),$$

respectively. Here we use the notation that

$$\bar{\tau}(x) \equiv \bar{\mu}_1(x) - \bar{\mu}_0(x),$$

and let $\bar{\mu}_d(x)$ be the probability limit of $\widehat{\mu}_d(x)$ as $n \to \infty$ if J_d stays fixed; that is

$$\bar{\mu}_d(x) \equiv \sum_{j=1}^{J_d} R_{dj}(x)'\beta_{dj} \quad \text{with} \quad \beta_{dj} \equiv \operatorname*{argmin}_{\beta} \mathbb{E}((Y - R_{dj}(X)'\beta)^2 | D = d, X \in A_{dj}).$$

The difference between the actual SATE τ_S and the B-SATE $\bar{\tau}_S$ can be made arbitrarily small by choosing J_1, J_0 and the components of K_1, K_0 sufficiently large; and the same applies to the difference between the PATE τ_P and the B-PATE $\bar{\tau}_P$. In particular, standard results from approximation theory suggest that if the volume of all cells in $\mathcal{A}_d$ is proportional to J_d^{-1} for $d \in \{0, 1\}$, then $\tau_j - \bar{\tau}_j = O(J_1^{-\min\{K_1\}/s} + J_0^{-\min\{K_0\}/s})$ for $j = \{S, P\}$ if $\tau(x)$ is sufficiently smooth. In practice, one might be willing to assume that the analyst is able to choose J and K such that the difference between the actual parameters and their biased version is of practically negligible magnitude for the purpose of inference in finite samples. The properties of confidence intervals for the B-SATE or the B-PATE should thus carry over if they are interpreted as confidence intervals for the SATE or the PATE instead.

For the remainder of this section, we focus on the B-SATE and the SATE as the parameter

of interest, but all results holds similarly for the B-PATE and PATE as well. We also introduce the notation that A_d^x denotes the cell of $\mathcal{A}_d$ that contains x, that is $A_d^x = A_{dj}$ if $x \in A_{dj}$, and write $\bar{f}_d(x) \equiv P(X \in A_d^x)$, $\bar{p}_d(x) \equiv P(D = d|X \in A_d^x)$, $\bar{\sigma}_d^2(x) \equiv \text{Var}(Y|D = d, X \in A_d^x)$. Now suppose for simplicity that $\sigma_d^2(x) = \bar{\sigma}_d^2(x)$ for all (d, x). In this case, it is easy to see that under "*fixed J asymptotics*" we have that

$$\bar{T}_{S,n} \equiv \frac{\sqrt{n}(\widehat{\tau} - \bar{\tau}_S)}{\widehat{\omega}_S^2} \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$, The interval $\bar{\mathcal{I}}_{S,1}$ can therefore be interpreted as a confidence interval for the B-SATE, and thus also as an approximate confidence interval for the SATE. The following proposition suggests that under limited overlap the finite sample coverage properties of this interval are generally poor.

Proposition 6. *(i) Suppose that Assumptions 1–3 hold, and that $\sigma_d^2(x) = \bar{\sigma}_d^2(x)$ for all (d, x). Under regularity conditions (Hall, 1992; Hall and Martin, 1988), it holds that*

$$P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,1}) = 1 - \alpha + n^{-1}\phi(z_\alpha)\bar{q}_2(z_\alpha) + O\left(n^{-2}\right)$$

where ϕ denotes the standard normal density function, and $\bar{q}_2(t)$ is an odd function.

(ii) Consider a sequence of covariate densities $f(x)$ and generalized propensity scores $p_d(x)$ such that $\min_{d,x} n_d(x) \to \infty$ as $n \to \infty$. Then $n^{-1}\phi(z_\alpha)\bar{q}_2(z_\alpha) = O(n_{d^*}(x^*)^{-1})$, where $(d^*, x^*) \in \{0,1\} \times \mathcal{X}$ is such that $\bar{p}_{d^*}(x^*)/\bar{f}_{d^*}(x^*) = \min_{d,x} \bar{p}_d(x)/\bar{f}_d(x)$.

This result is a minor variation of Propositions 1–2 above, showing that, just as in the case of discrete covariates, the accuracy of the confidence interval is driven by the local sample size $n_{d^*}(x^*)$ instead of the total sample size n. The coverage error of $\bar{\mathcal{I}}_{S,1}$ can thus be substantial under limited overlap.

5.5. Robust Confidence Intervals. In order to derive confidence intervals that are robust to limited overlap, we consider the following generalization of Assumption 5.

