Valid Inference for Treatment Effect Parameters under Irregular Identification and Many Extreme Propensity Scores

Phillip Heiler\textsuperscript{a},\textsuperscript{*}, Ekaterina Kazak\textsuperscript{b}

\textsuperscript{a}Aarhus University, Department of Economics and Business Economics, CREATES, TrygFonden’s Centre for Child Research, Fuglesangs Allé 4, 8210 Aarhus V, Denmark.
\textsuperscript{b}University of Manchester, Department of Economics, Arthur Lewis Building, Oxford Road, Manchester M13 9PL, United Kingdom.

Abstract
This paper provides a framework for conducting valid inference for causal parameters without imposing strong variance or support restrictions on the propensity score. In particular, it covers the case of irregularly identified treatment effect parameters. We provide limit theorems for inverse probability weighting and doubly robust estimation of causal or counterfactual parameters that do not rely on trimming approaches. By construction the limiting distributions of these estimators belong to the alpha-stable class which implies that standard inference methods such as the nonparametric bootstrap are inconsistent. We propose an adaptive version of the $m$-out-of-$n$ bootstrap that is robust to all types of identification and a bootstrap aggregation method for the optimal $m$ choice. Monte Carlo simulations suggest that the modified resampling method compares favorably to conventional methods in finite samples. The method is applied to a re-analysis of the causal impact of right heart catheterization on survival rates.

Keywords: Inverse probability weighting, Irregular identification, Propensity score, Stable distribution, Treatment effect

\textit{JEL:} C12, C21

1. Introduction

Statistical models for causal parameters such as the average treatment effect or other counterfactual parameters often invoke two major assumptions: First, the potential outcomes have to be independent of the treatment after conditioning on an appropriate set of confounding variables. Second, the conditional probability of being selected into treatment, the propensity score, cannot be equal to zero or one. While these assumptions are sufficient for point identification (Rosenbaum and Rubin, 1983), they do not guarantee standard estimators to be root-n consistent because they allow for infinite information bounds and variances

\textsuperscript{*}Corresponding author: 
Email address: pheiler@econ.au.dk (Phillip Heiler)
even when propensity scores are known. Invoking a more restrictive support condition sometimes referred to as strong overlap, i.e. assuming that the propensity scores are supported on \((\delta, 1 - \delta)\) for some \(\delta > 0\), or directly imposing variance restrictions is a common practice in the literature. It assures that estimators converge to a normal distribution at the usual root-n rate which consequently can be used for asymptotically valid inference. However, these are both very restrictive assumptions that rule out a variety of empirically relevant data generating processes.

In this paper we show how to conduct valid inference for causal parameters without imposing variance or support restrictions on the propensity score, i.e. when treatment effect parameters are irregularly identified. We provide several limit theorems for typical estimators of causal or counterfactual parameters such as inverse probability weighting and doubly robust estimation. In particular, we cover the most challenging case of many extreme propensity scores. We show that through the inverse probabilities the limiting distributions of the estimators belong to the alpha-stable class by construction. This class allows for heavy-tailed, skewed limiting distributions but also contains the normal distribution as a special case. The result implies that common approaches for inference such as analytical confidence intervals based on asymptotic normality or resampling methods such as the nonparametric bootstrap can yield incorrect p-values and test decisions. This paper explicitly addresses the link between identification conditions and distributional behavior in a general setup, i.e. without changing the sample and target parameter of interest via trimming and/or weighting (Hirano et al., 2003, Crump et al., 2009, Li et al., 2018) or explicit distributional assumptions (Khan and Tamer, 2010). Our limit theorems allow for the application of modified bootstrap methods that are robust with respect to the different limiting distributions. Furthermore, we provide evidence that changes in the limiting distribution can be caused by misspecification of the propensity score model. In particular, it is possible that misspecification does not affect the consistency of the estimator for the treatment effect but causes standard inferential methods based on asymptotic normality or the nonparametric bootstrap to fail. This is particularly relevant for doubly robust estimation that guarantees consistency under partial misspecification of either propensity score or potential outcome models. We also propose a simple model selection criterion for decreasing the likelihood of misspecification induced irregularities in applied work.

The main assumption behind our limit theorems restricts the occurrence of extreme inverse probability weighted potential outcomes or conditional mean errors. For relevant special cases, it reduces to regular variation (Feller, 1968, Bingham et al., 1989) of the density of the propensity score at zero or one. We show that the rate of convergence of the treatment effect estimator is directly linked to the behavior of the propensity score density at these boundaries, i.e. the faster its divergence, the slower the rate of convergence. Our results also raise questions regarding the validity of mean squared error comparisons common in the literature as estimates of sample variances do not converge to finite numbers if treatment parameters are irregularly identified.

We propose to use a modified version of the \(m\)-out-of-\(n\) bootstrap by Bickel et al. (1997) that replaces the
denominator of the test statistic by a normalization that is adaptive to the degree of identification. For the optimal $m$-choice we introduce bootstrap aggregation (bagging) version of the method by Bickel and Sakov (2008) which substantially improves coverage rates in finite samples. We also provide clear guidance on how to choose the involved hyperparameters for the $m$-choice in empirical applications. The proposed resampling method is not only consistent for the stable limiting case but also yields asymptotically correct coverage rates under regular identification and thus the researcher can apply it without knowledge of the regularity of his or her identification problem. Under regular identification, it is superior to fixed $m$-rules as it can select $m = n$ in large samples which helps to achieve the same second-order properties as pivotal bootstrap methods. Our Monte Carlo simulations suggest that for doubly robust estimation, the adaptive $m$-out-of-$n$ bootstrap yields sizes close to the nominal level in finite samples while conventional methods or trimming can produce rejection rates up to 100% under the null hypothesis. For inverse probability weighting all methods tend to over-reject in finite samples, however the proposed modified bootstrap is closest to the nominal rejection rates for all sample sizes.

The method is applied to a re-evaluation of the causal impact of right heart catheterization (RHC) on 30-day survival rates of initial care patients, see Murphy and Cluff (1990), Connors et al. (1996), and Hirano and Imbens (2001). Typically, these studies report a significant negative effect of RHC on survival rates. However, the empirical distributions of the propensity scores indicate a potential violation of strong overlap which advocates for a robust inferential method. Our robust resampling technique suggests that the significant effects reported in the literature are mostly robust with respect to a more relaxed overlap assumption. However point estimates obtained by inverse probability weighting are no longer significant at the 1%-level with a p-value of 0.015.

The literature has been concerned with the presence of extreme propensity scores for many years. Papers concerned with the statistical properties of different estimation approaches such Robins and Rotnitzky (1995), Wooldridge (2002, 2007), Hirano et al. (2003), Belloni et al. (2014), Farrell (2015), Chernozhukov et al. (2017), or Rothe and Firpo (2018) either impose the support restriction or directly assume finiteness of the asymptotic variance. Khan and Tamer (2010) show that without knowledge of the propensity score, the semiparametric efficiency bound is infinite if the propensity score is supported on $(0, 1)$ which implies non-regular convergence of the inverse probability weighting estimator.

A standard solution to assure a finite variance is to trim or to reweigh the sample based on the propensity scores. Heckman et al. (1998) remove observations with a propensity score density below a threshold. Dehejia and Wahba (1999) trim observations outside the range of propensity score values in the complementary treatment group. Imbens (2004) and Huber et al. (2013) suggest to drop observations that have a relative impact above a certain threshold on the final estimate. Crump et al. (2009) derive an optimal trimming rule for the average treatment effect that minimizes the critical term in the semiparametric efficiency bound by restricting the sample and show that a good rule of thumb for trimming is the interval $[0.1, 0.9]$ if propensity
scores follow a beta distribution. Alternative approaches that do not trim but reweigh the sample are usually based on the propensity score as well, see e.g. Fröhlich (2004), Pohlmeier et al. (2016) and Li et al. (2018). There is also a variety of studies that investigate the effects of extreme propensity scores from a small sample perspective using (empirical) Monte Carlo simulations (Busso et al., 2014, Lechner and Strittmatter, 2017). They show that trimming and reweighing can often be beneficial in terms of point estimation risk and that the relative performance of the different strategies depends heavily on the degree of overlap assumed. However, all these studies only look at (root) mean squared error differences and do not consider the impact of extreme propensity scores on distributional properties and statistical inference. Moreover, while trimming or reweighing approaches can help to improve point estimation risk in finite samples for certain designs, they either vanish asymptotically or affect the distribution of the treatment effect estimator, i.e. instead of estimating a parameter for the whole population only the parameter for a reweighted population in the sense of Hirano et al. (2003) is considered and/or the effects are only valid within a subpopulation. This is particularly questionable if there is strong heterogeneity in the conditional treatment effects at the boundary, i.e. for observations that have propensity scores close to zero or one. Neglecting the heterogeneity can have a large distorting impact on both point estimates and inference. The method proposed in this paper is also complementary to trimming as it can be used to evaluate the robustness of the statistical inference with regards to disregarding observations.

Rothe (2017) argues that for a given sample there is no conceptual difference between an arbitrarily small $\delta$ or simply a very small $\delta$ (“limited overlap”). He shows that under limited overlap standard coverage rates are driven by small local sample sizes and proposes a finite-sample correction following the literature on the Behrens-Fisher problem and two-sample testing that relies on distributional assumptions. The local sample size argument, however, cannot directly be translated to the case of irregular identification as the asymptotic behavior is then driven by the thin region arbitrarily close to the boundary of the support.\textsuperscript{1} Moreover, for the repeated sequence from experiment/data-generation to inference, varying support conditions will eventually lead to different unconditional distributions and test decisions. Going beyond limited overlap, our approach differs from Rothe (2017) as it provides an accurate characterization of the large sample behavior of the treatment effect estimators under irregular identification that can be used for asymptotically valid inference without restrictive distributional assumptions. In independent work Ma and Wang (2019) also consider heavy-tail robust inference for the standard inverse probability weighting estimator based on subsampling. They focus on trimming and do not consider optimal subsampling choices or misspecified propensity score models.

The paper is organized as follows: Section 2 motivates and outlines the main theoretical results using

\textsuperscript{1}In principle, the expansions in Rothe (2017) could be used to model a case for which $\delta$ goes to zero as the sample size increases. This would require a parameterization of local-to-zero sample sizes and should affect the limiting behavior of the estimator in a non-trivial fashion. This is beyond the scope of this paper and an interesting topic for further research.
a stylized example of an inverse probability weighting estimator under simplified assumptions. Section 3 introduces and discusses the main assumptions and provides the large sample theory for inverse probability weighting and doubly robust estimation. It also contains some examples including an irregular limiting domain caused by misspecification of the propensity score model and the model selection criterion. Section 4 provides the basic theory for inference and outlines the inferential procedure. Section 5 contains a Monte Carlo study that compares the performance of the robust to the standard methods of constructing confidence intervals. Section 6 contains the empirical application. Section 7 concludes.

Notation. Unless stated otherwise, all statements about random variables are with probability one. \( l(\cdot) \) denotes (different) slowly varying functions at infinity, i.e. \( \lim_{t \to \infty} \frac{l(tz)}{l(t)} = 1 \) for all \( z > 0 \). \( f(z) \sim g(z) \) means that \( \lim_{t \to \infty} \frac{f(t)}{g(t)} = 1 \). \( f \) and \( g \) denote probability density and cumulative distribution function of a random variable \( X \). We say that a random variable \( X \) belongs to the domain of attraction of a (class of) distribution(s) \( F \) if there exist normalizing sequences \( c_n, b_n \) such that \( c_n \sum_{i=1}^n X_i - nb_n \overset{d}{\to} F \) with \( F \in \mathcal{F} \). \( V \) is a placeholder for finite variance terms while \( V[\cdot] \) denotes the variance of the term in the brackets. \( \lambda_1(\cdot) \) denotes the maximum eigenvalue of a matrix.

2. Motivating Example

Assume we observe independent identically distributed data \((Y_i, D_i, X_i')\) for \( i = 1, \ldots, n \) where \( Y_i \) is a real-valued outcome variable of interest, \( D_i \) a treatment indicator and \( X_i \) a \( k \)-dimensional vector of real-valued confounding variables. Let \( Y_i(d), d \in \{0, 1\} \) denote the potential outcomes after receiving treatment status \( d \) for unit \( i \) and let \( p(X_i) = P(D_i = 1|X_i) \) denote the propensity score. For the remainder of this paper assume that unconfoundedness and overlap hold:

Unconfoundedness.

\[ Y_i(1), Y_i(0) \perp D_i|X_i. \]

Overlap.

\[ 0 < p(X_i) < 1 \quad \text{with probability one.} \]

The fundamental problem of causal inference (Holland, 1986) yields the observational rule \( Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \). Say, we would like the estimate the potential outcome mean \( E[Y_i(1)] = \theta_0 \). A typical inverse probability weighting estimator \( \hat{\theta}_n \) for estimating \( \theta_0 \) is then given by

\[ \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{p(X_i)}, \quad (2.1) \]

see e.g. Hirano et al. (2003). For simplicity, assume that propensity scores \( p(X_i) \) are observed and that the conditional second moment of the potential outcome is bounded, i.e. \( C_1 < E[Y_i(1)^2|X_i] < C_2 < \infty \) almost
everywhere.\footnote{Neither boundedness nor the specific estimator nor knowledge of the propensity scores are required for the core results of the paper. It serves a merely illustrative purpose at this stage. For a generalization consider Section 3.} Under unconfoundedness and overlap one can show that $\hat{\theta}_n \overset{D}{\to} \theta_0$. For constructing confidence bands for $\theta_0$ we are typically using a stabilizing transformation, i.e. we investigate the limiting behavior of

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{D_i Y_i}{p(X_i)} - E[Y_i(1)] \right). \quad (2.2)$$

By invoking a standard central limit theorem, one can show that the stabilizing transformation converges in distribution to a normal random variable if the variance of the components of the sum is finite, i.e. if

$$V\left[ \frac{D_i Y_i}{p(X_i)} \right] = E\left[ \frac{E[Y_i(1)^2|X_i]}{p(X_i)} \right] - E[Y_i(1)]^2 < \infty. \quad (2.3)$$

We are concerned about finiteness of the first component on the right-hand side as it contains components divided by propensity scores which can be arbitrarily close to zero. In the literature, (2.3) is often invoked directly or the identification condition overlap is strengthened to strong overlap:

\textit{Strong Overlap.}

$$\delta < p(X_i) < 1 - \delta \quad \text{with probability one for some } \delta > 0. \quad (2.4)$$


$$C_1 E\left[ \frac{1}{p(X_i)} \right] < E\left[ \frac{E[Y_i(1)^2|X_i]}{p(X_i)} \right] < C_2 E\left[ \frac{1}{p(X_i)} \right]. \quad (2.5)$$

Thus, for a finite asymptotic variance of (2.2) finiteness of $E[1/p(X_i)]$ is necessary and sufficient. Strong overlap is enough to bound (2.3) from above since

$$E\left[ \frac{1}{p(X_i)} \right] < \frac{1}{\delta} < \infty. \quad (2.6)$$

Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N\left( 0, V\left[ \frac{D_i Y_i}{p(X_i)} \right] \right), \quad (2.7)$$

which allows for construction of asymptotically valid confidence intervals and thus valid statistical inference. However imposing (2.3) or (2.4) is not harmless since it severely restricts the class of distributions for the
propensity score. For example, consider a simple threshold crossing model for selection into treatment

\[ D_i = \mathbb{1}(X_i - u_i > 0), \]  

which implies that \( p(x) = F_u(x) \). Hence if \( X_i \) and \( u_i \) are continuously distributed on the real line then for any \( \delta > 0 \) there exists a value \( x \) in the support of \( X_i \) such that \( p(x) < \delta \) violating strong overlap. Then \( E[1/p(X_i)] \) and hence the variance will only be finite if the density for \( X_i \) vanishes fast enough for these extreme propensity scores, a point also made by Khan and Tamer (2010). \( X_i \) and \( u_i \) just having equally tailed distributions is sufficient for an infinite variance which rules out e.g. a probit model with a normally distributed regressor \( X_i \) that enters (2.8) with a unit coefficient or larger. Another example where the strong overlap assumption is violated are propensity scores that have tails like a beta distribution, i.e. if \( p \sim B(a, b) \) then

\[
E\left[\frac{1}{p(X_i)}\right] = \int_0^1 \frac{p^{a-2}(1-p)^{b-1}}{B(a,b)} dp,
\]

where \( B(a,b) \) is the beta function. Thus the sum of the variance components in (2.3) is finite if and only if \( a > 1 \) as otherwise the integral in (2.9) is divergent. Note that this also excludes propensity score densities that have uniform like tails \( (a = 1) \). For more intuition consider Figure 2.1. We focus on the part of the
distribution around zero. Shapes like these are very common in empirical research.\(^3\) Note that the dotted (yellow) and the dash-dotted (purple) curves are examples for which the propensity score density goes to zero at the bound while the uniform (dashed, red) or the solid (blue) exhibit a nonzero limit from the

\(^3\)Using the keywords inverse probability weighting and propensity score, we went through a selection of research papers published in major journals after the year 2000. 17 papers included histograms or densities of their estimated propensity scores. 35.29% of these articles show distributions with relevant probability mass close to the boundaries of the support, see e.g. Simonsen and Skipper (2006) or Lechner and Strittmatter (2017).
right. While all densities are supported on (0,1) and violate strong overlap, only tails like the solid (blue) or the dashed (red) curve will not yield a finite variance. Hence the asymptotic normality will break down only if extreme propensity scores have sufficient probability. However, a limit theorem for a stabilizing transformation without any trimming can still be obtained under additional assumptions. It is necessary to restrict the behavior of the density of the propensity score as \( p \) goes to zero from the right. In particular, if the propensity score density varies regularly at zero, a suitable normalized version of \( \hat{\theta}_n \) converges in distribution to an alpha-stable distribution with location parameter zero and tail index \( \alpha > 1 \). Regular variation of the propensity score density \( f_p(z) \) at zero means that

\[
\lim_{t \to 0^+} \frac{f_p(zt)}{f_p(t)} = h(z) \leq \infty
\]  

(2.10)

exists on a dense set and is finite for some \( z \). Then, \( h(z) \) must be of shape \( z^{\alpha-2} \) with \( 1 < \alpha \leq 2 \), which is equivalent to \( f_p(z) \sim z^{\alpha-2}l(1/z) \) as \( z \to 0 \). Regular variation rules out densities that oscillate infinitely often with increasing peak amplitude or that grow at super fast rates (e.g. factorial) on an even faster vanishing range as \( p \) goes to zero. We do not expect these extreme patterns to be relevant for empirical applications. The beta distribution is one example for a density that has tails according to (2.10). Thus, even if \( E[1/p(X_i)] \) is infinite, under (2.10) there exists a normalizing constant \( [n^{\alpha-1}/l(n)]^{1/\alpha} = o(\sqrt{n}) \) such that

\[
[n^{\alpha-1}/l(n)]^{1/\alpha} (\hat{\theta}_n - \theta_0) \xrightarrow{d} S(\alpha, \beta, 0, c),
\]  

(2.11)

where \( S(\alpha, \beta, 0, c) \) denotes a random variable with an alpha-stable distribution with tail index \( 1 < \alpha \leq 2 \), skewness parameter \( -1 \leq \beta \leq 1 \), location parameter 0, and scale parameter \( c > 0 \). For \( 1 < \alpha \leq 2 \), an alpha-stable random variable \( S(\alpha, \beta, \mu, c) \) has the characteristic function

\[
\varphi(t; \alpha, \beta, \mu, c) = \exp \left( it \mu - c^\alpha |t|^\alpha \left( 1 - i \beta \tan \frac{\alpha \pi}{2} \text{sgn}(t) \right) \right).
\]  

(2.12)

For \( \alpha < 2 \), the corresponding distributions are heavy-tailed with polynomial tail decay. The smaller the tail index \( \alpha \), the slower the tail decay. For \( \alpha = 2 \), the distribution collapses to the Gaussian. \( \beta \) determines the relative probability mass of left versus right tail. Thus, under irregular identification, the limiting distributions of estimators for counterfactual or treatment effect parameters have more mass in the tails compared to the normal distribution and are potentially asymmetric.