Assumption 6 (Auxiliary Model)**.** $Y(d) = \bar{\mu}_d(X) + \bar{\sigma}_d(X) \cdot \varepsilon_d(X) \cdot \eta_d(X)$, *where* $\varepsilon \equiv \{\varepsilon_d(x) : (d,x) \in \{0,1\} \times \mathcal{X}\}$ *is a collection standard normal random variables,* $\eta \equiv \{\eta_d(x) : (d,x) \in \{0,1\} \times \mathcal{X}\}$ *is a collection of positive random variables with unit variance, and the components of* ε *and* η *are all independent of the data and of each other.*

This assumption postulates that within a typical cell A_{dj} of $\mathcal{A}_d$ a polynomial regression model of order K_{dj} with homoskedastic errors following a scale mixture of normals is correctly specified. This assumption is clearly unrealistic, and we do not believe that it literally holds in our setting. If J_d and the components of K_d are sufficiently large, however, it might constitute a reasonable approximation to a large class of data generating processes. We will proceed as in Section 4 and construct confidence intervals for our parameters of interest under this auxiliary assumption. We will then argue that these new intervals are robust to limited overlap if Assumption 6 is literally correct, and should thus perform well if the assumption is at least approximately true. The new intervals are also at least not worse than ones like $\bar{\mathcal{I}}_{S,1}$, which use critical values based on "$J \to \infty$ *asymptotics*".

Since the motivation is similar to the one used in Section 4, we present our new confidence intervals in rather concise form. Recall the definition that $M = \{(X_i, D_i), i \leq n\}$, and put

$$\begin{aligned}
\bar{\delta}_{\min} &= \min_{d,j} N_{dj} - K_{dj}, \qquad \bar{\delta}_{dj} = \min_{d,j} N_{dj} - K_{dj}, \\
\bar{\delta}_* &= \left(\sum_{d,j} \widehat{\lambda}_{dj} \widehat{\sigma}^2_{dj} \right)^2 \Bigg/ \left(\sum_{d,j} \widehat{\lambda}^2_{dj} \widehat{\sigma}^4_{dj} / \bar{\delta}_{dj} \right), \text{ and} \\
\bar{\rho}_\alpha &= \left(\frac{\sum_{d,j} (c_\alpha(\bar{\delta}_{dj}) / c_\alpha(\bar{\delta}_{\min}))^2 \cdot \widehat{\lambda}_{dj} \widehat{\sigma}^2_{dj} / N_{dj}}{\sum_{d,j} \widehat{\lambda}_{dj} \widehat{\sigma}^2_{dj} / N_{dj}} \right)^{1/2},
\end{aligned}$$

where $\widehat{\lambda}_{dj} = \widehat{L}'_{dj} \widehat{\Sigma}^{-1}_{dj} \widehat{L}_{dj}$. It then follows from elementary results on linear regression with fixed regressors and homoskedastic normal errors that conditional on M and η we can write the statistic $\bar{T}_{S,n}$ as the ratio of a standard normally distributed random variable and the square root of a linear combination of $J_1 + J_0$ independent χ^2-distributed random variables

scaled by the respective degrees of freedom. Arguing as in Section 4.1, we obtain again the conservative heuristic approximation that

$$P(\bar{T}_{S,n} \leq u | M, \eta) \approx F_t(u/\bar{\rho}_{\alpha(u)}, \bar{\delta}_{\min}),$$

On the other hand, when applying the Welch–Satterthwaite approach to the distribution of $T_{S,n}|M, \eta$, we obtain the approximation that

$$P(\bar{T}_{S,n} \leq u | M, \eta) \approx F_t(u, \bar{\delta}_*),$$

which as we mentioned above is very accurate in $\min_{d,j}(N_{dj} - K_{dj}) \geq 3$. Conservative and Welch-type confidence intervals with nominal level $1 - \alpha$ for the SATE are then given by

$$\bar{\mathcal{I}}_{S,2} = \left(\widehat{\tau} - c_\alpha(\bar{\delta}_{\min})\bar{\rho}_\alpha \times \frac{\widehat{\omega}_S}{\sqrt{n}}, \widehat{\tau} + c_\alpha(\bar{\delta}_{\min})\bar{\rho}_\alpha \times \frac{\widehat{\omega}_S}{\sqrt{n}} \right) \text{ and}$$
$$\bar{\mathcal{I}}_{S,3} = \left(\widehat{\tau} - c_\alpha(\bar{\delta}_*) \times \frac{\widehat{\omega}_S}{\sqrt{n}}, \widehat{\tau} + c_\alpha(\bar{\delta}_*) \times \frac{\widehat{\omega}_S}{\sqrt{n}} \right),$$

respectively. We have the following result about their coverage properties under "*fixed J asymptotics*".

Proposition 7. *(i) Suppose that Assumptions 1–3 and 6 hold. Then*

$$P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,2}) \geq 1 - \alpha \quad \text{and} \quad P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,3}) = 1 - \alpha + n^{-2}\phi(z_\alpha)\bar{\bar{q}}_2(z_\alpha) + O\left(n^{-3}\right),$$

where ϕ denotes the standard normal density function, and $\bar{\bar{q}}_2(t)$ is an odd function.