For illustration, we simulate an estimator for the potential outcome \( E[Y_i(1)] \) as in (2.1) using beta distributed propensity scores. Note again that the beta distribution is continuously distributed on (0,1) therefore violating strong overlap. The variance of the estimator in (2.2) is determined by the parameterization of the beta distribution. Consider first the density corresponding to the dash-dotted (purple) curve in
Figure 2.2: Distribution of Simulated $\hat{\theta}_n$

The Figure depicts a histogram for $\hat{\theta}_n$ based on 50000 Monte Carlo replications and probability density function estimates based on maximum likelihood estimation for the normal distribution and the alpha-stable distribution. The potential outcome is generated as $Y_i(1) = X_i + \epsilon_i$ with $X_i \sim U(1, 2)$ and $\epsilon_i \sim N(0, 1/12)$ for an equal signal to noise ratio. $p \sim B(2, 5)$ for the regular (left) and $p \sim B(0.5, 5)$ for irregular (right) design. The sample size is 1000.

Figure 2.1: $p \sim B(2, 5)$ which implies that $E[1/p(X_i)]$ is finite. Figure 2.2 depicts the simulated distribution of estimator (2.1) together with parametric density estimates assuming a normal and an alpha-stable distribution respectively. On the left panel one can see that the estimator is approximately normally distributed in line with the standard central limit theorem. The right panel of Figure 2.2 shows the distribution of the estimator and the density fit for $p \sim B(0.5, 5)$. One can show that for this design $E[1/p(X_i)]$ is not finite but the density of the propensity score is regularly varying with tail index $\alpha = 1.5$, i.e. the parameter $E[Y_i(1)]$ is irregularly identified. In this case, the rate of convergence will be approximately $n^{1/3}$ up to a slowly varying function.\textsuperscript{4} One can see that the normal distribution does not match the actual behavior of the treatment effect estimator while the stable distribution accurately captures the skewness and heavy tails that are induced through the probability weighting.\textsuperscript{5}

Although irregular identification still allows the conventional estimators to be consistent, it is not obvious how to conduct valid inference in cases where the limiting distribution of the treatment effect estimators belongs to the alpha-stable class. The nonparametric bootstrap by Efron (1979) is invalid if the data belongs to the domain of attraction of a stable law as the limiting distribution of the bootstrap mean is not the same as the limiting distribution of the mean from the original sample, i.e. the bootstrap distribution of the sample mean does not converge to a fixed distribution, but to a random probability distribution (Athreya, 1987, Knight, 1989). One possible way how to conduct valid inference for the mean with a limiting stable

\textsuperscript{4}For more details regarding the beta distribution and the associated tail index and convergence rates consider Supplementary Appendix B.1.3.

\textsuperscript{5}The wide fit of the normal distribution is explained by extreme observations that are outside the range of this plot. The axis limit was chosen for a better visual comparison of the distributions around the true value of $E[Y_i(1)]$. 

\hspace{1cm}
law is a modified $m$-out-of-$n$ bootstrap by (Bickel et al., 1997) that uses a test statistic with a normalization that adapts to the type of identification. The key parameter is the choice of resampling size $m$. We employ a bagging algorithm for the optimal $m$-choice based on the method by Bickel and Sakov (2008) that yields robust $m$-choices in finite samples and has superior second-order properties compared to fixed $m$ rules under regular identification. As the proposed resampling method is not only consistent for the stable limiting case but also yields asymptotically correct coverage rates under regular identification, the researcher can apply it without knowledge of the regularity of his or her identification problem. Further details on the inferential methods are presented in Section 4 below.

3. Large Sample Estimation Theory

This section provides the assumptions and limiting theory for both inverse probability weighting and doubly robust estimation. The distinction between these two classes is necessary as some assumptions on the underlying data generating process can be substantially weakened or even completely omitted when moving from inverse probability weighting to doubly robust estimation. For the remainder of this section, we focus on regular parametric binary response models for the propensity scores:

Assumption 3.1. Let $D_i(d) = \mathbb{1}(D_i = d)$. The propensity scores $P(D_i = d|X_i) = p_d(X_i, \gamma_0)$ are estimated as $p_d(X_i, \hat{\gamma})$ via a regular likelihood model:

$$\hat{\gamma} = \arg \max_{\gamma \in \Gamma} \sum_{i=1}^n \sum_d D_i(d) \ln p_d(X_i, \gamma)$$

with $\Gamma$ being a nonempty compact subset of $\mathbb{R}^k$ and

1. $p_d(X_i, \gamma_0)$ twice continuously differentiable around $\gamma_0$,
2. $E[s_{ij}^4] < \infty$ for all $j = 1, \ldots, k$,
3. $E[S_iS_i'] = -E[\partial S_i/\partial \gamma']$ with $S_i = S_i(\gamma_0) = (s_{i1}, \ldots, s_{ik})'$ being the score function evaluated at the true parameter $\gamma_0 \in \Gamma$.

Assumption 3.1 allows for methods commonly used in applied research such as logistic or probit regression. The information equality in iii) is not necessary for any of the main results but simplifies the asymptotic properties in the case of regular identification (Wooldridge, 2002, 2007). Alternatively, the asymptotic variance in the following subsection can be adjusted equivalently to quasi maximum likelihood estimation.

3.1. Inverse Probability Weighting

We consider inverse probability weighting with normalized weights (IPWII) which implies that the potential outcome mean estimates are within the range of the observable outcome variables in finite samples. It is commonly preferred over the non-normalized inverse probability weighting (IPWI) due to its superior finite sample properties (Busso et al., 2014, Pohlmeier et al., 2016). The theory for IPWI can be constructed
along the same lines. The IPWII estimator for the average treatment effect using estimated propensity scores is given by

$$
\hat{\theta}_n = \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(1)}{p_d(X_i, \tilde{\gamma})} \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(1)Y_i}{p_d(X_i, \tilde{\gamma})} - \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(0)}{p_d(X_i, \tilde{\gamma})} \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(0)Y_i}{p_d(X_i, \tilde{\gamma})}. \tag{3.1}
$$

We impose the following assumptions:

**Assumption 3.2 (Moments, IPW).** Let $k_{id} = (Y_i(d) - E[Y_i(d)])D_i(d)/p_d(X_i)$. For both $d = 0, 1$, there exist (potentially different) neighborhoods $N_{\gamma_0}$ around $\gamma_0$ such that

i) $E\left[ \sup_{\gamma \in N_{\gamma_0}} \left| \frac{D_i(d)}{p_d(X_i, \gamma)} \right| \right] < \infty$,

ii) $E\left[ \sup_{\gamma \in N_{\gamma_0}} \left| k_{id}S_i(\gamma) \right| \right] < \infty$,

iii) $\lambda_1(E[S_iS_i']^{-1}E[S_i k_{id} E[k_{id} S_i E[S_i S_i']^{-1}] < \infty$.

**Assumption 3.3 (Regular Identification, IPW).** The inverse probability weighted squared centered outcomes have finite second moment

$$
E\left[ \frac{(Y_i(d) - E[Y_i(d)])^2}{p_d(X_i)} \right] < \infty
$$

for both $d = 0, 1$.

**Assumption 3.4 (Irregular Identification, IPW).** The tails of the weighted outcomes are regularly varying with tail index $\alpha_d > 1$ for both $d = 0, 1$ and balanced

i) $U_d(z) \equiv E\left[ \frac{(Y_i(d) - E[Y_i(d)])^2}{p_d(X_i)} \left( \frac{|Y_i(d) - E[Y_i(d)]|}{p_d(X_i)} < z \right) \right] \sim z^{2-\alpha_d} l_d(z)$,

ii) $\lim_{z \to \infty} E\left[ \frac{|Y_i(d) - E[Y_i(d)]|}{p_d(X_i)} \right] p_d(X_i) = q_d$,

$$
\lim_{z \to \infty} E\left[ \frac{|Y_i(d) - E[Y_i(d)]|}{p_d(X_i)} \right] p_d(X_i) = r_d
$$

for $q_d + r_d = 1$; and the unweighted potential outcomes have finite second moment,

iii) $E[Y_i(d)^2] < \infty$.

Assumption 3.2 puts some regularity conditions on the score of the propensity score likelihood function and its interaction with the inverse probability weights. As the scores depend on inverse probability weights, the moment assumptions might seem restrictive. However, we provide some sufficient conditions in Supplementary Appendix B.1.1 for single index models that demonstrate that the assumption usually reduces to a simple moment condition for the confounding variables $X_i$. In general, these assumptions are substantially weaker than strong overlap. Assumptions 3.3 and 3.4 are mutually exclusive, i.e. we assume only one to hold. Assumption 3.3 is a standard regular identification assumption implied by the strong overlap condition. Assumption 3.4 restricts the behavior of the truncated second moment of the squared weighted outcome in i) and assures that the tail probabilities for positive and negative values, $q_d$ and $r_d$, have a constant ratio in the limit in ii), i.e. right and left tail are “balanced”. Note that there is a nontrivial impact of the potential outcome variable $Y_i(d)$ on the tail index $\alpha_d$. In particular, if the potential outcome diverges
as the propensity score goes to zero, the index of regular variation can be affected as well, i.e. $\alpha_d$ for both $d = 0, 1$ can potentially be closer to one compared to the index of the inverse propensity score only. Thus Assumption 3.4 is not only concerned with the behavior of the propensity score at zero but also with its interplay with the potential outcome. This is a particular feature of the IPW methodology in the context of irregular identification. One could in principle restrict the range of the potential outcomes conditional on the propensity score, however this would not be meaningful for many applications as it rules out commonly applied functional forms such as linear mean functions with continuous, unbounded regressors. The tail balance condition 3.4(ii) effectively restricts the distributions of the potential outcomes conditional on the propensity score. It is only concerned with how its left and right tail behave relative to each other as the propensity scores go to zero, i.e. at the extremes the occurrence of positive and negative deviations from the mean need to be in a constant ratio in the limit. A symmetric conditional distribution is sufficient but not necessary. Violation occurs if there is oscillation between different tail decay patterns at the extremes.\footnote{See Supplementary Appendix B.1.2 for more details and an example.}

The following proposition characterizes the large sample behavior of the IPWII estimator under general identification conditions:

**Proposition 3.1.** Let $\hat{\theta}_n$ be chosen according to (3.1). If Assumptions 3.1, 3.2, and either 3.3 or 3.4 with $\alpha_d < \alpha_{d'}$ for some $d \in \{0, 1\}, d' = 1 - d$ hold, then there exists a slowly varying function $l(\cdot)$ such that

$$[n^{\alpha_d - 2}/l(n)]^{1/\alpha} (\hat{\theta}_n - \theta_0) \rightarrow S(\alpha, \beta, 0, c),$$

where $S(\alpha, \beta, 0, c)$ denotes a random variable with an alpha-stable distribution with tail index $\alpha = \alpha_d \in (1, 2]$ and skewness parameter $\beta = (q_{d'} - r_{d'})/(q_{d'} + r_{d'}) \in [-1, 1]$, location parameter 0, and scale parameter $c > 0$.\footnote{The parameterization for the case of equal tail indices can be found in Appendix A.1.}

Proposition 3.1 shows that even if the propensity scores are allowed to be arbitrarily close to zero or if the second moments of the weighted potential outcomes are infinite, the IPWII estimator still has a well-defined limiting distribution in the alpha-stable class. Thus, weighting by the conditional probabilities is a transformation that leads to a potentially different limiting distribution than the Gaussian. Note that convergence can be slower than the usual root-n rate. Under irregular identification it only depends on the tail index of the inverse probability weighted potential outcome with the slower tail decay.\footnote{In this case $c = \sqrt{V/2}$ where $V$ denotes the asymptotic variance of the corresponding normal limiting distribution.} The smaller the tail index the slower the rate of convergence. The proposition also covers the case $\alpha = 2$, i.e. when the limiting distribution is normal. If Assumption 3.3 holds, the limiting distribution is normal with $[n^{\alpha_d - 2}/l(n)]^{1/2} = \sqrt{n}$.\footnote{However there are cases for which $\alpha = 2$ but $[n^{\alpha_d - 2}/l(n)]^{1/2} = o(\sqrt{n})$, see Supplementary Appendix B.1.3. Note that since $\alpha > 1$ the limiting stable distribution will always have a mean of 0, i.e. there is no asymptotic bias in the stabilizing transformation that occurs for some trimming estimators (Khan and Tamer, 2010). This limiting behavior is a consequence of the construction of the treatment effect estimators that use}
inverse probabilities and not driven by the tails of the potential outcome variables, which themselves are in the domain of attraction of the normal distribution. A thorough inspection of the proof for Proposition 3.1 in Appendix A.1 reveals that, contrary to a regularly identified problem, the first order asymptotics of the IPWII using estimated propensity scores are identical to using true propensity scores in the case of irregular identification. Thus, the result of increased efficiency using estimated propensities over true propensities as in Wooldridge (2002, 2007) has no parallel under irregular identification. This is due to the parametric propensity score model converging at the faster root-n rate. Thus, the projection of the weighted outcome onto the tangent space of the likelihood function, the source of the increased efficiency in the regular case, is dominated in probability by the weighted outcome with true weights and regularly varying tails.

3.2. Doubly Robust Estimation

Doubly robust estimation is a combination of regression adjustment for confounding variables and inverse probability weighting (Robins et al., 1994, Kang and Schafer, 2007). It yields consistent estimates of the causal effects even if either the propensity score or the regression function is misspecified. Moreover, the underlying moment function guards against learning bias and thus can yield root-n consistent and asymptotically normal treatment effect parameter estimates in the case of high-dimensional nuisance quantities and regular identification under weak conditions (Chernozhukov et al., 2017). Doubly robust estimates weigh the residuals of the potential outcome equations by the inverse propensity scores and hence are subject to the small denominator critique as well.

Let \( E[Y_i(d)|X_i] = \mu_d(X_i, \delta_d) \) for \( d = 0, 1 \) with \( \delta_d \) being in a nonempty compact subset of \( \mathbb{R}^k \). In the following we consider a typical doubly robust estimator based on the efficient influence function (Robins and Rotnitzky, 1995, Hahn, 1998) with unknown quantities replaced by sample estimates:

\[
\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} D_i(1)(Y_i - \mu_1(X_i, \hat{\delta}_1)) \frac{p_1(X_i, \hat{\gamma})}{p_1(X_i, \hat{\gamma})} - \frac{1}{n} \sum_{i=1}^{n} D_i(0)(Y_i - \mu_0(X_i, \hat{\delta}_0)) \frac{p_0(X_i, \hat{\gamma})}{p_0(X_i, \hat{\gamma})} + \frac{1}{n} \sum_{i=1}^{n} \left( \mu_1(X_i, \hat{\delta}_1) - \mu_0(X_i, \hat{\delta}_0) \right).
\]

(3.2)

We first present assumptions and limit characterizations parallel to those of the inverse probability weighting estimator and then consider additive separable outcome models with bounded heteroskedasticity. We show that for the latter, the conditions on the tail behavior can be substantially weakened and the limiting distributions can be characterized more precisely. Consider the following assumptions:

**Assumption 3.5 (Outcome Model, DR).** The mean functions \( \mu_d(X_i, \delta_d) \) are estimated as \( \mu_d(X_i, \hat{\delta}_d) \) via a regular parametric model:

i) \( \mu_d(X_i, \delta_d) \) continuously differentiable around \( \delta_d \),

ii) \( \delta_d - \hat{\delta}_d = O_p(n^{-1/2}) \)

for both \( d = 0, 1 \).
Assumption 3.6 (Moments, DR). For both $d = 0, 1$, there exist (potentially different) neighborhoods $N_{\gamma_0}$ around $\gamma_0$ and $N_{(\gamma_0, \delta_d)}$ around $(\gamma_0, \delta_d)$ such that

$$E \left[ \sup_{\gamma \in N_{\gamma_0}} \left\| \frac{D_d(\gamma)}{p_d(X_i, \gamma)} \right\| \right] < \infty,$$

$$E \left[ \sup_{\gamma \in N_{(\gamma_0, \delta_d)}} \left\| \frac{D_d(\gamma) - E[Y_i(d) | X_i]}{p_d(X_i, \gamma)} S_i(\gamma) \right\| \right] < \infty,$$

$$E \left[ \sup_{(\gamma, \delta) \in N_{(\gamma_0, \delta_d)}} \left\| \frac{\partial p_d(X_i, \delta_d)}{\partial \delta_d} \frac{D_d(\gamma)}{p_d(X_i, \gamma)} S_i(\gamma)' \right\| \right] < \infty.$$  

Assumption 3.7 (Regular Identification, DR). The weighted conditional mean errors have finite second moment

$$E \left[ \frac{(Y_i(d) - E[Y_i(d) | X_i])^2}{p_d(X_i)} \right] < \infty$$

for both $d = 0, 1$.