(ii) Consider a sequence of covariate densities $f(x)$ and generalized propensity scores $p_d(x)$ such that $\min_{d,x} n_d(x) \to \infty$ as $n \to \infty$. Then $\bar{\bar{q}}_2(z_\alpha) = O(n_{d^*}(x^*)^{-2})$, where $(d^*, x^*) \in \{0,1\} \times \mathcal{X}$ is such that $\bar{p}_{d^*}(x^*)/\bar{f}_{d^*}(x^*) = \min_{d,x} \bar{p}_d(x)/\bar{f}_d(x)$.

(iii) Under Assumptions 1–3 and the regularity conditions of Proposition 6, we have that

$$P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,2}) = P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,1}) + O(n^{-2}) \quad \text{and} \quad P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,3}) = P(\bar{\tau}_S \in \bar{\mathcal{I}}_{S,1}) + O(n^{-2}).$$

The proposition shows that the robust confidence intervals $\bar{\tau}_S \in \bar{\mathcal{I}}_{S,2}$ and $\bar{\tau}_S \in \bar{\mathcal{I}}_{S,3}$ achieve improvements over the standard interval $\bar{\tau}_S \in \bar{\mathcal{I}}_{S,1}$ that are qualitatively analogous to their counterparts in a setting with discrete covariates.[10] A similar result could also be obtained for the following conservative and Welch-type confidence intervals with nominal level $1-\alpha$ for the PATE:

$$\bar{\mathcal{I}}_{P,2} = \left(\widehat{\tau} - g_{C,\alpha}(\bar{\delta}_{\min}, \bar{\rho}_\alpha)\frac{\widehat{\omega}}{\sqrt{n}}, \widehat{\tau} + g_{C,\alpha}(\bar{\delta}_{\min}, \bar{\rho}_\alpha) \times \frac{\widehat{\omega}}{\sqrt{n}}\right) \text{ and}$$
$$\bar{\mathcal{I}}_{P,3} = \left(\widehat{\tau} - g_{W,\alpha}(\bar{\delta}_{\min}) \times \frac{\widehat{\omega}}{\sqrt{n}}, \widehat{\tau} + g_{W,\alpha}(\bar{\delta}_{\min}) \times \frac{\widehat{\omega}}{\sqrt{n}}\right),$$

where the critical values $g_{C,\alpha}(\cdot)$ and $g_{W,\alpha}(\cdot)$ are exactly as defined in Section 4. We omit the details in the interest of brevity.

6. Numerical Evidence

In this section, we report the results of a small simulation study, and of the application of our data to the LaLonde (1986) data on the evaluation of a labor market program.

6.1. A Small Simulation Study. We conducted several Monte Carlo experiments to investigate the performance of our proposed robust confidence intervals in finite samples. For simplicity, here we only report results for inference on the SATE in a setting where X is binary, and thus $\mathcal{X} = \{0, 1\}$. In order to ensure that the SATE remains constant across simulation runs, we hold the data $M = \{(D_i, X_i), i \leq n\}$ on covariates and treatment indicators constant in each repetition, and only simulate new values of the outcome variables. Specifically, with a total sample size of $n = 1,000$, we construct the set M such that $N_0 = N_1 = 500$ and $N_0(0) = N_0(1) = 250$. We then vary the value of $N_1(1) = N_1 - N_1(0)$ over the set $\{250, 125, 75, 25, 15, 10, 8, 6, 4, 3, 2\}$. This is equivalent to setting $\widehat{f}(0) = \widehat{f}(1) = \widehat{p}(0) =$

[10] In view of Ibragimov and Müller (2013), we conjecture that the result concerning $\bar{\mathcal{I}}_{S,2}$ might also continue to hold if Assumption 6 is weakened to allow for within-cell heteroskedasticity, although a formal proof of this is far beyond the scope of this paper.

1/2, and letting $\widehat{p}(1)$ range over the set $\{0.5, 0.25, \ldots, 0.006, 0.004\}$. Our simulations thus include settings with good, moderate and extremely limited overlap. We conduct 100,000 replications for every value of the propensity score. We also put $\mu_d(x) \equiv 0$, $\sigma_0^2(0) = \sigma_0^2(1) = \sigma_1^2(0) = 1$, and consider the cases $\sigma_1^2(1) = 4$ and $\sigma_1^2(1) = .25$ for our study. We generate outcomes as $Y_i = \mu_{D_i}(X_i) + \sigma_{D_i}(X_i) \cdot \varepsilon_{D_i}(X_i)$, where the distribution of the error term is a mixture of a standard normal distribution and a standard exponential distribution centered at zero. That is, we have $\varepsilon_d(x) \sim \lambda \cdot \mathcal{N}(0,1) + (1-\lambda) \cdot (\text{Exp}(1) - 1)$, where $\lambda \in [0,1]$ is the mixture weight. For our simulations, we consider the cases $\lambda = 1$ and $\lambda = .5$. The first type of error distribution satisfies out Assumption 5, whereas the second one does not and is included to check the robustness of our methods against deviations from the auxiliary distributional assumption.