Assumption 3.8 (Irregular Identification, DR). The tails of the weighted conditional mean errors are regularly varying with tail index $\alpha_d > 1$ for both $d = 0, 1$ and balanced

$$U_d(z) \equiv E \left[ \frac{(Y_i(d) - E[Y_i(d) | X_i])^2}{p_d(X_i)} I[\frac{Y_i(d) - E[Y_i(d) | X_i]}{p_d(X_i)} < z] \right] \sim z^{2 - \alpha_d} I_d(z),$$

$$\lim_{z \to \infty} \frac{E[|Y_i(d) - E[Y_i(d) | X_i]|/p_d(X_i) > z] p_d(X_i)}{p_d(X_i)} = q_d,$$

$$\lim_{z \to \infty} \frac{E[|Y_i(d) - E[Y_i(d) | X_i]|/p_d(X_i) < -z] p_d(X_i)}{p_d(X_i)} = r_d$$

for $q_d + r_d = 1$; and the unweighted potential outcomes have finite second moment,

$$E|Y_i(d)|^2 < \infty.$$

Assumption 3.6 puts some constraints on the interactions between the score of the likelihood function and both weighted conditional mean error and conditional mean function. As for IPW, the assumptions are substantially weaker than strong overlap. Assumption 3.5 imposes a regular parametric model for the outcome. Assumptions 3.7 and 3.8 collect the cases of regular and irregular identification. Note that the tail conditions in Assumption 3.8 are usually weaker than the corresponding Assumption 3.4 used for inverse probability weighting. In particular, instead of regular variation of the truncated second moment of the weighted potential outcome, only regular variation of the truncated second moment of the weighted conditional mean error $\{Y_i(d) - E[Y_i(d) | X_i]\}$ is required. As the latter usually has a smaller truncated second moment, the index of regular variation $\alpha_d$ for doubly robust estimation will be larger or equal to the case of inverse probability weighting. Thus estimation will concentrate around the true values at an equal or faster rate of convergence.\(^9\) If the potential outcome or its conditional mean error converges to a nonzero bound at the extreme propensities, then the index of regular variation $\alpha_d$ in Assumptions 3.4, i) or 3.8, i) is equivalent to the index of the more simple $E[1/p_d(X_i) I(1/p_d(X_i) < z)]$. We will exploit this insight in Section 4 for optimizing the inferential procedure.

\(^9\)The only exception is the very hypothetical case of potential outcome means converging to zero as propensities go to zero as this can imply slower divergence of the truncated variance in Assumption 3.8, i) and thus a larger tail index.
The following proposition contains the limiting distribution of the doubly robust estimator of the average treatment effect under general identification conditions:

**Proposition 3.2.** Let \( \hat{\theta}_n \) be chosen according to (3.2). If Assumptions 3.1, 3.5, 3.6, and either 3.7 or 3.8 hold with \( \alpha_d < \alpha_{d'} \) for some \( d \in \{0, 1\}, d' = 1 - d \) hold, then there exists a slowly varying function \( l(\cdot) \) such that

\[
[n^{-1}(l(n))]^{\frac{1}{\alpha}} (\hat{\theta}_n - \theta_0) \xrightarrow{d} S(\alpha, \beta, 0, c),
\]

where \( S(\alpha, \beta, 0, c) \) denotes a random variable with an alpha-stable distribution with tail index \( \alpha = \alpha_d \in (1, 2) \) and skewness parameter \( \beta = (q_d - r_d)/(q_d + r_d) \in [-1, 1] \), location parameter 0, and scale parameter \( c > 0 \).

Similarly to inverse probability weighting, the doubly robust estimator has a well-defined limiting distribution that can differ from the Gaussian in the case of irregular identification with convergence rates possibly slower than the parametric rate. It is also possible to relax Assumption 3.5 such that convergence of the parameter estimates occurs at a slower than root-n rate in the case of irregular convergence, i.e. if \( [n^{-1}(l(n))]^{\frac{1}{\alpha}} \) is not equal to \( \sqrt{n} \). For this case it is sufficient that the product of estimated propensity score parameters \( (\hat{\gamma} - \gamma_0) \) and estimated outcome model parameters \( (\hat{\delta}_d - \delta_d) \) is \( o([n^{-1}(l(n))]^{\frac{1}{\alpha}}) \). Thus, for a small \( \alpha \) and hence a slow convergence rate, estimation of the nuisance parameters can be slower as well. Depending on the tail index, this allows for less restrictive conditions on nonparametric estimation of the nuisance quantities (Rothe and Firpo, 2018) or on learning rates for high dimensional nuisance functions and model selection methods such as \( \ell_1 \)-regularization with sparsity conditions (Belloni et al., 2014) or general machine learning methods (Chernozhukov et al., 2017). However without knowledge of the tail index or the normalizing constant we still recommend to stick with the conditions and methods from the aforementioned papers as they work reliably in the case of regular identification. We leave a potential extension of our method to the high-dimensional or nonparametric case along the lines of Chernozhukov et al. (2017) for future research.

It is important to note that Assumptions 3.8 i) and ii) can be simplified substantially if one is willing to impose an additive separable structure with bounded multiplicative conditional heteroskedasticity for the potential outcomes, i.e. if

\[
Y_i(d) = \mu_d(X_i, \delta_d) + \sigma_d(X_i)\varepsilon_i(d)
\]

with \( \varepsilon_i(d) \) and \( X_i \) independent and \( \sigma_d(X_i) \) bounded, see Supplementary Appendix B.1.3 for more details.

Under these assumptions the tail balance condition can be completely omitted and regular variation is only required for the propensity score density, i.e. we simply need to assume that \( f_{\rho_d}(z) \sim z^{\alpha_d - 1}l(1/z) \). Then the skewness parameter of the limiting distribution is solely determined by the ratio between negative and positive part \( \alpha \)-th moment of the independent component \( \varepsilon_i(d) \) and does not depend on the conditional potential outcome mean. For inverse probability weighting estimators, however, this simplification would only occur in the case of homogeneous, i.e. constant potential outcomes. Thus, under additive separability, the tail conditions permit a much larger class of outcome structures if a doubly robust approach is used.

In the following we provide an example about the relationship between the behavior of the propensity
score density at the boundary, the limiting distribution, and the corresponding convergence rate. Assume that the propensity scores follow a beta distribution, i.e. \( p_d(X_i) \sim \mathcal{B}(a, b) \) for some \( a, b > 0 \), see also Figure 2.1. Under the assumptions above one obtains for \( a \leq 1 \)

\[
\left[ n^a/l(n) \right]^{1/\alpha} (\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{S}(a + 1, \beta, 0, c),
\]

while for \( a > 1 \)

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, V),
\]

which is in line with the previous propositions as \( a > 0 \) and thus \( \alpha > 1 \) for the limiting stable distribution.\(^{10}\)

From (3.4) and (3.5) we can see how the tail behavior of the propensity score is linked to the rate of convergence. The faster the density diverges, i.e. the smaller the \( a \) parameter, the slower the rate of convergence. The latter can be almost arbitrarily slow if the \( a \) parameter gets close to zero or equivalently if the left tail of the beta distribution approaches a function resembling the behavior of \( 1/p_d \).

An interesting special case is \( a = b = 1 \), i.e. when the beta distribution is equal to the uniform. It then follows that

\[
\sqrt{n} \left( \frac{\hat{\theta}_n - \theta_0}{\ln(n)} \right) \xrightarrow{d} \mathcal{S}(2, \beta, 0, c) \xrightarrow{d} \mathcal{N}(0, V).
\]

Note that this coincides with the rate and distribution obtained by Khan and Tamer (2010) for a trimmed IPW estimator under distributional assumptions for confounder and latent error of a selection equation that also yield a uniform distribution for the propensity score. Khan and Tamer (2010) assume a logistic propensity score model with a single logistic regressor that enters with a unit coefficient\(^{11}\) and apply a theoretically optimal trimming rate to get a limiting theorem similar to (3.6) but containing a limiting bias. Our limit theorem is valid without any trimming.\(^{12}\)

In general the propensity score densities that go to a nonzero constant from the right as \( p_d \) goes to zero, such as the uniform density, can be considered as interesting intermediate cases because the estimators do converge to a normal distribution but not at the usual root-n rate. For a zero limit as \( p_d \) goes to zero, regular convergence is achieved. A diverging propensity score density together with assumption of regular variation yields a stable limiting distribution at a strictly slower rate than root-n. If both tails diverge, the rate of convergence is dictated by the tail of the propensity score density that diverges faster.

\(^{10}\)See Supplementary Appendix B.1.4 for the derivations.

\(^{11}\)Any threshold crossing model with index and error that have identical tails yields a uniform distribution for the propensity score.

\(^{12}\)Khan and Tamer (2010) assume only bounded conditional average treatment effects. Our additive separable model restricts the class of distributions for the potential outcome. However the same result can also be replicated under a boundedness assumptions together with convergence of the conditional CDF of the potential outcome as \( p_d \to 0 \), see Supplementary Appendix B.1.2.
3.3. Misspecified Propensity Score induced Irregularity

Previously we only considered irregularities arising from the distribution of the true propensity scores under regularly varying tails. However, a change in the limiting domain can also be obtained through simple misspecification of the propensity score model even if the misspecified model still yields consistent estimates of the treatment effect parameters. The parameters of the propensity score model under misspecification usually converge to pseudo-true parameters often being the minimizers on the Kullback-Leibler information criterion. If the density of the corresponding pseudo-true propensity scores has regularly varying tails at the boundaries, then the limiting distribution will be affected accordingly independent of whether the treatment effect is estimated consistently or not. The latter is particularly relevant for homogeneous treatment effects and more importantly for doubly robust estimation with misspecified propensity scores. Under a misspecified propensity score model, the DR estimator remains consistent if the potential outcome models are chosen correctly. The limiting distribution, however, will be altered.

Consider the following example of a boundary case with a limiting normal distribution with uniformly distributed propensity scores and compare it to a misspecified model that uses a logistic regression. For the characterization of the limiting behavior, it is sufficient to only consider the tail of the density at zero, as the density of the misspecified propensity scores will be bounded from above for larger scores in this example. Let $X_i \sim U(0,1)$. Assume that $p_1(x) = x$, i.e. uniformly distributed propensity scores. As shown in the previous subsection, a treatment effect estimator using a correctly specified propensity score model yields a limiting normal distribution with convergence rate $\sqrt{n/\ln(n)}$. However now assume the propensity scores are estimated using the following misspecified logistic model with $\ln(x)$ as regressor:

$$p_1^{[m]}(x, \gamma) = \frac{1}{1 + \exp(-\gamma_0 - \gamma_1 \ln(x))}.$$  

(3.7)

Figure 3.1 contains the simulated distributions for the true and misspecified propensity scores estimated via maximum likelihood using $n = 10000$ observations. The densities of the propensity scores differ from each other particularly at the tails: the density of the misspecified scores is increasing as scores get closer to zero. In fact, the misspecified inverse probability weights have regular varying probability tails with tail index $\alpha = 2/\gamma_1$, see Supplementary Appendix B.2.1. However, $\gamma_1$ is not arbitrary as the misspecified likelihood estimator will yield parameter estimates concentrating around the pseudo-true parameters $(\gamma_0^*, \gamma_1^*)$ given by the minimizers of the Kullback-Leibler information criterion:

$$(\gamma_0^*, \gamma_1^*) = \arg \min_{(\gamma_0, \gamma_1)} KLIC(f_{p_1^{[m]}}, f_{p_1}) \approx (1.6807, 1.9835).$$  

(3.8)

Thus with probability approaching one, the tail index is $\alpha = 2/1.9835 \approx 1.008$ which corresponds to a heavy-tailed limiting distribution with a vastly slower convergence rate compared to the correctly specified
model that uses uniformly distributed scores with a limiting normal distribution or a tail index $\alpha = 2$.

In practice we will almost always use a somewhat misspecified model. Now if there are different candidate models for the propensity score available that seem to sufficiently balance the distributions of covariates across treatment and control regimes, then consistency is not a major concern. However models with similar balance could still lead to different limiting domains for the treatment effect estimator. Selecting a model that avoids or at least reduces irregularity would be beneficial as it leads to narrower confidence bands and thus more precise inference for the treatment effect parameters. Graphical inspection of the propensity score densities as discussed in Section 2 provides some insight into the (ir)regularity of the problem. However, for applied work, the behavior of estimated propensity score densities at the boundaries will be very sensitive to smoothing or bin parameter choices. Moreover, under irregular identification, the true propensity score density diverges at the boundaries and thus conventional methods for density estimation will be severely biased downwards (Bouezmarni and Rolin, 2007). However, the overlap robust limit characterizations in Propositions 3.1 and 3.2 allow for developing a model selection heuristic based on the tail behavior that favors limiting domains characterized by less irregularity, i.e. by a larger tail index. This helps to select a propensity score model more suitable for causal inference, in particular if conventional criteria such as predictive quality or standardized differences are inconclusive. It also applies naturally to any doubly robust approach with a correctly specified outcome model.

We propose to use an estimate of the tail index $\alpha$ for the inverse probability weights for model selection. Assume we have $M$ candidate models for the propensity scores:

1. For each $m = 1, \ldots, M$
   i) estimate the propensity scores $\hat{p}_{i[m]}(X_i)$. 
ii) construct the inverse probability weights

\[ \hat{w}_i^{[m]} = \frac{D_i(1) - \hat{p}_i^{[m]}(X_i)}{\hat{p}_i^{[m]}(X_i)(1 - \hat{p}_i^{[m]}(X_i))} \]

for \( i = 1, \ldots, n \).

iii) estimate the tail-index \( \hat{\alpha}^{[m]} \) of the inverse probability weights \( \hat{w}_1^{[m]}, \ldots, \hat{w}_n^{[m]} \).

2. Select model \( m \) with \( \hat{\alpha}^{[m]} = \max\{\hat{\alpha}^{[1]}, \ldots, \hat{\alpha}^{[M]}\} \).

For more details regarding the estimation of the tail index and an application to the right heart catheterization data used in Section 6 consider Supplementary Appendix B.2.2. The procedure selects the model with the least amount of irregularity with probability going to one. It is not monotone in the most extreme propensity score but takes into account the overall decay of probability mass at the extreme weights. The procedure chooses the maximum instead of conducting statistical tests between different models as there is no obvious relationship between covariate balance and irregularity, i.e. it could well be that a less irregular model is no worse or better in correcting for selectivity. Note that while the tail index model selection procedure can help to select a model that avoids irregularities it is not a replacement for criteria that evaluate the fit or more importantly covariate balance of the model. However, it serves as an additional tool for finding a more robust model for causal inference.

4. Statistical Inference

In the following we demonstrate how to use the results from Section 3 to construct asymptotically valid confidence bounds and p-values for testing the statistical significance of treatment effect parameters. The goal is to conduct a two-sided hypothesis test for the parameter of interest \( \theta \) against the null-value \( \theta_0 \), i.e. \( H_0 : \theta = \theta_0 \) versus \( H_1 : \theta \neq \theta_0 \). Consider the IPW or DR estimator \( \hat{\theta}_n \) chosen according to (3.1) or (3.2) respectively. In principle, one could invert the limiting alpha-stable distributions in Propositions 3.1 and 3.2 with unknown parameters replaced by consistent sample estimates. In the case of regular identification, this would yield the typical confidence bounds based on asymptotic normality of \( \sqrt{n}(\hat{\theta}_n - \theta_0) \). For the general case, the same could be done for the three unknown parameters in the stable limiting distributions \( \alpha, \beta, \) and \( c \).\footnote{See e.g. Cornea-Madeira and Davidson (2015) on guidance on estimating the parameters of the stable distribution for an i.i.d. location parameter model.}

Precise estimation of the tail index \( \alpha \) tends to require large samples and additional tuning parameters (Hill, 1975, Hall, 1990, Danielsson et al., 2001). Moreover, the unknown normalizing constant \( [n^{\alpha-1}/l(n)]^{\frac{1}{2}} \) depends not only on the true tail index \( \alpha \) but on a slowly varying function which can be very
difficult to estimate in finite samples (Bertail et al., 1999).\textsuperscript{14} Therefore instead we exploit the properties of self-normalizing sums following Logan et al. (1973). Consider the following test statistic:

\[ T_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\hat{V}_n[\hat{\theta}_n]}}, \]  

(4.1)

with \( \hat{V}_n[\hat{\theta}_n] \) being a consistent estimate for the asymptotic variance of the stabilizing transformation under regular identification, see Supplementary Appendix B.3.1. The following theorem shows that under regular identification, the statistic (4.1) converges in distribution to a Gaussian while under irregular identification it is suitably normalized to have a well defined limiting-distribution:

**Theorem 4.1.** Let \( \theta_0 \) be the true average treatment effect and \( \hat{\theta}_n \) be chosen according to (3.1) \((3.2)\). Under Assumptions 3.1, 3.2 \([3.1, 3.5, 3.6]\),

i) and 3.3 \([3.7]\), the normalized statistic (4.1) has limiting distribution

\[ T_n \overset{d}{\rightarrow} N(0,1). \]

ii) and 3.4 \([3.8]\), the normalized statistic (4.1) has limiting distribution

\[ T_n \overset{d}{\rightarrow} T_{\alpha,\beta} \]

with \( T_{\alpha,\beta} \) being a random variable with continuous density \( f(x;\alpha,\beta) \) that has local maxima at \( x = \pm 1 \) and symmetry for \( |x| < 1 \).

The distribution of the limit under irregular identification converges smoothly to the normal distribution as \( \alpha \to 2 \) but is generally complicated to evaluate (Logan et al., 1973).\textsuperscript{15} However, since independently of the identification assumption the self-normalizing sum has a well-defined limit, Theorem 4.1 suggests the use of resampling methods that are valid within the stable domain of attraction. The nonparametric bootstrap by Efron (1979) is not applicable as it fails to approximate the distribution of \( T_n \) in the case of regular varying tails with a tail index smaller than two. In particular, Athreya (1987) and Knight (1989) show that for a sequence of i.i.d. random variables belonging to the alpha-stable domain of attraction with \( 1 < \alpha < 2 \), the bootstrap distribution of the test statistic \( T_n \) converges in distribution to a random limit. This implies that the nonparametric bootstrap can produce almost arbitrary and unpredictable distortions in the coverage rate of the corresponding confidence bounds under irregular identification, i.e. significant over- or under-rejections can occur.

The failure of the nonparametric bootstrap can be overcome by evaluating the distribution of resampled test statistic with a sample size \( m = o(n) \). We consider the \( m \)-out-of-\( n \) bootstrap as a feasible alternative.

\textsuperscript{14}In principle the rate of convergence (including its slowly varying function part) can be approximated asymptotically with an additional resampling step under mild regularity conditions, see Romano and Wolf (1999) and Bertail et al. (1999). However, for statistical inference, methods using these estimates have poor second order properties and typically require extreme sample sizes to produce nominal coverage rates.

\textsuperscript{15}The case distinction for i) is included only for expositional reasons but could well be summarized under \( T_{\alpha,\beta} \) for \( \alpha = 2 \).
to a direct inversion of the analytical limiting distributions. Alternatively subsampling approaches (Politis et al., 1999, Romano and Wolf, 1999) can be applied. However, they have worse second-order properties under regular identification, see the discussion below. The $m$-out-of-$n$ bootstrap approximates the null-distribution of the self-normalizing sum (4.1) by a data-centered bootstrapped test statistic computed over bootstrap samples of size $m$:

$$T_{m,n} = \sqrt{m} \frac{\hat{\theta}_n^m - \hat{\theta}_n}{\sqrt{V_m[\hat{\theta}_n^m]}}$$

(4.2)

where $\hat{\theta}_n^m$ denotes the treatment effect estimator using the $m$-out-of-$n$ bootstrap sample and $V_m[\hat{\theta}_n^m]$ corresponds to a consistent estimator of the asymptotic variance under regular identification using only $m$ observations. The detailed procedure is given below. The results in Bickel et al. (1997) together with Theorem 4.1 imply that if $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$ and if $m/n \to 0$ for $m,n \to \infty$, then the bootstrap distribution of the $m$-out-of-$n$ test statistic (4.2) converges uniformly to the limiting distribution of (4.1). Note that this includes the case of regular identification, i.e. asymptotic validity of the inferential procedure is guaranteed without knowledge of the degree of overlap. Uniformly here means that for all $(\alpha, \beta) \in (1,2] \times [-1,1]$, $\lim_{n \to \infty} \sup_x |F_{n,\alpha,\beta}(x) - F_{\alpha,\beta}(x)| = 0$ with $F_{n,\alpha,\beta}(\cdot)$ and $F_{\alpha,\beta}(\cdot)$ denoting the cumulative distribution function of $T_{m,n}$ and $T_{\alpha,\beta}$ respectively. Thus convergence is pointwise in $(\alpha, \beta)$.\(^{16}\)

It is important to note that asymptotic validity of the $m$-out-of-$n$ bootstrap under irregular identification is guaranteed only if $m/n \to 0$. Therefore, the choice of $m$ is crucial as many sequences fulfill the asymptotic requirement $m = o(n)$ but yield basically zero power in finite samples. Moreover, fixed $m$-rules are not adaptive to the different identification assumptions. Thus, we propose to select $m$ by a modified version of the method of Bickel and Sakov (2008).\(^{17}\) They propose to compare (a simulated version of the) finite sample distributions of the test statistic of interest along a grid of resampling sizes $m_j$. The optimal resampling size $\hat{m}^*$ is then chosen as the argument that minimizes a metric consistent with convergence in distribution such as the Kolmogorov distance. Evaluating the exact empirical finite sample distributions for resampling sizes $m_j$ is generally computationally infeasible as the number of possible data permutations is of order $n^m$. However, using a smaller number of Monte Carlo iterations for the method by Bickel and Sakov (2008) can yield unstable $\hat{m}$-choices for a given sample. In order to reduce the variance of the $\hat{m}$-choice procedure for a given sample we propose to use bootstrap aggregation, i.e. to simulate the choice of the $m$ repeatedly and to use the mean of the selected resampling sizes as the optimal $\hat{m}$. This yields estimates closer to the infeasible solution that uses the empirical distributions of all possible permutations. Using the full set of

\(^{16}\)This, together with the fact that convergence is generally slower for smaller tail indices, implies that large sample sizes are required for a good approximation if tail indices are very close to one. For more details regarding uniformity and sample requirements for different tail indices please consider Appendix B.3.2.

\(^{17}\)Note that under correctly chosen $m$, the $m$-out-of-$n$ resampling would be valid even for partial domains of attraction (Feller, 1968, Hall and LePage, 1996). However, the method by Bickel and Sakov (2008) requires that the bootstrap distribution of the test statistic for the sequence of resampling sizes has a well-defined limit. Thus Theorem 4.1 allows for a consistent application of their method.
permutations, Bickel and Sakov (2008) show that for the estimated resampling size $\hat{m}^*$, $\hat{m}^*/n \to 0$ if the nonparametric bootstrap fails and $\hat{m}^*/n \to 1$ if the bootstrap is valid. The same applies straightforwardly to our method as it is constructed by taking averages over $\hat{m}^*$ that are drawn from the same distribution. Note that this implies that, in contrast to subsampling, the method can adopt the desirable second-order properties of the nonparametric bootstrap for the pivotal test statistic (4.1) under regular identification (Hall, 1992). In finite samples under regular identification selection of $\hat{m}$ is unlikely to happen with probability one. Thus, we expect some power loss compared to the nonparametric bootstrap in this case.

In a nutshell, the inference procedure consists of the following steps:

1. Given sample $(Y_i, D_i, X'_i)$, $i = 1, ..., n$, compute $T_n$.
2. Choose $m$-grid of the form $m_j = [q^j n]$, $0 < q < 1$, $j = 0, 1, 2, ...$
3. for $s$ from 1 to $S$
   
   Select $m^{[s]}$ according to:
   
   i) For $b_1 = 1, ..., B_1$ and each element $m_j$ in the $m$-grid, draw samples of size $m_j$ with replacement from $(Y_i, D_i, X'_i)$ and calculate $T_{m_j}^{[b_1]}$.
   
   ii) For each $j$, let $L_{[m_j]}$ denote the empirical distribution of $(T_{m_j}^{[b_1]}, ..., T_{m_j}^{[B_1]})$.
   
   iii) Choose $m^{[s]} = \arg \min_{m_j} \rho(L_{[m_j]}, L_{[m_{j+1}]}), \rho$ being a metric consistent with convergence in distribution.
4. Calculate the optimal $\hat{m} = \frac{1}{S} \sum_{s=1}^{S} m^{[s]}$
5. For $b_2 = 1, ..., B_2$, draw $\hat{m}$ observations with replacement from $(Y_i, D_i, X'_i)$ and calculate $T_{\hat{m}, n}^{[b_2]}$.
6. Calculate the p-value according to $\frac{1}{B_2} \sum_{b_2=1}^{B_2} \mathbb{1} (|T_{\hat{m}, n}^{[b_2]}| > |T_n|)$.

It is important to note that within each iteration, $b_1$ and $b_2$, the nuisance parameters are reestimated as well, i.e. draws are not using fixed propensity scores or potential outcome predictions. Instead of calculating the p-value directly, one can also invert the empirical quantiles of the second bootstrap step 6 to construct confidence intervals in the usual manner. For the grid, we usually set $q = 0.75$. In practice, we also recommend to set a lower bound for the grid in step 2 such that $m_j > k[\ln(n)/\max\{1, \ln(k)\}]^2$ with $k$ being the dimensionality of the confounding vector. This assures that $m_j > k$ for all $j$ and does not affect the asymptotic behavior as the optimal $m$ is polynomial in $n$ for tail index $\alpha > 1$. Instead of the mean, other aggregation schemes for step 3 can also be applied. The number of bootstrap replications $B_2$ and especially $B_1$ should generally be large. For estimating $\hat{m}$, results tend to stabilize after $S \approx 100$ iterations.

Note that the test statistic in step 3 which is used to to determine the $m$ for which the empirical distributions are close to each other does not necessarily have to coincide with the test statistic used in step 5 for the causal parameter inference. For instance, if the potential outcome mean functions are bounded, the tail index and thus the rate of convergence for both IPW and DR estimators can be recovered from the inverse probability weights only, see the discussion in Section 3.2. This allows for more precise finite sample
$m$-choices as it removes the noise coming from the outcome data $Y_i$. For more guidance on alternative choices for $T_{m_j}$ see Appendix B.3.3.

5. Monte Carlo Simulations

In this section we investigate the finite sample coverage rates of the confidence intervals based on the conventional and the overlap robust inference methods outlined in previous sections. We consider three cases: regular identification, irregular identification, and irregular identification originating from misspecification. For additional designs and power properties simulations consider Appendix B.3.2. We use a simulation design with heterogeneous treatment effects depending on the propensity scores similar to Frölich (2004) and Busso et al. (2014):

$$
Y_i(0) = \ln (0.02 + 0.05 p(X_i)) + \varepsilon_i(0), \quad D_i = 1(\gamma X_i - u_i > 0),
$$
$$
Y_i(1) = \ln (0.01 + 0.25 p(X_i)) + \varepsilon_i(1), \quad Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0),
$$

where $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are independent standard normal errors and $u_i$ are standard logistic. Therefore $p(X_i) = \Lambda(\gamma X_i)$ with $\Lambda(\cdot)$ being the standard logistic cdf. The support of the propensity score distribution and the limiting distribution of the treatment effect estimator $\hat{\theta}_n$ crucially depends on the relative tail behavior of the scaled regressor $\gamma X_i$ and the error term $u_i$. We consider a standard logistic $\Lambda(0,1)$ distribution for $X_i$ together with different scaling parameters $\gamma$, which determine the shape of the propensity score distribution and thus the (ir)regularity of the identification.

Figure 5.1: DGP: Propensity Score Densities and Conditional Treatment Effect

The left panel depicts the theoretical density of propensity score depending on $\gamma$ in $p(X_i) = \Lambda(\gamma X_i)$ and $X_i \sim \Lambda(0,1)$. The right panel depicts the conditional average treatment effect for different propensity scores as defined in (5.1).
The left panel in Figure 5.1 depicts the true density functions of the propensity score for different values of \( \gamma \). The dotted (black) line for \( \gamma = 0.5 \) corresponds to the regularly identified case for which the limiting distribution of \( \hat{\theta}_n \) is root-\( n \) normal. The solid (red) line for \( \gamma = 1.5 \) depicts the distribution of the propensity score corresponding to an alpha-stable limiting law for the average treatment effect estimator with a tail index \( \alpha = 5/3 \). The right panel in Figure 5.1 depicts the conditional average treatment effect as a function of the propensity score. The functional form is chosen to introduce some heterogeneity in treatment effects for extreme propensity scores.

We consider several methods of testing the null hypothesis \( H_0 : \theta = \theta_0 \): symmetric nonparametric percentile bootstrap of Efron and Tibshirani (1994), denoted as i.i.d., and the \( m \)-out-of-\( n \) bootstrap with aggregated \( m \)-choice as described in Section 4. We also consider three trimming approaches: fixed trimming of propensity scores outside \([0.1, 0.9]\), data-driven efficiency optimal trimming as in Crump et al. (2009), and relative weight trimming by Imbens (2004) and Huber et al. (2013), which removes observations whose inverse probability weights are larger than 4\% of the sum of the weights over the whole sample. The p-values of the trimming procedures are computed via i.i.d. bootstrap ignoring the uncertainty in the trimming as suggested by Crump et al. (2009). For all inference methods we use \( B_2 = 500 \) bootstrap iterations and for the modified bootstrap aggregation method for choosing an optimal block length \( \hat{m} \) we set \( q = 0.75 \), \( S = 100 \), and \( B_1 = 10000 \).

First, we estimate the propensity scores using a correctly specified logistic propensity score model. For the doubly robust estimator defined in (3.2) we also assume a correctly specified mean function estimated with linear GMM. In the first step we repeatedly draw data \((Y_i, D_i, X_i')\) with replacement to obtain bootstrap samples of the corresponding length. In the second step, for a given random sample, we estimate a logistic regression and compute the estimate of the average treatment effect with IPW estimator (3.1) or DR estimator (3.2) based on \( p(X_i, \hat{\gamma}) \) and the appropriately normalized test statistic, see Supplementary Appendix B.3.1. Therefore, within the \( m \)-out-of-\( n \) bootstrap the propensity score will be estimated less precisely using only \( \hat{m} \leq n \) observations compared to the i.i.d. bootstrap.

Table 5.1 reports the empirical null rejection rates over 12000 simulations for the nominal 5\% level of significance for the case of regular identification (\( \gamma = 0.5 \)). The upper block corresponds to the rejection rates of the IPW estimator and the lower block to the rejection rates of the DR estimator of the average treatment effect under the null hypothesis. All resampling methods have rejection rates close to the nominal 5\% level, except for the slight over-rejection of some trimming methods. For trimming-based approaches the average treatment effect of the trimmed sample no longer corresponds to the ATE of the full population.

---

18 We have also considered alternative methods for heavy-tail robust inference such as the alpha-parametric bootstrap by Cornea-Madeira and Davidson (2015). However, it is not suited to incorporate the estimation of the propensity scores sufficiently, thus results are not reported here. The results are available from the authors upon request.

19 For the i.i.d. bootstrap the length of the bootstrap sample is \( n \), for the trimming approaches it is the length of the original trimmed sample and for the \( m \)-out-of-\( n \) bootstrap it equals to the optimally selected \( \hat{m} \).
Table 5.1: Null Rejection Rates for Nominal 5% Level under Regular Identification

<table>
<thead>
<tr>
<th></th>
<th>n=500</th>
<th>n=1000</th>
<th>n=5000</th>
<th>n = 10000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IPW</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.i.d.</td>
<td>4.60%</td>
<td>5.25%</td>
<td>5.27%</td>
<td>4.94%</td>
</tr>
<tr>
<td>trim. fix.</td>
<td>4.46%</td>
<td>5.24%</td>
<td>6.25%</td>
<td>6.56%</td>
</tr>
<tr>
<td>trim. Crump (2009)</td>
<td>4.26%</td>
<td>5.36%</td>
<td>7.03%</td>
<td>8.06%</td>
</tr>
<tr>
<td>trim. Huber (2013)</td>
<td>4.40%</td>
<td>5.25%</td>
<td>5.28%</td>
<td>4.94%</td>
</tr>
<tr>
<td>m-out-of-n</td>
<td>4.17%</td>
<td>4.72%</td>
<td>5.16%</td>
<td>5.05%</td>
</tr>
<tr>
<td><strong>DR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.i.d.</td>
<td>4.55%</td>
<td>5.18%</td>
<td>4.96%</td>
<td>5.05%</td>
</tr>
<tr>
<td>trim. fix.</td>
<td>4.41%</td>
<td>5.31%</td>
<td>6.07%</td>
<td>6.56%</td>
</tr>
<tr>
<td>trim. Crump (2009)</td>
<td>4.46%</td>
<td>5.61%</td>
<td>7.07%</td>
<td>8.37%</td>
</tr>
<tr>
<td>trim. Huber (2013)</td>
<td>4.50%</td>
<td>5.18%</td>
<td>4.96%</td>
<td>5.05%</td>
</tr>
<tr>
<td>m-out-of-n</td>
<td>4.13%</td>
<td>4.99%</td>
<td>4.85%</td>
<td>5.00%</td>
</tr>
</tbody>
</table>

Regular identification: $\gamma = 0.5$, correctly specified models for propensity score and mean function. The numbers in the table correspond to the share of Monte Carlo draws where $H_0: \theta = \theta_0$ was rejected against the two-sided alternative. Columns correspond to different sample sizes $n$. All simulations are based on 12000 Monte Carlo repetitions. For all methods the number of bootstrap iterations is set to $B_2 = 500$. In the $\hat{m}$-choice algorithm: $q = 0.75$, $S = 100$, $B_1 = 10000$ ($B_1 = 5000$ for $n = 10000$), $j = 0, 1, \ldots, 40$.

However, in case of regular identification the probability for occurrence of extreme propensity scores is very low. Therefore in our simulations the trimmed sample usually happens to cover about 98% of the original sample and the bias of the ATE estimates only affects the rejection rates in the range of 1% to 3%.