The top panels of Figure 1 shows the simulated finite sample coverage probabilities of the three confidence intervals $\mathcal{I}_{S,1}$ (standard; black line), $\mathcal{I}_{S,2}$ (conservative; blue line) and $\mathcal{I}_{S,3}$ (Welch; red line) for the different values of the empirical propensity score $\widehat{p}(1)$ and $\lambda = 1$. The top left panel reports results for $\sigma_1^2(1) = 4$, where as the top right panel reports results for $\sigma_1^2(1) = .25$ In both cases, the standard interval's coverage rate is close to the nominal level for $\widehat{p}(1) \geq 0.05$, which corresponds to realized local sample sizes such that $N_1(1) \geq 25$. For smaller values of the propensity score, $\mathcal{I}_{S,1}$ becomes more and more distorted, eventually deviating from the nominal level by almost 25 percentage points. As suggested by its construction, the coverage probability of our conservative interval $\mathcal{I}_{S,2}$ exceeds its nominal level for all values of the propensity score. However, the deviations are surprisingly minor, becoming noticeable only for $\widehat{p}(1) \leq 0.04$ and $\sigma_1^2(1) = .25$, and even then never exceed one percentage point. Our Welch-type interval $\mathcal{I}_{S,3}$ also has correct coverage probability for most values of the propensity score. However, it shows some distortions for $\widehat{p}(1) \leq 0.02$, which corresponds to settings with realized local sample sizes such that $N_1(1) \leq 4$.

In the bottom panel of Figure 1, we report the results of our simulation experiments in