Table 5.2: Null Rejection Rates for Nominal 5% Level under Irregular Identification

<table>
<thead>
<tr>
<th></th>
<th>n=500</th>
<th>n=1000</th>
<th>n=5000</th>
<th>n = 10000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IPW</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.i.d.</td>
<td>17.59%</td>
<td>17.89%</td>
<td>16.15%</td>
<td>15.56%</td>
</tr>
<tr>
<td>trim. fix.</td>
<td>17.22%</td>
<td>32.63%</td>
<td>93.96%</td>
<td>99.90%</td>
</tr>
<tr>
<td>trim. Crump (2009)</td>
<td>17.52%</td>
<td>31.37%</td>
<td>94.04%</td>
<td>99.89%</td>
</tr>
<tr>
<td>trim. Huber (2013)</td>
<td>16.44%</td>
<td>17.11%</td>
<td>15.90%</td>
<td>15.47%</td>
</tr>
<tr>
<td>m-out-of-n</td>
<td>13.31%</td>
<td>13.12%</td>
<td>12.37%</td>
<td>11.93%</td>
</tr>
<tr>
<td><strong>DR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.i.d.</td>
<td>2.78%</td>
<td>3.23%</td>
<td>4.00%</td>
<td>3.29%</td>
</tr>
<tr>
<td>trim. fix.</td>
<td>17.80%</td>
<td>37.38%</td>
<td>96.93%</td>
<td>99.99%</td>
</tr>
<tr>
<td>trim. Crump (2009)</td>
<td>17.63%</td>
<td>37.62%</td>
<td>97.82%</td>
<td>99.99%</td>
</tr>
<tr>
<td>trim. Huber (2013)</td>
<td>2.92%</td>
<td>3.37%</td>
<td>3.95%</td>
<td>3.31%</td>
</tr>
<tr>
<td>m-out-of-n</td>
<td>2.34%</td>
<td>2.37%</td>
<td>2.97%</td>
<td>1.91%</td>
</tr>
</tbody>
</table>

Irregular identification: $\gamma = 1.5$, correctly specified models propensity score and mean function. The numbers in the table correspond to the share of Monte Carlo draws where $H_0: \theta = \theta_0$ was rejected against the two-sided alternative. Columns correspond to different sample sizes $n$. All simulations are based on 12000 Monte Carlo repetitions. For all methods the number of bootstrap iterations is set to $B_2 = 500$. In the $\hat{m}$-choice algorithm: $q = 0.75$, $S = 100$, $B_1 = 10000$ ($B_1 = 5000$ for $n = 10000$), $j = 0, 1, \ldots, 40$.

Table 5.2 reports empirical null rejection rates for the case of irregular identification $\gamma = 1.5$, i.e. many extreme propensity scores. For the IPW estimator, fixed and optimal trimming rules by Crump et al. (2009) have an asymptotic bias which results in implausible rejection rates, getting worse with increasing sample size up to 100%. Rejection rates for the nonparametric bootstrap are in the range of 16% to 18%. The $m$-out-of-$n$ bootstrap generates 4% - 5% smaller size distortions across all sample sizes. For the DR estimator,
the size distortions are not that pronounced. For this design, the i.i.d. bootstrap and trimming by Huber et al. (2013) produce under-rejection of 1%-2%. The $m$-out-of-$n$ bootstrap with improved $\hat{m}$-selection has under-rejections of about 2%-3%.

Table 5.3: Null Rejection Rates for Nominal 5% level under Misspecified Propensity Score Model

<table>
<thead>
<tr>
<th></th>
<th>n=500</th>
<th>n=1000</th>
<th>n=5000</th>
<th>n=10000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IPW</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.i.d.</td>
<td>5.18%</td>
<td>4.07%</td>
<td>14.45%</td>
<td>30.29%</td>
</tr>
<tr>
<td>trim. fix.</td>
<td>28.05%</td>
<td>51.37%</td>
<td>99.85%</td>
<td>100.00%</td>
</tr>
<tr>
<td>trim. Crump (2009)</td>
<td>33.60%</td>
<td>61.93%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>trim. Huber (2013)</td>
<td>5.08%</td>
<td>4.08%</td>
<td>14.45%</td>
<td>30.42%</td>
</tr>
<tr>
<td>$m$-out-of-$n$</td>
<td>4.68%</td>
<td>3.45%</td>
<td>12.00%</td>
<td>25.08%</td>
</tr>
<tr>
<td><strong>DR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.i.d.</td>
<td>2.97%</td>
<td>2.47%</td>
<td>6.43%</td>
<td>12.86%</td>
</tr>
<tr>
<td>trim. fix.</td>
<td>27.70%</td>
<td>51.58%</td>
<td>99.52%</td>
<td>100.00%</td>
</tr>
<tr>
<td>trim. Crump (2009)</td>
<td>31.30%</td>
<td>56.83%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>trim. Huber (2013)</td>
<td>3.27%</td>
<td>2.83%</td>
<td>6.83%</td>
<td>13.10%</td>
</tr>
<tr>
<td>$m$-out-of-$n$</td>
<td>1.32%</td>
<td>1.32%</td>
<td>2.25%</td>
<td>4.52%</td>
</tr>
</tbody>
</table>

Irregularity through misspecification. In the true DGP $\gamma = 0.9$, $X_i \overset{i.i.d.}{\sim} U(0,1)$, $p(X_i) = \gamma X_i$. Mean functions remain the same as in (5.1) and are correctly specified for estimation. Misspecified propensity score model: $p^{(m)}(X_i) = 1/(1 + \exp(-\hat{\gamma}_0 - \hat{\gamma}_1 \ln(X_i)))$. The numbers in the table correspond to the share of Monte Carlo draws where $H_0: \theta = \theta_0$ was rejected against the two-sided alternative. Columns correspond to different lengths of sample size $n$. All simulations are based on 6000 Monte Carlo repetitions. For all methods the number of bootstrap iterations is set to $B_2 = 500$. In the $m$-choice algorithm: $q = 0.75$, $S = 200$, $B_1 = 1000$, $j = 0, 1, \ldots, 40$.

Consider now the case of propensity score misspecification as discussed in Section 3.3. We use a regularly identified model with $\gamma = 0.9$ and standard uniform $X_i \overset{i.i.d.}{\sim} U(0,1)$, but propensity scores estimated using a logit model with $\ln(X_i)$ as regressor as in (3.7). Under a misspecified propensity score model the IPW estimator will be biased independently of the regularity of the problem. The DR estimator is robust against misspecification and produces consistent estimates of the treatment effect. As discussed in Section 3.3, misspecification of the propensity score model changes the identification of the problem to an irregularly identified treatment effect parameter with an alpha-stable instead of normal limiting distribution. Table 5.3 contains the null rejection rates for IPW and DR for the different sample sizes. For IPW, i.i.d. bootstrap, $m$-out-of-$n$ bootstrap, and relative weight trimming show implausible rejection rates up to 30% while all other trimming approaches yield rejection rates up to 100%. For DR, the over-rejection rates of the i.i.d. bootstrap and trimming approaches increase with the sample size $n$. Only the $m$-out-of-$n$ bootstrap combined with the bootstrap aggregated $\hat{m}$-selection method shows no over-rejections approaching the nominal 5% level for larger sample sizes. Thus, it provides the empirical researcher with some safeguard against the consequences of misspecification-induced irregularity on statistical inference for doubly robust estimation.

---

20 This is due to the fact that the quantiles of the test statistic for this particular design at 5% are close to the ones from the normal distribution, see Figure 5 in Supplementary Appendix B.3.4. It is easy to find designs for which the i.i.d. bootstrap produces rejections far from the nominal level.
6. Empirical Application: The Effect of Right Heart Catheterization on Survival Rates

In this section we re-analyze the causal impact of right heart catheterization (RHC) on survival rates of ill adult patients. The analysis is based on the observational study by Connors et al. (1996) that uses data from the study to understand prognoses and preferences for outcomes and risks of treatments, see Murphy and Cluff (1990). RHC is a diagnostic tool for monitoring cardiovascular activity such as right atrium pressure to supervise and guide therapy. The data set contains 2184 treatment and 3551 control observations. An observation is defined as treated if the patient was subject to RHC within 24 hours of admission. Health is evaluated using 30-day survival of the patients. The data set comprises over 72 covariates including a large set of medical information about the subjects such as blood pressure, heart rate, and various cancer diagnoses. Moreover, socio-economic information such as sex, education, and income are recorded as well, for more details consider Connors et al. (1996). As selection of initial care patients into treatment depends on observable, medically relevant information that is also related to survival chances, Connors et al. (1996) use a propensity score pair matching approach to control for potential selection bias. They find that RHC leads to a lower 30-day survival rate compared to no RHC. The data set has also been studied with propensity score based methods by Hirano and Imbens (2001), Crump et al. (2009), and Rothe (2017).

We estimate the average treatment effect using both inverse probability weighting and a doubly robust approach similar to Hirano and Imbens (2001) and Crump et al. (2009). The propensity score model uses the same regressors as Hirano and Imbens (2001) in a probit model. Figure 6.1 shows the distributions of the estimated propensity scores for treatment and control group. For both groups, the propensity scores cover almost the full support from zero to one indicating a violation of strong overlap. Moreover, for the control group the share of observations in the propensity score bins increases as the scores get closer to zero which is a first indicator for irregular identification with regularly varying propensity score density at zero.
Table 6.1: Average Treatment Effects and Confidence Intervals

<table>
<thead>
<tr>
<th>Method</th>
<th>Point Estimate</th>
<th>95% CI</th>
<th>99% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPWII</td>
<td>-0.0582</td>
<td>[-0.0950, -0.0213]</td>
<td>[-0.1066, -0.0097]</td>
<td>0.0020</td>
</tr>
<tr>
<td>Asymptotic Normality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetric Percentile Bootstrap</td>
<td></td>
<td>[-0.0951, -0.0213]</td>
<td>[-0.1061, -0.0102]</td>
<td>0.0012</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overlap Robust (m = 871)</td>
<td></td>
<td>[-0.1050, -0.0114]</td>
<td>[-0.1201, 0.0040]</td>
<td>0.0150</td>
</tr>
<tr>
<td>DR</td>
<td>-0.0609</td>
<td>[-0.0936, -0.0282]</td>
<td>[-0.1039, -0.0179]</td>
<td>0.0002</td>
</tr>
<tr>
<td>Asymptotic Normality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetric Percentile Bootstrap</td>
<td></td>
<td>[-0.0942, -0.0275]</td>
<td>[-0.1049, -0.0169]</td>
<td>0.0004</td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overlap Robust (m = 712)</td>
<td></td>
<td>[-0.0993, -0.0224]</td>
<td>[-0.1116, 0.0102]</td>
<td>0.0032</td>
</tr>
</tbody>
</table>

Results are based on $B_2 = 5000$ bootstrap/m-out-of-n iterations. The aggregated $m$ choice is based on $S = 250$ iterations. The propensity score and potential outcome models are reestimated within each bootstrap iteration.

Using our model, the smallest estimated propensity score is about 0.0004 which corresponds to an inverse weight of 2500 to 1.

Table 6.1 contains the point estimates, 95% and 99% confidence intervals, and p-values for the average treatment effect of RHC on 30-day survival using asymptotic normality with analytical standard errors, nonparametric bootstrap with symmetric percentiles, and the overlap robust confidence intervals. For the latter, the resampling sizes $m$ are chosen according to the optimal $m$-choice method with bootstrap aggregation as in Section 4. The suggested $m$-choices are 871 and 712 for inverse probability weighting and doubly robust estimation respectively which is about 12% to 15% of the full sample size of 5735.$^{21}$ Using the respective $m$-estimates from the alternative estimation approach does not yield qualitatively different results for the confidence bounds.

Our point estimates are similar or equivalent to the ones in Hirano and Imbens (2001): RHC decreases the expected 30-day survival probability by 6% for both inverse probability weighting and doubly robust estimation.$^{22}$ Without controlling for the presence of extreme propensity scores, the results are significant

---

$^{21}$Inspection of the simulated distribution of $T_{m,n}$ reveals some of the features that would be expected from the limiting law of $T_n$ under irregular identification, i.e. symmetry around $[-1,1]$ and small modes at $±1$ which argues in favor of using robust inference.

$^{22}$The results differ somewhat from Rothe (2017) as he uses a stratification approach and a slightly different model for the propensity score that does not fully exploit the information on missing body weight entries for some patients.
at the 1% level using both normal approximation and conventional percentile bootstrap. Depending on the estimator and significance level, the robust confidence intervals are about 17% to 29% wider. Our adaptive inference method suggests that the findings are robust with respect to the weaker overlap assumption at a 5% level. However, using inverse probability weighting, the treatment effect is no longer significant at a 1% level with an asymptotic p-value of 0.015.

7. Concluding Remarks

In this paper we provide a method for conducting asymptotically valid inference for treatment effect estimators without imposing restrictive support conditions on the propensity score, distributional assumptions on the data, and/or trimming observations. Using the key assumption of regular variation, we show that treatment effect parameter estimates converge in distribution to alpha-stable random variables. Thus, conventional inference methods based on asymptotic normality or the nonparametric bootstrap are invalid. We propose a modified version of the $m$-out-of-$n$ bootstrap that not only yields asymptotically correct coverage rates for the stable limiting domain but also covers the regularly identified case. Thus the researcher can apply it without knowledge of the regularity of his or her identification problem. In addition it provides some safeguard against the effects of propensity score misspecification on statistical inference. Promising directions for future research are extensions to other parameters and general missing data problems that rely on inverse probability weights and propensity score estimation in the high-dimensional or nonparametric case. Moreover, it would be interesting to see whether finite sample approximations for the test statistic in the sense of Rothe (2017) could be used under irregular identification based on expansions for the self-normalizing sum as in Logan et al. (1973) to further improve coverage in finite samples.
Appendix A

A.1 Proof of Proposition 3.1

Proof: Under Assumption 3.1 for the propensity scores we have

\[ \hat{\gamma} - \gamma_0 = \left( E[S_i(\gamma_0)S_i(\gamma_0)']^{-1} + o_p(1) \right) \frac{1}{n} \sum_{i=1}^{n} S_i(\gamma_0) \]

with

\[ S_i(\gamma) = \sum_d \frac{D_i(d)}{p_d(X_i, \gamma)} \frac{\partial p_d(X_i, \gamma)}{\partial \gamma}. \]

Let \( \mu_d = E[Y_i(d)] \) and \( \hat{\mu}_d \) be the corresponding IPW estimator. Expanding around the true propensity score yields

\[ \hat{\mu}_d - \mu_d = \left( \frac{1}{n} \sum_{i=1}^{n} D_i(d) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} D_i(d) \left( Y_i(d) - \mu_d \right) \]

\[ = \left( 1 + o_p(1) \right) \left( \frac{1}{n} \sum_{i=1}^{n} D_i(d) \right) \left( Y_i(d) - \mu_d \right) \]

\[ - \left( Y_i(d) - \mu_d \right) \left( D_i(d) \right) \frac{p_d(X_i, \gamma)}{p_d(X_i, \gamma_0)} \frac{\partial p_d(X_i, \gamma)}{\partial \gamma} \left( \gamma - \gamma_0 \right) \]

\[ = \left( 1 + o_p(1) \right) \left( \frac{1}{n} \sum_{i=1}^{n} D_i(d) \right) \left( Y_i(d) - \mu_d \right) \]

\[ - \frac{1}{n} \sum_{i=1}^{n} \left( Y_i(d) - \mu_d \right) \left( D_i(d) \right) \frac{p_d(X_i, \gamma)}{p_d(X_i, \gamma_0)} \left( E[S_i(\gamma_0)S_i(\gamma_0)']^{-1} + o_p(1) \right) \frac{1}{n} \sum_{i=1}^{n} S_i(\gamma_0) \]

with \( \hat{\gamma} \) being on the line-segment between \( \gamma_0 \) and \( \hat{\gamma} \). The second equality follows from continuity and the moment bound in Assumption 3.2, together with Newey and McFadden (1994), Lemma 4.3. By the second moment bound in Assumption 3.2 it follows equivalently that

\[ \hat{\mu}_d - \mu_d = \left( 1 + o_p(1) \right) \left( \frac{1}{n} \sum_{i=1}^{n} k_{id}^{IPW} \right) - \left( E[k_{id}^{IPW} S_i(\gamma_0)'] + o_p(1)E[S_i(\gamma_0)S_i(\gamma_0)']^{-1} + o_p(1)S_i(\gamma_0) \right) \]

with

\[ k_{id}^{IPW} = \frac{D_i(d)}{p_d(X_i, \gamma_0)} \left( Y_i(d) - \mu_d \right). \]

Thus under the regular identification according to Assumption 3.3 we have

\[ \sqrt{n}(\hat{\mu}_d - \mu_d) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} k_{id}^{IPW} - b_d S_i(\gamma_0) + o_p(1) \]

with \( b_d = E[k_{id}^{IPW} S_i']E[S_i S_i']^{-1} \) for both \( d = 0, 1 \). Using a standard CLT yields

\[ \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \kappa_1 + \kappa_0 - 2\omega) \]
with

\[ \kappa_d = E[(k_{IPW} | S_i - b_d S_i)^2] \]

\[ \omega = E[k_{IPW}^1 S_i^1 E[S_i^1 S_i^1]^{-1} E[S_i^1 k_{IPW}^1]]. \]

Now assume Assumption 3.3 to be violated, i.e. the second moments are singular, but instead there is regular variation of the truncated second moment plus tail balance for the components of the first two partial sums according to Assumption 3.4.

Let \( u \) be independent uniform on \((0, 1)\) and \( F \) be the probability measure with respect to all other random variables. Note that

\[
E \left[ \frac{D_i(d)(Y_i(d) - \mu_d)}{p_d(X_i)} \right] \begin{cases} 
\omega, & \text{if } 0 < \omega, \\
1, & \text{if } \omega = 0.
\end{cases}
\]

Thus, the conditions in Assumption 3.4 are equivalent to the ones in Feller (1968) Theorem XVII.5.2. The latter implies that both \( k_{IPW}^1 \) and \( k_{IPW}^0 \) are in the domain of attraction of an alpha-stable distribution with tail index being equal to \( \alpha_1 \) and \( \alpha_0 \) respectively. Thus, there are two sequences \( a_n \) and \( b_n \) with \( c_n = \min\{a_n, b_n\} = o(n^{1/2}) \)
such that
\[
c_n(\hat{\theta}_n - \theta_0) = \left( a_n \frac{1}{n} \sum_{i=1}^{n} k_{i1}^{IPW} \right) c_n - \left( b_n \frac{1}{n} \sum_{i=1}^{n} k_{i0}^{IPW} \right) c_n
- \left( b_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\gamma_0) - b_0 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\gamma_0) \right) \frac{a_n}{\sqrt{n}} + o_p(1)
\]
\[= S_{n,1} - S_{n,0} - S_{n,a} + o_p(1)
= S_{n,1} - S_{n,0} + O_p(1)O(\frac{c_n}{\sqrt{n}}) + o_p(1)
= S_{n,1} - S_{n,0} + o_p(1).
\]

If \(U_0(z) = o(U_1(z))\), \(a_n = o(b_n)\) thus
\[S_{n,1} = \left( a_n \frac{1}{n} \sum_{i=1}^{n} k_{i1}^{IPW} \right) \xrightarrow{d} S(\alpha_1, \beta_1, 0, c_1)
S_{n,0} = \left( b_n \frac{1}{n} \sum_{i=1}^{n} k_{i0}^{IPW} \right) \frac{a_n}{b_n} = O_p(1)O(a_n/b_n) = o_p(1)
\]
\[\Rightarrow c_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} S(\alpha_1, \beta_1, 0, c_1).
\]

Similarly if \(U_1(z) = o(U_0(z))\), \(b_n = o(a_n)\) thus
\[S_{n,1} = \left( a_n \frac{1}{n} \sum_{i=1}^{n} k_{i1}^{IPW} \right) \frac{b_n}{a_n} = O_p(1)O(b_n/a_n) = o_p(1)
S_{n,0} = \left( b_n \frac{1}{n} \sum_{i=1}^{n} k_{i0}^{IPW} \right) \xrightarrow{d} S(\alpha_0, \beta_0, 0, c_0)
\]
\[\Rightarrow c_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} S(\alpha_0, \beta_0, 0, c_0).
\]

However, if \(U_1(z) = O(U_0(z))\), \(a_n \sim b_n\), thus
\[S_{n,1} = \left( a_n \frac{1}{n} \sum_{i=1}^{n} k_{i1}^{IPW} \right) \xrightarrow{d} S(\alpha_1, \beta_1, 0, c_1)
S_{n,0} = \left( b_n \frac{1}{n} \sum_{i=1}^{n} k_{i0}^{IPW} \right) \xrightarrow{d} S(\alpha_0, \beta_0, 0, c_0)
\]

with \(\alpha_1 = \alpha_0 = \alpha\). Note that by construction of the independent samples, the two components are asymptotically independent and hence
\[c_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} S(\alpha, \beta, 0, c)
\]
with
\[ c = (|c_1|^\alpha + |c_0|^\alpha)^{1/\alpha}, \quad \beta = \frac{\beta_1 |c_1|^\alpha - \beta_0 |c_0|^\alpha}{|c_1|^\alpha + |c_0|^\alpha} \]
by the closure property under convolution for stable random variables with equivalent \( \alpha \). Thus, depending on the tail index of both components, the limiting behavior of the stabilizing transformation is only driven by the one with the slower decaying tails, i.e. the limiting distribution of the average treatment effect will then be determined by the tail properties of one of the potential outcome means except for the case of equivalent tail indices. The exact rate for the normalizing constant \( c_n \) follows directly from the rate of the truncated second moment, see Feller (1968), Section XVII.5, equation (5.23).

### A.2 Proof of Proposition 3.2

**Proof:** Under parametric estimators according to Assumptions 3.1 and 3.5 for the first-stage we have
\[
\hat{\gamma} - \gamma_0 = O_p(n^{-1/2})
\]
\[
\hat{\delta}_d - \delta_d = O_p(n^{-1/2})
\]
for both \( d = 0, 1 \). Expanding the differentiable mean function yields
\[
\mu_d(X_i, \hat{\delta}_d) = \mu_d(X_i, \delta_d) + \frac{\partial \mu_d(X_i, \hat{\delta}_d)}{\partial \delta_d} (\hat{\delta}_d - \delta_d)
\]
with \( \hat{\delta} \) being on the line segment between \( \hat{\delta}_d \) and \( \delta_d \). For the doubly robust estimator we obtain the following expansion
\[
\hat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^n \frac{D_i(1)(Y_i - \mu_1(X_i, \hat{\delta}_1))}{p_1(X_i, \hat{\gamma})} - \frac{1}{n} \sum_{i=1}^n \frac{D_i(0)(Y_i - \mu_0(X_i, \hat{\delta}_0))}{p_0(X_i, \hat{\gamma})} + \frac{1}{n} \sum_{i=1}^n \mu_1(X_i, \hat{\delta}_1) - \mu_0(X_i, \hat{\delta}_0) - \theta_0
\]
\[
= \frac{1}{n} \sum_{i=1}^n \frac{D_i(1)(Y_i - \mu_1(X_i, \delta_1))}{p_1(X_i, \gamma_0)} - \frac{1}{n} \sum_{i=1}^n \frac{D_i(0)(Y_i - \mu_0(X_i, \delta_0))}{p_0(X_i, \gamma_0)}
\]
\[
- \frac{1}{n} \sum_{i=1}^n \frac{D_i(1)(Y_i - \mu_1(X_i, \delta_1) \cdot \frac{\partial p_1(X_i, \hat{\gamma})}{\partial \gamma}}{p_1(X_i, \hat{\gamma})^2} (\hat{\gamma} - \gamma_0) + \frac{1}{n} \sum_{i=1}^n \frac{D_i(0)(Y_i - \mu_0(X_i, \delta_0) \cdot \frac{\partial p_0(X_i, \hat{\gamma})}{\partial \gamma}}{p_0(X_i, \hat{\gamma})^2} (\hat{\gamma} - \gamma_0)
\]
\[
+ \frac{1}{n} \sum_{i=1}^n \frac{D_i(1)}{p_1(X_i, \gamma_0)} \frac{\partial p_1(X_i, \hat{\gamma})}{\partial \gamma} (\hat{\gamma} - \gamma_0) \frac{\partial \mu_1(X_i, \hat{\delta}_1)}{\partial \delta_1} (\delta_1 - \delta_1)
\]
\[
- \frac{1}{n} \sum_{i=1}^n \frac{D_i(0)}{p_0(X_i, \gamma_0)} \frac{\partial p_0(X_i, \hat{\gamma})}{\partial \gamma} (\hat{\gamma} - \gamma_0) \frac{\partial \mu_0(X_i, \hat{\delta}_0)}{\partial \delta_0} (\delta_0 - \delta_0)
\]
\[
+ \frac{1}{n} \sum_{i=1}^n \left( 1 - \frac{D_i(1)}{p_1(X_i, \gamma_0)} \right) \frac{\partial \mu_1(X_i, \hat{\delta}_1)}{\partial \delta_1} (\delta_1 - \delta_1) - \frac{1}{n} \sum_{i=1}^n \left( 1 - \frac{D_i(0)}{p_0(X_i, \gamma_0)} \right) \frac{\partial \mu_0(X_i, \hat{\delta}_0)}{\partial \delta_0} (\delta_0 - \delta_0)
\]
\[
+ \frac{1}{n} \sum_{i=1}^n \mu_1(X_i, \hat{\delta}_1) - \mu_0(X_i, \hat{\delta}_0) - \theta_0.
\]
Note that by the law of iterated expectations and conditional independence

\[ E \left[ \frac{D_i(d)Y_i - \mu_d(X_i, \delta_d)}{p_d(X_i, \gamma_0)^2} \frac{\partial p_d(X_i, \gamma_0)}{\partial \gamma} \right] = E \left[ \frac{E[Y_i(d)X_i] - \mu_d(X_i, \delta_d)}{p_d(X_i, \gamma_0)} \frac{\partial p_d(X_i, \gamma_0)}{\partial \gamma} \right] = 0 \]

and

\[ E \left[ 1 - \frac{D_i(d)}{p_d(X_i, \gamma_0)} \frac{\partial p_d(X_i, \delta_d)}{\partial \delta_d} \right] = E \left[ 1 - \frac{E[D_i(d)|X_i]}{p_d(X_i, \gamma_0)} \frac{\partial p_d(X_i, \delta_d)}{\partial \delta_d} \right] = 0. \]

Since \( \hat{\gamma} - \gamma_0 = O_p(n^{-1/2}) \) and \( \hat{\delta}_d - \delta_d = O_p(n^{-1/2}) \), Newey and McFadden (1994), Lemma 4.3 in conjunction with Assumption 3.6 implies that

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{D_i(d)(Y_i - \mu_d(X_i, \delta_d))}{p_d(X_i, \gamma_0)^2} \frac{\partial p_d(X_i, \hat{\gamma})}{\partial \gamma} (\hat{\gamma} - \gamma_0) = o_p(n^{-1/2})
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \left( 1 - \frac{D_i(d)}{p_d(X_i, \gamma_0)} \right) \frac{\partial p_d(X_i, \hat{\delta}_d)}{\partial \delta_d} (\hat{\delta}_d - \delta_d) = o_p(n^{-1/2})
\]

while for the cross-terms we have

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{D_i(d)}{p_d(X_i, \gamma_0)^2} \frac{\partial p_d(X_i, \hat{\gamma})}{\partial \gamma} (\hat{\gamma} - \gamma_0) \frac{\partial p_d(X_i, \hat{\delta}_d)}{\partial \delta_d} (\hat{\delta}_d - \delta_d) = O_p(n^{-1}).
\]

Thus the expansion simplifies to

\[
\hat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(1)(Y_i - \mu_1(X_i, \delta_1))}{p_1(X_i, \gamma_0)} - \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(0)(Y_i - \mu_0(X_i, \delta_0))}{p_0(X_i, \gamma_0)} + \frac{1}{n} \sum_{i=1}^{n} \mu_1(X_i, \delta_1) - \mu_0(X_i, \delta_0) - \theta_0 + o_p(n^{-1/2}).
\]

Note that the cross-terms have zero correlation, i.e.

\[
E \left[ \frac{D_i(1)(Y_i - \mu_1(X_i, \delta_1))}{p_1(X_i, \gamma_0)} \frac{D_i(0)(Y_i - \mu_0(X_i, \delta_0))}{p_0(X_i, \gamma_0)} \right] = 0
\]

\[
E \left[ \frac{D_i(d)(Y_i - \mu_d(X_i, \delta_d))}{p_d(X_i, \gamma_0)} \right] \frac{(\mu_1(X_i, \delta_1) - \mu_0(X_i, \delta_0) - \theta_0)^2}{p_0(X_i, \gamma_0)} = 0
\]

by the law of iterated expectations. Thus under regular identification according to Assumption 3.7, the central limit theorem implies that

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N} \left( 0, \sum_{d} E \left[ (Y_i(d) - \mu_d(X_i, \delta_d))^2 \right] + E \left[ (\mu_1(X_i, \delta_1) - \mu_0(X_i, \delta_0) - \theta_0)^2 \right] \right).
\]
Under irregular identification \( D_i(d)(Y_i - \mu_d(X_i, \delta_d))/p_d(X_i, \gamma_0) \) are of regular variation with potentially different tail indices. By equivalent arguments as in Section A.1, let \( c_n = \min\{a_n, b_n\} = o(n^{1/2}) \). We have

\[
e_n(\hat{\theta}_n - \theta_0) = c_n \frac{1}{n} \sum_{i=1}^{n} D_i(1)(Y_i - \mu_1(X_i, \delta_1)) p_1(X_i, \gamma_0) - c_n \frac{1}{n} \sum_{i=1}^{n} D_i(0)(Y_i - \mu_0(X_i, \delta_0)) p_0(X_i, \gamma_0)
+ O_p(c_n n^{-1/2}) + o_p(1)
\]

\[
e_n \frac{1}{n} \sum_{i=1}^{n} D_i(1)(Y_i - \mu_1(X_i, \delta_1)) p_1(X_i, \gamma_0) - c_n \frac{1}{n} \sum_{i=1}^{n} D_i(0)(Y_i - \mu_0(X_i, \delta_0)) p_0(X_i, \gamma_0)
+ O_p(1)
\]

\[
\Rightarrow S(\alpha, \beta, 0, c)
\]

according to Feller (1968). Note again that depending on the tail index of the two weighted components, one will prevail and determine the convergence rate of the stabilizing transformation if the tails are different.

A.3 Proof of Theorem 4.1

(Part I, IPW). For the following let

\[
k_i - bS_i \equiv k_{1IPW}^i - b_1 S_i - k_{0IPW}^i + b_0 S_i
\]

\[
T_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (k_i - bS_i)^2}}
\]

i) follows from the approximation of Proposition 3.1 using a CLT for independent data under finite variance and Slutsky’s theorem. For ii) note that

\[
\sqrt{n}(\hat{\theta}_n - \theta_0)
\sqrt{\frac{1}{n} \sum_{i=1}^{n} (k_i - bS_i)^2}
\frac{\sum_{i=1}^{n} (k_i - bS_i)}{\sqrt{\sum_{i=1}^{n} (k_i - bS_i)^2}} + o_p(1).
\]

From the proof of Proposition 3.1 one can infer that there exists a sequence \( c_n = o(n^{1/2}) \) such that

\[
\frac{c_n}{n} \sqrt{\frac{1}{n} \sum_{i=1}^{n} bS_i} = O_p(c_n n^{-1/2})
\]

\[
\frac{c_n}{n} \sum_{i=1}^{n} k_i = O_p(1)
\]

and thus

\[
\sum_{i=1}^{n} bS_i = o_p(|\sum_{i=1}^{n} k_i|).
\]
For the denominator we have
\[ \sum_{i=1}^{n} (k_i - bS_i)^2 = \sum_{i=1}^{n} k_i^2 - 2 \sum_{i=1}^{n} k_ibS_i + b \sum_{i=1}^{n} S_iS'_i. \]

Let \( \lambda_1(A) \) be the largest eigenvalue of a matrix \( A \). Note that
\[ b \sum_{i=1}^{n} S_iS'_i = \text{tr}(b \sum_{i=1}^{n} S_iS'_i) \]
\[ = \text{tr}(b'b \sum_{i=1}^{n} S_iS'_i) \]
\[ \leq \lambda_1(b'b) \sum_{j} n \sum_{i=1}^{n} s_{ij}^2 \]

with \( s_{ij}^2 \) being the \( j \)-th component of the score \( S_i \) at the true parameter. By Assumption 3.2, \( \lambda_1(b'b) \) is bounded from above. By Assumption 3.1, the squared scores are (regular) means, i.e. a weak law of large numbers applies. Note that \( k_i^2 \) are distributed on the positive half-axes and by Assumption 3.4, their truncated variance is also regularly varying with tail index \( \tilde{\alpha} = \alpha/2 \). As \( \tilde{\alpha} < 1 \), it follows from Feller (1968), Theorem XVII.5.2 and Theorem XVII.5.3, that (without centering) the means are in the domain of attraction of a stable random variable that does not possess a mean, i.e. for some \( \zeta_n = o(1) \) with \( \zeta_n n^{1/2} \to \infty \)
\[ \frac{\zeta_n}{n} \sum_{i=1}^{n} k_i^2 \overset{d}{\to} S(\alpha/2, 1, \tilde{\mu}, \tilde{c}). \]
This implies that \( \frac{1}{n} \sum_{i=1}^{n} k_i^2 = O_p(\zeta_n^{-1}) \). Moreover, by the WLLN \( \frac{1}{n} \sum_{i=1}^{n} s_{ij}^2 = O_p(1) \) and thus \( \sum_{i=1}^{n} s_{ij}^2 = \Theta_p(\sum_{i=1}^{n} k_i^2) \) which implies that
\[ b \sum_{i=1}^{n} S_iS'_i = \Theta_p(\sum_{i=1}^{n} k_i^2). \]

Using the intermediate results yields the order for the cross-product by the Cauchy-Schwarz inequality:
\[ \sum_{i=1}^{n} k_ibS_i \leq \sqrt{\sum_{i=1}^{n} k_i^2b \sum_{i=1}^{n} S_iS'_i} \]
\[ = \Theta_p(\sqrt{(\sum_{i=1}^{n} k_i^2)^2}) \]
\[ = \Theta_p(\sum_{i=1}^{n} k_i^2). \]
Plugging in the rates into the normalized sum

\[
\frac{\sum_{i=1}^{n}(k_i - bS_i)}{\sqrt{\sum_{i=1}^{n}(k_i - bS_i)^2}} = \frac{\sum_{i=1}^{n} k_i}{\sqrt{\sum_{i=1}^{n} k_i^2}} + o_p(1)
\]

\[
= \frac{n^{-1/\alpha} \sum_{i=1}^{n} k_i}{\sqrt{n^{-2/\alpha} \sum_{i=1}^{n} k_i^2}} + o_p(1)
\]

\[
\equiv \frac{U_n}{V_n} + o_p(1).
\]

By Assumption 3.4 and Feller (1968), Theorem XVII.5.2 it follows that \( k_i \) are in the domain of a stable distribution with tail parameter \( \alpha > 1 \). By Logan et al. (1973), for an appropriate sequence \( c_n \), there is joint convergence in distribution of \( (c_n U_n, c_n^2 V_n^2) \). The smoothness and symmetry properties in Theorem 4.1 are also shown in Logan et al. (1973).