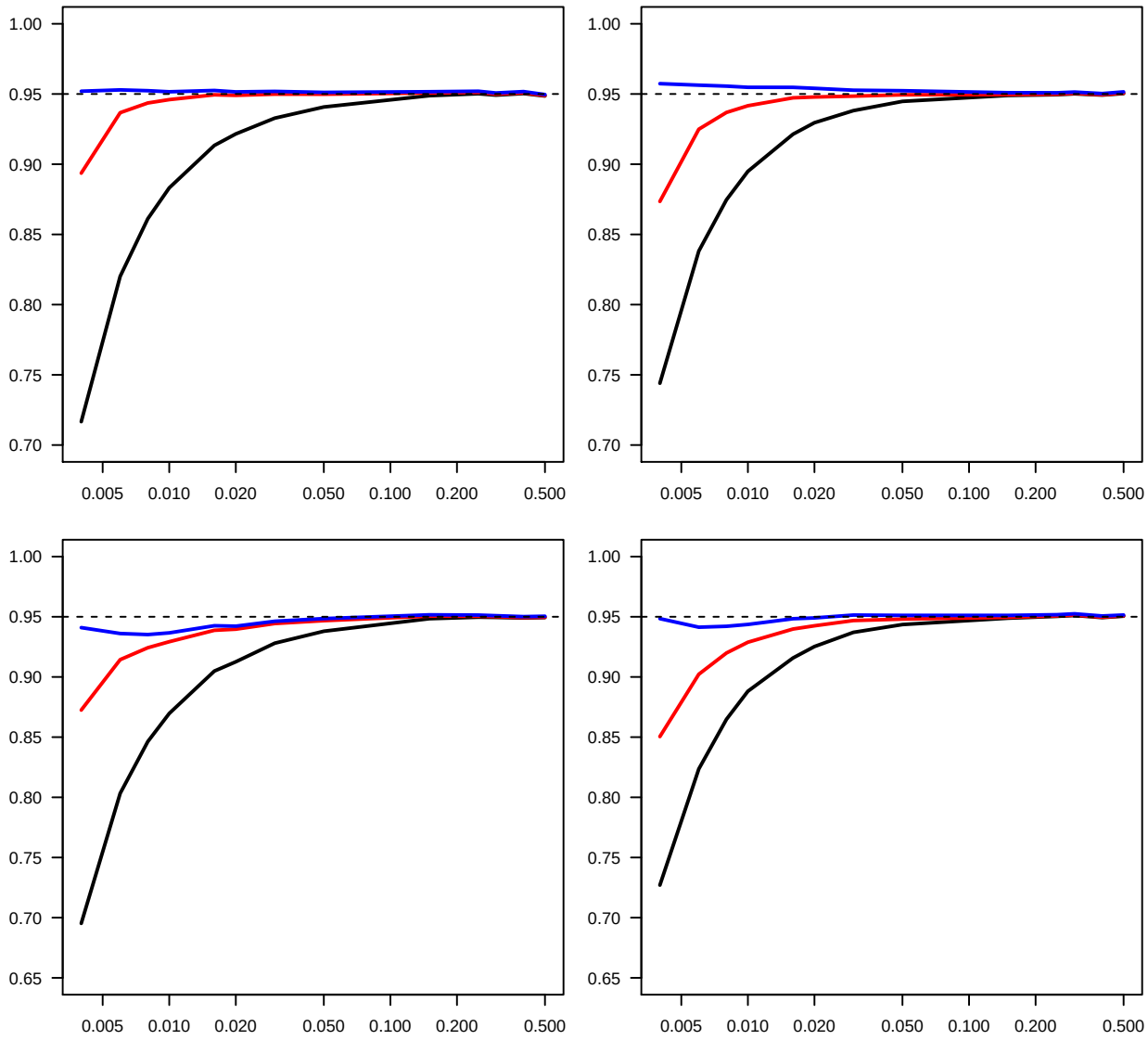

Figure 1: Empirical coverage probabilities of $\mathcal{I}_{S,1}$ (standard; black line), $\mathcal{I}_{S,2}$ (conservative; blue line) and $\mathcal{I}_{S,3}$ (Welch; red line) for the different values of the empirical propensity score $\widehat{p}(1)$ between 0.004 and 0.5 (or, equivalently, values of realized local sample size $N_1(1)$ between 2 and 250). The parameters being used are $\lambda = 1, \sigma_1^2(1) = 4$ (top left panel), $\lambda = 1, \sigma_1^2(1) = .25$ (top right panel), $\lambda = .5, \sigma_1^2(1) = 4$ (bottom left panel), and $\lambda = .5, \sigma_1^2(1) = .25$ (bottom right panel).

which $\lambda = .5$. The bottom left panel reports results for $\sigma_1^2(1) = 4$, whereas the bottom right panel reports results for $\sigma_1^2(1) = .25$. These are both settings in which our Assumption 5 does not hold. Following Propositions 3 and 4, our robust confidence intervals formally only

have the same asymptotic coverage error as the standard interval in this case. However, since the distribution of the errors is not "too different" from Gaussian in this experiment, one would hope that some of the robustness properties are preserved. Our simulations show that this is indeed the case. The results for all three confidence intervals are qualitatively very similar to the case where $\lambda = 1$. Our robust confidence intervals suffer from a slight additional distortion for low values of the propensity score, but those are very mild relative to those of the standard intervals. This suggests our constructions remains beneficial even if our stringent distributional assumptions are substantially violated.

6.2. An Empirical Illustration. In this subsection, we apply the methods proposed in this paper to data from the National Supported Work (NSW) demonstration, an evaluation of an active labor market program first analyzed by LaLonde (1986), and then subsequently by Dehejia and Wahba (1999) and many others. The NSW demonstration was a federally and privately funded program implented in the mid-1970's in which hard-to-employ people were given work experience for 12 to 18 months in a supportive but performance-oriented environment. The data set that we use here (taken from Dehejia and Wahba, 1999) is a combination of a sample of 185 participants from a randomized evaluation of the NSW program, and a sample of 2490 non-participants taken from the Panel Study of Income Dynamics (PSID). For the purpose of illustration, we ignore the fact that the data combine two populations, and treat them as being a single sample from "pseudo-population" for which we wish to determine the average treatment effect of the NSW program.

The left panel of Table 1 presents some summary statistics for the data used in our analysis. Note that there are major differences in pre-treatment characteristics between individuals that participate in the program and those who do not. A practical concern is thus that there might be *no* overlap in larger parts of the covariate space. We therefore first estimated the propensity score using a partitioning approach to investigate this issue. Since

Table 1: Descriptive Statistics

	Original Data				Trimmed Data			
	Treated (185)		Control (2490)		Treated (178)		Control (475)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Covariates								
Age	25.81	7.15	34.85	10.44	25.66	7.18	35.84	11.46
Education	10.34	2.01	12.12	3.08	10.33	2.01	11.44	3.34
Black	0.84	0.36	0.25	0.43	0.84	0.37	0.30	0.46
Hispanic	0.06	0.24	0.03	0.17	0.06	0.24	0.04	0.21
Married	0.19	0.39	0.87	0.34	0.17	0.38	0.78	0.42
Earnings '74	2.10	4.89	19.43	13.41	1.45	3.27	5.33	8.23
Earnings '75	1.53	3.22	19.06	13.59	1.06	1.88	3.41	7.29
Outcome								
Earnings '78	6.34	7.86	21.55	15.55	6.11	7.61	8.81	14.50

Note: Earnings data are in thousands of 1978 dollars.

the covariates we consider here include both binary and continuously distributed variables, we partition their support using the Classification and Regression Trees (CART) algorithm (Breiman, Friedman, Stone, and Olshen, 1984) as implemented in the library `rpart` of the statistical software package `R` (Ihaka and Gentleman, 1996). CART can be thought of as a local constant partitioning regression estimator with a data-dependent partition of the covariate space. Specifically, CART creates a partition through a series of binary splits, chosen such that at each step the greatest possible reduction in the total within-cell sum of squares is achieved.