**Part II: DR.** For the following let

\[
k_i \equiv D_i(1)(Y_i - \mu_1(X_i, \delta_1)) p_1(X_i, \gamma_0) - D_i(0)(Y_i - \mu_0(X_i, \delta_0)) p_0(X_i, \gamma_0)
\]

\[
+ \mu_1(X_i, \delta_1) - \mu_0(X_i, \delta_0) - \theta_0
\]

\[
\equiv k_{i1} - k_{i0} + a(X_i)
\]

(A.3)

\[
T_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} k_{i1}^2 - k_{i0}^2 + a(X_i)^2}}.
\]

(A.4)

\( i \) follows from the approximation of the proof in Section A.2 using a central limit theorem for independent data under finite variance and Slutsky’s theorem. For \( ii \) note that

\[
\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} k_{i1}^2 + k_{i0}^2 + a(X_i)^2}} = \frac{\sum_{i=1}^{n} k_{i1} - k_{i0} + a(X_i)}{\sqrt{\sum_{i=1}^{n} k_{i1}^2 + k_{i0}^2 + a(X_i)^2}} + o_p(1).
\]

Without loss of generality assume that \( a_n = o(b_n) \). Note that Assumption 3.8, iii) implies that \( E[\mu_d(X_i, \delta_d)^2] < \infty \) by the law of total variance. Thus, by regularity of the “regression component” and regular variation for the “weighted component” we have

\[
\sum_{i=1}^{n} a(X_i) = o_p(\sum_{i=1}^{n} k_{i1})
\]

\[
\sum_{i=1}^{n} k_{i0} = o_p(\sum_{i=1}^{n} k_{i1})
\]
and equivalently

\[ \sum_{i=1}^{n} a(X_i)^2 = o_p(\sum_{i=1}^{n} k_{i1}^2) \]

\[ \sum_{i=1}^{n} k_{i0}^2 = o_p(\sum_{i=1}^{n} k_{i1}^2). \]

Thus

\[ \frac{\sum_{i=1}^{n} k_{i1} - k_{i0} + a(X_i)}{\sqrt{\sum_{i=1}^{n} k_{i1}^2 + k_{i0}^2 + a(X_i)^2}} = \frac{\sum_{i=1}^{n} k_{i1}}{\sqrt{\sum_{i=1}^{n} k_{i1}^2}} + o_p(1) \]

\[ = \frac{n^{-1/2} \sum_{i=1}^{n} k_{i1}}{\sqrt{n^{-2/\alpha} \sum_{i=1}^{n} k_{i1}^2}} + o_p(1) \]

\[ \equiv \frac{U_n}{V_n} + o_p(1), \]

where, for appropriately \( c_n, (c_n U_n, c_n^2 V_n^2) \) converge jointly by the arguments in Logan et al. (1973). The same holds true for the reverse case with flipped indices and \( b_n = o(a_n) \). For \( a_n \sim b_n \) note that by construction

\[ k_{i1} k_{i0} = \frac{D_1(1)(Y_i - mu_1(X_i, \delta_1))}{p_1(X_i, \gamma_0)} \frac{D_1(0)(Y_i - mu_0(X_i, \delta_0))}{p_0(X_i, \gamma_0)} = 0 \]

and thus

\[ \frac{\sum_{i=1}^{n} k_{i1} - k_{i0} + a(X_i)}{\sqrt{\sum_{i=1}^{n} k_{i1}^2 + k_{i0}^2 + a(X_i)^2}} = \frac{\sum_{i=1}^{n} (k_{i1} - k_{i0})}{\sqrt{\sum_{i=1}^{n} (k_{i1} - k_{i0})^2}} + o_p(1), \]

as the “regression components” are still dominated by the regularly varying tails of the “weighting components”. Thus, under irregular identification the variance-like denominator in (A.4) serves as a self-normalizing sum. Since the leading term is now a proper self-normalizing sum with zero mean in the numerator, the Theorem follows similarly as for the IPW estimator from Logan et al. (1973).

Acknowledgments

We would like to thank the Editor, the Associate Editor, two anonymous referees, Patrik Guggenberger, Alastair Hall, Winfried Pohlmeier, Christoph Rothe, the members of the 7th Konstanz-St. Gallen CSS Workshop 2017, and the participants of the ESEM 2018 for their time, comments, and discussions that helped to greatly improve the paper. The authors gratefully acknowledge the financial support from the German Research Foundation through Project 321011636 (Ekaterina Kazak) and Project 219805061 (Phillip Heiler) and by the Graduate School of Decision Sciences at the University of Konstanz. Computations were supported by the Karlsruhe bwHPC-Cluster. All remaining errors are ours.
References


List of Figures

2.1 Beta Distribution: Examples .................................................. 7
2.2 Distribution of Simulated \( \hat{\theta}_n \) ....................................... 9
3.1 True and Misspecified Propensity Score Distributions .................. 18
5.1 DGP: Propensity Score Densities and Conditional Treatment Effect .. 23
6.1 Propensity Score Distributions ................................................ 27
B.1 Supplementary Material for Section 3.1 and 3.2

Supplementary Material to Heiler and Kazak (2020)

Contents

B.1 Supplementary Material for Section 3.1 and 3.2 . . . . . . . . . . . . . . . . . . . . . 1
  B.1.1 Sufficient Conditions for Assumptions 3.2 and 3.6 . . . . . . . . . . . . . . . . . . 1
  B.1.2 Notes on Tail Balance Assumptions 3.4 ii) and 3.8 ii) . . . . . . . . . . . . . . . 3
  B.1.3 Doubly Robust Estimation under Additive Separability . . . . . . . . . . . . . . 5
  B.1.4 Beta Distribution Calculus . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
B.2 Supplementary Material for Section 3.3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
  B.2.1 Index of Regular Variation for the Misspecified Propensity Scores . . . . . . . . . . . 8
  B.2.2 Notes on Tail Index Estimation and Model Selection . . . . . . . . . . . . . . . . . 9
B.3 Supplementary Material for Sections 4 and 5 . . . . . . . . . . . . . . . . . . . . . . . . . . 11
  B.3.1 Asymptotic Variance and Self-normalizing Sums . . . . . . . . . . . . . . . . . . 11
  B.3.2 Power Properties of m-out-of-n Bootstrap . . . . . . . . . . . . . . . . . . . . . . . 11
  B.3.3 m-choice Alternatives . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
  B.3.4 Monte Carlo Distribution of $T_n$ for Doubly Robust Estimation . . . . . . . . . . . 19
  B.3.5 Monte Carlo Distribution of $\hat{m}$ . . . . . . . . . . . . . . . . . . . . . . . . . . 20

B.1 Supplementary Material for Section 3.1 and 3.2

B.1.1 Sufficient Conditions for Assumptions 3.2 and 3.6

In the following we provide sufficient conditions for the critical moment assumptions in the case of popular parametric specifications of the propensity score model. The moment Assumptions 3.2 and 3.6 of inverse probability weighted treatment indicator times score might seem restrictive on first sight as the likelihood scores contain inverse probability weighted terms as well. However, as propensity scores are usually cumulative distribution functions, their derivatives are densities and thus will compensate these inverse weights and lead to rather simple moment bounds. Consider two popular single index models:

1. Logistic regression:

$$p_d(X_i, \gamma) = \frac{1}{1 + \exp(-X_i'\gamma)}$$

2. Probit regression:

$$p_d(X_i, \gamma) = \Phi(X_i'\gamma)$$
with $\Phi(\cdot)$ denoting the cumulative distribution function of the standard normal distribution. Recall that the corresponding score functions at some $\gamma \in \Gamma$ are given by

$$S_i(\gamma) = \sum_d \frac{D_i(d)}{p_d(X_i, \gamma)} \frac{\partial p_d(X_i, \gamma)}{\partial \gamma}.$$ 

Note that for the logit for any $x$:

$$\left| \frac{1}{p_d(x, \gamma)} \frac{\partial p_d(x, \gamma)}{\partial \gamma} \right| = |(1 - p_d(x, \gamma))x| < |x|.$$ 

Thus Assumption 3.2, ii) is implied by

$$E \left[ \sup_{\gamma \in \mathcal{N}_0} \left| \frac{D_i(d)}{p_d(X_i, \gamma)} S_i(\gamma) \right| \right] \leq E \left[ \sup_{\gamma \in \mathcal{N}_0} \left| \frac{p_d(X_i, \gamma_0)}{p_d(X_i, \gamma)} X_i \right| \right] < \infty.$$ 

For the probit, note that using the asymptotic properties of the inverse Mills ratio implies a linear bound for the score component

$$\left| \frac{1}{p_d(x, \gamma)} \frac{\partial p_d(x, \gamma)}{\partial \gamma} \right| = \left| \frac{x \phi(x' \gamma)}{\Phi(x' \gamma)} \right| \leq C |x' \gamma|$$

for some $C > 0$, and $x' \gamma$ small as asymptotically the inverse mills ratio is linear for $x' \gamma \to -\infty$. This is without loss of generality as large values of the single index would always yield large propensities away from zero and thus a bounded expectation by construction. Thus Assumption 3.2, ii) is implied by

$$E \left[ \sup_{\gamma \in \mathcal{N}_0} \left| \frac{D_i(d)}{p_d(X_i, \gamma)} S_i(\gamma) \right| \right] \leq C E \left[ \sup_{\gamma \in \mathcal{N}_0} \left| \gamma \right| \left| \frac{p_d(X_i, \gamma_0)}{p_d(X_i, \gamma)} X_i X'_i \right| \right] < \infty.$$ 

As the parameter lies in a compact finite-dimensional space, this is equivalent to

$$E \left[ \sup_{\gamma \in \mathcal{N}_0} \left| \frac{p_d(X_i, \gamma_0)}{p_d(X_i, \gamma)} X_i X'_i \right| \right] < \infty.$$ 

This implies that the “confounder-weighted” ratios of true propensity score over propensity score model around the true parameter are not permitted to diverge too fast. Thus, for both probit and logit, the moment conditions on the interaction between score and weighted outcome boil down to simple moment conditions on the covariates in conjunction with the propensity score model.

In the following we consider further simplifications also covering Assumption 3.2, i) and Assumption 3.6, i). For simplicity, assume again a logistic model and independent regressors $x_1, \ldots, x_K$ with marginals $f_k(\cdot)$.
for \( k = 1, \ldots, K \). We obtain the following bound for the single index model:

\[
E \left[ \sup_{\gamma \in \mathbb{N}} \left| \frac{p_d(X_i, \gamma)}{p_d(X_i, \gamma_0)} \right| \right] = E \left[ \sup_{\gamma \in \mathbb{N}} \left| \frac{1 + \exp(-X_i'\gamma)}{1 + \exp(-X_i'\gamma_0)} \right| \right] \\
\leq 1 + E \left[ \sup_{\gamma \in \mathbb{N}} \left| \exp(-X_i'(\gamma - \gamma_0)) \right| \right] \\
= 1 + \prod_k \int_{\gamma_k \in \mathbb{N}} \sup f_k(x) \exp(x(\gamma_k - \gamma_{0,k})) dx.
\]

Without loss of generality, we consider only negative deviations, i.e. \((\gamma_k - \gamma_{0,k}) < 0\). Then the moment is bounded if there exists an \( \varepsilon > 0 \) such that for all \( k \)

\[
\lim_{x \to \infty} f_k(x) \exp(x\varepsilon) = 0.
\]

Thus we effectively require an exponential moment bound on the regressors rescaled by an arbitrary constant. Note that this does not contradict irregular propensity score distributions. For example, if \( x \) is logistically distributed with scaling constant \( \sigma > 0 \) the condition simplifies to

\[
\lim_{x \to \infty} \exp(-x\sigma) \exp(x\varepsilon) = \lim_{x \to \infty} \exp(-x(\sigma - \varepsilon)) = 0
\]

which always holds true by picking e.g. \( \varepsilon = \sigma/2 \). As the tails of the marginal regressor distributions have to vanish at least an exponential rate, the condition automatically implies a bound for the same term multiplied by polynomials of \( X_i \) which can be used to simplify the conditions derived above for Assumptions 3.4 ii) and 3.8, ii). For the probit model or single index models, the condition is more difficult to evaluate analytically. Using different numerical simulation designs we were unable to find examples that violate the moment assumption and thus it does not seem to severely restrict the range of admissible data-generating processes.

**B.1.2 Notes on Tail Balance Assumptions 3.4 ii) and 3.8 ii)**

The tail balance conditions 3.4(ii) and 3.8(ii) effectively restrict the distributions of the potential outcomes conditional on the propensity score. However, it is only concerned with how left and right tail of the conditional distribution of the potential outcome behave relative to each other as the propensity scores go to zero, i.e. at the extremes the occurrence of positive and negative deviations from the mean need to be in a constant (but not necessarily equal) ratio at the limit. Let \( E_d(\cdot) = E[\cdot | D_i(d) = 1] \) and
Consider the probability of an extreme event for the weighted outcome

\[ P(D_i(d)(Y_i(d) - \mu_d)/p_d(X_i)) > z) \]

\[ = E_d[p(1 - \mu_d > zp_d(X_i))|D_i(d = 1)P(D_i(d = 1) \]

\[ = E_d[E[p(1 - \mu_d > zp_d(X_i))|D_i = 1, p_d(X_i)]P(D_i(d = 1) \]

\[ = E_d[F_{y_d - \mu_d|p_d}(zp_d(X_i))]P(D_i(d = 1) \]

where \( F_{y_d - \mu_d|p_d}(z) \) denotes the distribution of the demeaned potential outcome conditional on the propensity score. The last step follows from independence of the potential outcome conditional on the propensity score, see Rosenbaum and Rubin (1983). Therefore the first tail balance condition simplifies to

\[
\lim_{z \to \infty} \frac{1 - E_d[F_{y_d - \mu_d|p_d}(zp_d(X_i))]}{1 - E_d[F_{y_d - \mu_d|p_d}(zp_d(X_i))] + E_d[F_{y_d - \mu_d|p_d}(-zp_d(X_i))]} = q_d
\]

and similarly for the second condition. Existence of this limit does not require an equal or similar likelihood of positive or negative values of deviations of the potential outcome from its mean. It is sufficient for tail balance if the conditional distributions converge for positive and negative deviations of \( y_d \) as the propensity scores \( p_d \) go to zero from above. Another sufficient condition is a symmetric conditional distribution, i.e. \( F_{y_d - \mu_d|p_d}(-x) = 1 - F_{y_d - \mu_d|p_d}(x) \) which implies \( q_d = r_d = 0.5 \). These parameters yield a limiting stable law for the treatment effect estimator that is symmetric as well, i.e. \( \beta = 0 \). Moreover, as shown in the section on additive separability, for the additive separable model with bounded conditional heteroskedasticity, tail balance always holds.

Violation of tail-balance would occur if as \( p_d \downarrow 0 \), the tails of the conditional distribution of \( y_d \) would oscillate between different patterns of decays for either positive or negative values while for the complementary case it does not. Or similarly if both oscillate simultaneously between differently varying tails. Consider the following example: For \( (y_d - \mu_d)/p_d \) given \( d = 1 \) assume existence of a density function \( f(\cdot) \) (up to a normalizing constant). Let the tails for negative values have polynomial decay, i.e.

\[ f(x) = |x|^{-a_1} \]

for any \( x < 0 \) with \( a_1 \in (0, 1) \). Now let the right side tail oscillate between two power laws along the support:

\[ f(x) = c_k(x) + \begin{cases} 
(1 + x)^{-a_1} & \text{if } k(x) \text{ odd} \\
(1 + x)^{-a_2} & \text{if } k(x) \text{ even} 
\end{cases} \]

for \( x \geq 0 \) with \( a_2 \in (0, 1) \), \( k(x) = \lfloor x \rfloor \) and \( c_k = c(k - 1) + (-1)^k(k^{-a_1} + k^{-a_2}) \) with \( c_1 = 0 \). This right tail of
the density function is continuous with varying tail decays between integers. The tails are balanced iff

$$\lim_{z \to \infty} \frac{1 - P_d((Y_i(d) - \mu_d)/p_d(X_i)) < z)}{1 - P_d((Y_i(d) - \mu_d)/p_d(X_i) < z) + P_d((Y_i(d) - \mu_d)/p_d(X_i) < -z)} = \lim_{z \to \infty} \frac{f(z)}{f(-z)}$$

exists. However, consider now the subsequences along only even or only odd $z$. We have that for odd $z$

$$\lim_{z \to \infty} \frac{f(z)}{f(-z)} = \lim_{z \to \infty} \left( \frac{z}{1 + z} \right)^{a_1 - a_2} = \begin{cases} \infty & \text{if } a_1 > a_2 \Rightarrow \beta = 1 \\ 1 & \text{if } a_1 = a_2 \Rightarrow \beta = 0 \\ 0 & \text{if } a_1 < a_2 \Rightarrow \beta = -1 \end{cases}$$

while along sequences of even $z$

$$\lim_{z \to \infty} \frac{f(z)}{f(-z)} = 1$$

which implies a skewness parameter $\beta = 0$ as the tail decay is 'symmetric' in its convergence. As the two subsequences have different limits if $a_1 \neq a_2$, the full sequence does not converge. While counterexamples like these are technically possible to construct, we do not consider them as empirically relevant for applied problems of causal inference.

### B.1.3 Doubly Robust Estimation under Additive Separability

The conditions in Section 3.2 can be simplified and become intuitively more appealing when applied to doubly robust estimation in an additive separable model. Moreover, it allows for an explicit characterization of the skewness parameter of the limiting distribution. Consider the following assumption:

**Assumption B.1.** For both $d = 0, 1$, the outcomes are generated by a location scale model

$$Y_i(d) = \mu_d(X_i, \delta_d) + \sigma_a(X_i) \varepsilon_i(d)$$

with

i) $E[\mu_d(X_i, \delta_d)^2] < \infty$ and $E[\varepsilon_i(d)^2] < \infty$,

ii) $\sigma \leq \sigma_d(X_i) \leq \overline{\sigma}$ with probability one for $\sigma, \overline{\sigma} \in \mathbb{R}^+$ and continuous density $f_\sigma(\sigma)$,

iii) $\varepsilon_i(d)$ and $X_i$ are independent,

iv) $\lim_{t \to 0^+} \frac{f_{\sigma_d}(zt)}{\mu_d(t)} = h(z) \sim z^{a_d - 1}$ exists on a dense set and is finite for some $z$.