Figure 2 shows the CART estimate of the propensity score; that is the structure of the partition, the resulting cell sizes, and the local estimates $\widehat{p}(x)$. The graph indicates that there is indeed a large group of 2022 units, defined as those with earnings in 1974 greater than 2.3 and earnings in 1975 greater than 6.1, in which the share of treated individuals is extremely low at 0.0035. This indicates that estimating the function $\mu_1(x)$ would require a rather coarse partition of the covariate space in this region, which in turn would likely lead to a substantial partitioning bias. We therefore remove these observations from the sample, and consider estimating the ATE for the remaining 178 treated and 475 control units. Summary

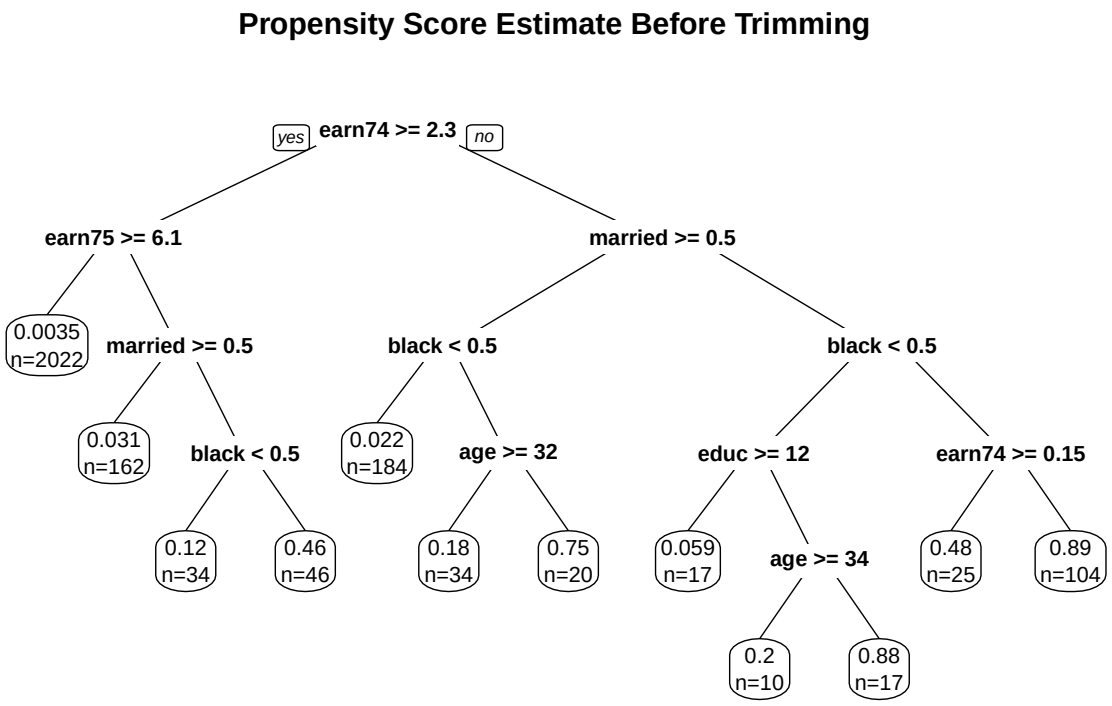

Figure 2: Estimated propensity score in the full sample. Tree represents the partition of the covariate space as determined by the CART algorithm for treated and untreated individuals. Numbers in boxes denote the respective local estimate of the propensity score $p(x)$, and the corresponding realized local sample size $N(x)$.

statistics for this trimmed sample are given in the right panel of Table 1.

Note that while the use of CART is somewhat unusual in treatment effects applications, their structure makes regression trees particularly suitable for identifying regions on the covariate space with no or limited overlap. If the propensity score had been estimated by, say, a Logit or Probit model, it would have been much harder to characterize the regions of the covariate space in which treated observations are only sparsely located.

In the next step, we then estimated the function $\mu_d(x)$ on the trimmed data, using again the CART algorithm to determine the partition of the covariate space. This was

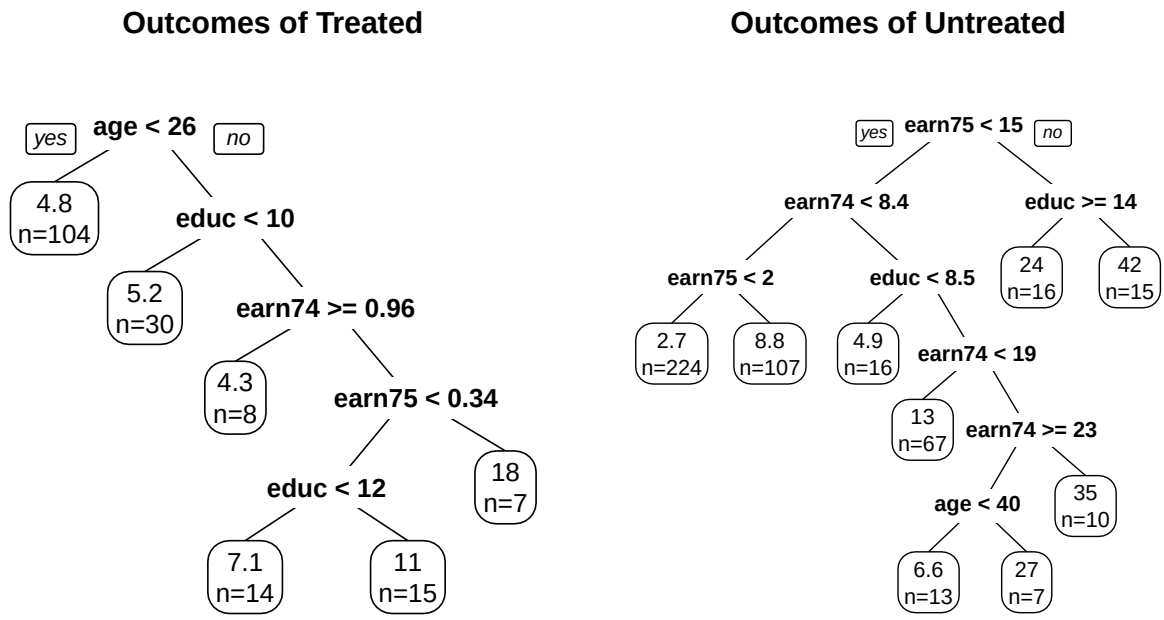

Figure 3: Partition of the covariate space as determined by the CART algorithm for treated and untreated individuals. Numbers in boxes denote the respective local estimate of the function $\mu_d(x)$, and the corresponding realized local sample size $N_d(x)$.

done separately for treated and untreated units.[11] Figure 3 shows the findings, namely the structure of the partition, the cell sizes $N_d(x)$, and the estimates of $\mu_d(x)$ obtained by fitting a constant to the data on earnings in 1978 within each cell (in multiples of \$1,000, and rounded to the nearest decimal). The graph is to be read as follows: CART created a cell containing all 104 treated units with less than 26 years of age (with average 1978 earnings of 4.8); another cell containing all 30 treated units that are 26 or older and have less than 10 years of education (with average 1978 earnings of 5.2); and so forth. In total, we partition the treatment and non-treatment groups into 6 and 9 cells, respectively. Cell sizes show a substantial amount of heterogeneity, ranging from 7 to 224, and there are a number of

[11]There is no need for the partition to be constant across treatment status, in the same way that there is no need to use the same smoothing parameter for the treated and untreated samples if any other type of nonparametric estimator was being used to estimate $\mu_d(x)$.

Table 2: Effect of NSW program on Earnings '78

	SATE Inference	PATE Inference
Estimation results		
Point Estimate	-0.74	-0.74
Standard Error	0.96	1.03
Critical value (nominal 95% level)		
Standard	1.96	1.96
Welch	2.01	2.01
Conservative	2.25	2.20
Two-sided confidence interval (nominal 95% level)		
Standard	[-2.62, 1.25]	[-2.75, -1.27]
Welch	[-2.67, 1.20]	[-2.80, -1.32]
Conservative	[-2.91, 1.44]	[-3.00, -1.52]

Note: The outcome is earning in 1978 in thousands of 1978 dollars.

small cells with less than 20 observations. This suggests that limited overlap might be a concern for inference.[12] Note that the covariates used by the CART algorithm to define the partition differ between the two treatment groups, and some covariates are not used at all. The non-inclusion of a covariate means that a split based on its realizations does not lead to sufficiently large improvement in fit according to CART's default stopping criteria.

In Table 2, we report the final results of applying our methods to the data.[13] We consider both inference on the SATE and the PATE. The point estimate of both parameters is -0.74, which is larger than the unadjusted difference in outcomes of treated and untreated individuals of -2.70 we obtain from the right panel of Table 1. Our results show that in order to conduct overlap-robust inference on the SATE using the Welch correction, one should use a critical value of $c_\alpha(\bar{\delta}_*) = 2.01$ (45.8 degrees of freedom) for $\alpha = 0.05$ in this case, which translates into a confidence interval that is only 3% longer than the standard one using the critical value 1.96. Using our conservative approach leads to a critical value of $c_\alpha(\bar{\delta}_{\min})\bar{\rho}_\alpha = 2.25$,

[12]Some additional calculations show that the cell with the lowest ratio of the estimated generalized propensity score and the estimated covariate density, i.e. the sample analogue of the point (d^*, x^*) we defined above, is the one containing treated units with `age` ≥ 26, `educ` ≥ 10 and `earn74` ≥ 0.96, which contains 8 observations.

[13]Note that our results formally do not cover the case of a data-driven partition, but we ignore this additional source of variation for simplicity here.

which gives a confidence interval that is about 15% wider. Given our simulation results, this would be our preferred confidence interval. Inference results for the PATE are qualitatively similar, with robust confidence intervals being a little less wide relative to the standard one.

7. Conclusions

Limited overlap creates a number of challenges for empirical studies that wish to quantify the average effect of a treatment under the assumption of unconfounded assignment. In addition to point estimates being rather imprecise, an important practical problem is that standard methods for inference can break down. For example, commonly used confidence intervals of the form "point estimate$\pm$1.96$\times$standard error" can have coverage probability substantially below their nominal level of 95% even in very large, but finite, samples. This paper has provided some insights for why this phenomenon occurs, and proposed new robust confidence intervals that have good theoretical and practical properties in many empirically relevant settings.

A. Proofs

A.1. Proof of Propositions 1 and 2. Proposition 1 can be shown by 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 M by $H_n(\cdot|M)$, it follows from their reasoning that under the conditions of the proposition $H_n(\cdot|M)$ satisfies the following Edgeworth expansion:

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

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 confidence interval $\mathcal{I}_{S,n}$ given M is given by

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

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

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

The result of Proposition 1 then follows from the fact that $\mathbb{E}(\widehat{q}_2(z_\alpha)) = q_2(z_\alpha) + O(n^{-1})$, the relationship that $P(\tau_S \in \mathcal{I}_{S,n}) = \mathbb{E}(P(\tau_S \in \mathcal{I}_{S,n}|M))$, and dominated convergence. Proposition 2 follows from some simple algebra.

A.2. Proof of Proposition 3. To show part (i) we first prove the following auxiliary result, which is similar to a finding of Hayter (2014).

Lemma 1. *Let X be be normally distributed with mean zero and unit variance, 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 $u > 0$ it*

holds that

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

for all $\gamma \in \Gamma$.

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 $P(V_{e_k} \leq u) \geq \max_{k=1,\ldots,K} F_t(u/a_k^{1/2}, s_k)$. This completes our 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 (M, η), by putting (with a slight abuse of notation) that

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

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

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

and by noting that since the inequality holds conditional on (M, η) 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 Propositions 4 and 5. Proposition 4(i) is follows from a classical result in Welch (1947) and arguments similar to those used to prove Proposition 1. Proposition 4(ii) follows from the fact that $c_\alpha(\delta) = z_\alpha + O(\delta^{-1})$, which implies that $c_\alpha(\delta_*) = z_\alpha + O(n^{-1})$. Finally, Proposition 5 follows from some simple algebra in the same way as Proposition 2.

A.4. Proof of Propositions 6 and 7. The results follow from minor modifications of the arguments used to prove Propositions 1–5 and standard results for concerning the finite sample properties of least squares estimators in correctly specified linear regression models with homoskedastic Gaussian errors. Details are omitted.

References

Abadie, A. and G. W. Imbens (2006): "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74, 235–267.

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.

Beran, R. (1988): "Prepivoting test statistics: a bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, 687–697.

Best, D. and J. Rayner (1987): "Welch's approximate solution for the Behrens–Fisher problem," *Technometrics*, 29, 205–210.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984): *Classification and regression trees*, CRC press.

Cattaneo, M. D. and M. H. Farrell (2011): "Efficient estimation of the dose–response function under ignorability using subclassification on the covariates," *Advances in Econometrics*, 27, 93–127.

——— (2013): "Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators," *Journal of Econometrics*, 174, 127–143.

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

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics*, 36, 808–843.

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

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.

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

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

Györfi, L., A. Krzyzak, M. Kohler, and H. Walk (2002): *A distribution-free theory of nonparametric regression*, Springer Verlag.

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.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an econometric evaluation estimator," *Review of Economic Studies*, 65, 261–294.

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

Ibragimov, R. and U. K. Müller (2013): "Inference with Few Heterogenous Clusters," *Working Paper*.

Ihaka, R. and R. Gentleman (1996): "R: a language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

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

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

Imbens, G. W. and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Khan, S. and D. Nekipelov (2013): "On Uniform Inference in Nonlinear Models with Endogeneity," *Working Paper*.

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

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

Lee, A. F. and J. Gurland (1975): "Size and power of tests for equality of means of two normal populations with unequal variances," *Journal of the American Statistical Association*, 70, 933–941.

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*, 639–642.

Pfanzagl, J. (1974): "On the Behrens-Fisher problem," *Biometrika*, 61, 39–47.

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

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

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

Satterthwaite, F. (1946): "An approximate distribution of estimates of variance components," *Biometrics Bulletin*, 2, 110–114.

Scheffé, H. (1970): "Practical solutions of the Behrens-Fisher problem," *Journal of the American Statistical Association*, 1501–1508.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Wang, Y. Y. (1971): "Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem," *Journal of the American Statistical Association*, 66, 605–608.

Welch, B. (1938): "The significance of the difference between two means when the population variances are unequal," *Biometrika*, 29, 350–362.

——— (1947): "The generalization of Student's' problem when several different population variances are involved," *Biometrika*, 28–35.

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