Assumption B.1 puts regularity conditions on the moments of the mean function and the independent error and allows for bounded conditional heteroskedasticity. Note that compared to Assumption 3.8, the additional structure on the potential outcome process allows to substantially weaken the tail conditions for doubly robust estimation in the case of irregular identification. In particular, iv) only requires regular variation of the propensity score density at the boundaries instead of the weighted conditional mean error. Thus, in contrast to IPW, any dependence between outcome model and propensity scores through confounding variables in accordance with simple moment condition i) is admitted. Put differently, the DR approach under
heterogeneous treatment effects is comparable to the IPW in terms of the tail index under constant potential outcomes. Moreover, the tail balance condition can be omitted completely and relative tail probabilities will only be determined by the relative tails of the independent component \( \varepsilon_i(d) \). The following proposition provides the characterization of the limiting behavior of the doubly robust estimator under the additive separable model:

**Proposition B.1.** Let \( \hat{\theta}_n \) be chosen according to (3.2). If Assumptions 3.1, 3.6, 3.5, and either 3.7 or B.1 hold with \( \alpha_d < \alpha_{d'} \) for some \( d \in \{0, 1\} \) hold, then there exists a slowly varying function \( l(\cdot) \), such that

\[
[n^{\alpha-1}/l(n)]^{1/2} (\hat{\theta}_n - \theta_0) \xrightarrow{d} S(\alpha, \beta, 0, c),
\]

where \( S(\alpha, \beta, 0, c) \) denotes a random variable with an alpha-stable distribution with tail index \( \alpha = \alpha_d \in (1, 2) \), location parameter 0, scale parameter \( c > 0 \), and skewness parameter \( \beta \) given by

\[
\beta = \frac{1 - E[\varepsilon_{i-}^\alpha]/E[\varepsilon_{i+}^\alpha]}{1 + E[\varepsilon_{i-}^\alpha]/E[\varepsilon_{i+}^\alpha]}
\]

with \( \varepsilon_{-} = \min(\varepsilon, 0) \) and \( \varepsilon_{+} = \max(\varepsilon, 0) \) and \( \varepsilon_i = \varepsilon_i(d') \)

The proposition reveals that for the doubly robust estimator, the skewness parameter of the limiting distribution is determined by the ratio between negative and positive part \( \alpha \)-th moment of the independent component \( \varepsilon_i(d) \). Again, this is qualitatively different from inverse probability weighting estimators for which the tail balance condition determines the effective skewness parameter. Furthermore, the tail index is potentially affected by the mean function if the latter diverges or converges to zero as the propensity score approaches zero. Thus under additive separability, the tail conditions permit a much larger class of outcome models if a doubly robust approach is used.

**Proof of Proposition B.1.** The case of Assumption 3.7 is equivalent to the previous Section, therefore we only consider \( \alpha < 2 \) in the following. Note that since \( D_i(d)\sigma_d(X_i)/p_d(X_i) \) is nonnegative and independent from \( \varepsilon_i(d) \) it follows by Breiman (1965), Proposition 3 that \( D_i(d)\sigma_d(X_i)/p_d(X_i)\varepsilon_i(d) \) is in the domain of attraction of an alpha-stable distribution with \( \alpha \) being equal to the index of regular variation of \( D_i(d)\sigma_d(X_i)/p_d(X_i) \). Note that regular variation of the propensity score density in Assumption B.1 implies regular variation of any truncated moment and thus of the cumulative distribution function of \( D_i(d)/p_d(X_i) \) as well. Therefore

\[
P(D_i(d)\sigma_d(X_i)/p_d(X_i) \leq z) = P(D_i(d)/p_d(X_i) \leq z/\sigma_d(X_i))
\]

\[
= E[1(D_i(d)/p_d(X_i) \leq z/\sigma_d(X_i))]
\]

\[
= E[E[1(D_i(d)/p_d(X_i) \leq z/\sigma_d(X_i))\sigma_d(X_i)]]
\]

\[
= E[E[F_{D/p}(z/\sigma_d(X_i))]]
\]

\[
= \int_{-\infty}^{\infty} \left( \frac{z}{\sigma} \right)^{-\alpha_d} l\left( \frac{z}{\sigma} \right) f_\sigma(\sigma) d\sigma,
\]
where the last equality follows from regular variation of the cdf of $D_i(d)/p_d(X_i)$. Now for some $z > 0$ consider the limit of the ratio of the CDFs

$$\lim_{t \to \infty} \frac{P(D_i(d)\sigma_d(X_i)/p_d(X_i) \leq tz)}{P(D_i(d)\sigma_d(X_i)/p_d(X_i) \leq t)} = \lim_{t \to \infty} \left( \frac{tz}{t} \right)^{-\alpha_d} \frac{\int \sigma^\alpha_d l(tz/\sigma)f_\sigma(\sigma)d\sigma}{\int \sigma^\alpha_d l(t/\sigma)f_\sigma(\sigma)d\sigma} = \lim_{t \to \infty} \frac{tz}{t} \frac{\int \sigma^\alpha_d l(tz/\sigma)f_\sigma(\sigma)d\sigma}{\int \sigma^\alpha_d l(t/\sigma)f_\sigma(\sigma)d\sigma} \frac{l(tz)}{l(t)}.$$

Now note that by Seneta (1973) slowly varying functions are bounded on finite intervals far enough to the right. Thus, for $t$ large $l(t/\sigma)$ is bounded uniformly since $\sigma$ is compactly supported. The inverse of slowly varying functions are also slowly varying and thus bounded on the compact set for large $t$. By Assumption B.1 $\sigma^\alpha_d$ and $f_\sigma(\sigma)$ are uniformly bounded and hence dominated convergence can be applied. Note that by compactness, Karamata’s uniform convergence theorem applies to the ratios of the slowly varying functions, i.e. $l(t\sigma)/l(t)$ and $l(tz/\sigma)/l(tz)$ both converge to one uniformly over the support of $\sigma$. Thus it follows that

$$\lim_{t \to \infty} \frac{P(D_i(d)\sigma_d(X_i)/p_d(X_i) \leq tz)}{P(D_i(d)\sigma_d(X_i)/p_d(X_i) \leq t)} = z^{-\alpha_d} \frac{\int \sigma^\alpha_d \lim_{t \to \infty} l(tz/\sigma)/l(tz)f_\sigma(\sigma)d\sigma}{\int \sigma^\alpha_d l(t/\sigma)f_\sigma(\sigma)d\sigma} \lim_{t \to \infty} \frac{l(tz)}{l(t)} = z^{-\alpha_d} \frac{\int \sigma^\alpha_d f_\sigma(\sigma)d\sigma}{\int \sigma^\alpha_d f_\sigma(\sigma)d\sigma} = z^{-\alpha_d},$$

which implies regular variation of the cdf of $D_i(d)\sigma_d(X_i)/p_d(X_i)$ with same tail index $\alpha_d$ as the cdf of $D_i(d)/p_d(X_i)$. Moreover, one can use Breiman (1965), equation (3.2) and (3.3) to show that for $\alpha \neq 2$ the corresponding skewness parameter of the stable law is given by

$$\lim_{z \to \infty} \frac{1 - P(D_i(d)\sigma_d(X_i)\epsilon_i(d)/p_d(X_i) < z) - P(D_i(d)\sigma_d(X_i)\epsilon_i(d)/p_d(X_i) > -z)}{1 - P(D_i(d)\sigma_d(X_i)\epsilon_i(d)/p_d(X_i) < z) + P(D_i(d)\sigma_d(X_i)\epsilon_i(d)/p_d(X_i) > -z)} = \frac{1 - E[\epsilon_i^\alpha (d)]/E[\epsilon_i^\alpha (d)]}{1 + E[\epsilon_i^\alpha (d)]/E[\epsilon_i^\alpha (d)]},$$

which completes the proof.

### B.1.4 Beta Distribution Calculus

Assume that the propensity scores follow a beta distribution, i.e. $p_1(X_i) \sim B(a, b)$ for some $a, b > 0$. In order to characterize the limiting distribution of the DR estimator under Assumption B.1 it is sufficient to
only consider the truncated second moment of the inverse propensity scores (see Appendix B.1.3) given by

\[ E[1/p_1(X_i) \mathbb{1}(1/p_1(X_i) < z)] = \int_{1/p_1 > z} \frac{1}{p_1} \frac{p_1^{a-1}(1 - p_1)^{b-1}}{B(a, b)} dp_1, \]  

(B.1)

with \( B(a, b) \) being the beta function. Note that for \( 1/p_1 \to \infty \)

\[ \lim_{p_1 \to 0} \frac{(1 - \lambda p_1)^{b-1} B(a, b)}{(1 - p_1)^{b-1} B(a, b)} = 1, \]  

(B.2)

which by definition corresponds to slow variation as \( p_1 \to 0 \) and hence

\[ \int_{1/p_1 > z} \frac{1}{p_1} \frac{p_1^{a-1}(1 - p_1)^{b-1}}{B(a, b)} dp_1 = \int_{1/p_1 > z} p_1^{a-2}(1/p_1) dp_1. \]  

(B.3)

From Feller (1968), Section VIII.9, Lemma I, it follows that \( E[1/p_1(X_i) \mathbb{1}(1/p_1(X_i) < z)] \) is regularly varying with exponent \( a = a + 1 \), i.e. \( E[1/p_1(X_i) \mathbb{1}(1/p_1(X_i) < z)] \sim z^{1-a} \). Thus, for \( a \leq 1 \), we obtain the desired result. For \( a = b = 1 \), i.e. when the beta distribution is equal to the uniform, it follows that

\[ E[1/p_1(X_i) \mathbb{1}(1/p_1(X_i) < z)] = \int_{1/z}^{1} \frac{1}{p_1} dp_1 = \ln(z) \]  

(B.4)

which implies that

\[ \sqrt{\frac{n}{\ln(n)}} (\hat{\theta}_n - \theta_0) \overset{d}{\to} S(2, \beta, 0, c). \]  

(B.5)

B.2 Supplementary Material for Section 3.3

B.2.1 Index of Regular Variation for the Misspecified Propensity Scores

The example is chosen for its analytical simplicity, similar results can also be obtained through other variants of misspecification. For \( \gamma = (\gamma_0, \gamma_1)' \), the truncated variance of the misspecified inverse probability weights at truncation point \( z \) in then given by

\[
U_1(z) = E\left[ \frac{D_i(1)}{p_i^{[m]}(X_i, \gamma)^2} \mathbb{1}(D_i(1)/p_i^{[m]}(X_i, \gamma) < z) \right] \\
= \int_{0}^{1} \frac{p_1(x)}{p_i^{[m]}(x, \gamma)^2} \mathbb{1}(p_i^{[m]}(x, \gamma) > 1/z) dx \\
= \int_{g(z)}^{1} x[1 + \exp(-\gamma_0 - \gamma_1 \ln(x))]^2 dx \\
\equiv \int_{g(z)}^{1} a(x) dx
\]
with \( g(z) \equiv (z - 1)^{-1/\gamma} \exp(\gamma_0^{-1/\gamma}) \). By the Leibniz integral rule,

\[
\frac{\partial U_1(z)}{\partial z} = -a(g(z)) \frac{\partial g(z)}{\partial z}.
\]

Moreover, note that

\[
a(g(z)) = g(z) z^2 \frac{\partial g(z)}{\partial z} = g(z) \left( -\frac{1}{\gamma_1} \frac{1}{z - 1} \right) \exp(\gamma_0^{-1/\gamma}).
\]

To verify regular variation of the propensity score density at zero, we consider the limit of the ratio of the truncated variances as in Feller (1968), Section VIII.9:

\[
\lim_{t \to \infty} \frac{U_1(tz)}{U_1(t)} = \lim_{t \to \infty} \frac{a(g(tz)) g'(tz)}{a(g(t)) g'(t)} z
\]

\[
= \lim_{t \to \infty} \frac{g'(tz)^2}{g'(t)} \frac{g(tz)}{g(t)} \frac{t - 1}{t z - 1} z^3
\]

\[
= \lim_{t \to \infty} \left[ \frac{g'(tz)}{g'(t)} \right]^2 z^2
\]

for \( z > 1 \), where the first equality follows from L’Hospital’s rule as both numerator and denominator are assumed to diverge. Note that since

\[
\lim_{t \to \infty} \frac{g(tz)}{g(t)} = \lim_{t \to \infty} \exp \left( \frac{1}{\gamma_1} \ln \left( \frac{t - 1}{t z - 1} \right) \right)
\]

\[
= \lim_{t \to \infty} \left( \frac{t - 1}{t z - 1} \right)^{1/\gamma_1}
\]

\[
= z^{-1/\gamma_1}
\]

it follows that

\[
\lim_{t \to \infty} \frac{U_1(tz)}{U_1(t)} = z^{2 - 2/\gamma_1},
\]

which implies that \( U_1(z) \) is of regular variation with index \( 2 - 2/\gamma_1 \) or equivalently, the misspecified inverse probability weights have regular varying probability tails with tail index \( \alpha = 2/\gamma_1 \).

\[\text{B.2.2 Notes on Tail Index Estimation and Model Selection}\]

Instead of using the weights for the treatment effect of interest, estimation of the tail-index can also be done separately with potential outcome weights as the resulting indices are asymptotically equivalent. If both groups are equally irregular, using the combined inverse probability weights will be more efficient as more observations are used within a single estimation step. If extreme observations are mostly within the treated or the control group, the two approaches will generally be close to each other as the tail-index is
driven by the more extreme inverse probability weights. In principle, one could use the weighted outcome
data instead of the simple inverse probability weights. This, however, relies on outcome data and thus could
lead to additional problems regarding post-selection inference and goes against the design phase philosophy
outlined by Rubin (2007) that is widely appreciated in applied research.

For the tail index \( \alpha \) we suggest to use the estimator of Hill (1975), which is based on the \( k \) order statistics.
Let \( w_{n,i}[m] \leq \cdots \leq w_{n,n}[m] \) denote the order statistics of \( w_1^{[m]}, \ldots, w_n^{[m]} \). Hill’s estimator is given by

\[
\hat{\alpha}^{[m]}(k) = \frac{1}{k} \sum_{i=1}^{k} \log(w_{n,n-i+1}[m]) - \log(w_{n,n-k}[m]).
\]

The estimator is very sensitive to the choice of \( k \) in finite samples. Danielsson et al. (2001) suggest a double
bootstrap procedure, which provides a consistent choice of \( k \). It is based on minimizing a modified MSE
criterion, which has the same asymptotic behavior as the MSE minimizer used by Hall (1990) but generally
tends to provide more accurate estimates in finite samples.

We apply the model selection to the RHC data used in Section 6. Table B.1 contains point estimates,
conventional model selection criteria, and the estimated tail indices for the case of probit and logit together
with an empirical balancing approach. The empirical balancing approach uses a modified loss function to
equate the means of all covariates to be equal for the inverse probability weighted distributions, see e.g. Imai
and Ratkovic (2014), Zhao (2019), or Heiler (2020).

<table>
<thead>
<tr>
<th></th>
<th>Probit</th>
<th>Logit</th>
<th>Balancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPW estimate</td>
<td>-.0582</td>
<td>-.0601</td>
<td>-.0602</td>
</tr>
<tr>
<td>DR estimate</td>
<td>-.0609</td>
<td>-.0619</td>
<td>-.0608</td>
</tr>
<tr>
<td>Share of Correct Predictions</td>
<td>0.7310</td>
<td>0.7308</td>
<td>0.7276</td>
</tr>
<tr>
<td>Mean Absolute Standardized Difference</td>
<td>0.1068</td>
<td>0.1138</td>
<td>0.0000</td>
</tr>
<tr>
<td>Min. Generalized Propensity Score</td>
<td>0.0004</td>
<td>0.0033</td>
<td>0.0018</td>
</tr>
<tr>
<td>Max. ATE Weight</td>
<td>70.470</td>
<td>43.417</td>
<td>46.895</td>
</tr>
<tr>
<td>Tail Index (ATE Weight)</td>
<td>1.8912</td>
<td>1.8915</td>
<td>1.6292</td>
</tr>
</tbody>
</table>

The results in Table B.1 suggest that probit performs slightly better than logit in terms of predictive
accuracy and mean balancing but produces the most extreme propensity score. The logistic model is
very similar to the probit. Both have a tail-index of about 1.89 which suggests moderate deviations from
asymptotic normality and an estimated convergence rate of about \( n^{-0.47} \) up to a slowly varying function. By
construction, empirical balancing produces zero standardized differences for all covariates. Its most extreme
propensity score and ATE weight are more moderate than for the probit but slightly more extreme than for
the logit. However, it produces overall more propensities at the boundary leading to an estimated tail index of 1.63 which corresponds to an estimated convergence rate of about $n^{-0.39}$. Thus for given $n = 5735$, the normalizing constant is only about half the size for probit/logit compared to empirical balancing.

B.3 Supplementary Material for Sections 4 and 5

B.3.1 Asymptotic Variance and Self-normalizing Sums

For IPWII according to (3.1), the estimated variance/normalizing sum is given by

$$
\hat{V}_n[\hat{\theta}_{IPW_n}] = \hat{\kappa}_1 + \hat{\kappa}_0 - 2\hat{\omega}
$$

with

$$
\hat{\kappa}_d = \frac{1}{n} \sum_{i=1}^{n} (\hat{k}_{id}^{IPW} - \hat{b}_d S_i(\hat{\gamma}))^2
$$

$$
\hat{\omega} = \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{k}_{id}^{IPW} S_i(\hat{\gamma}) \right] \left[ \frac{1}{n} \sum_{i=1}^{n} S_i(\hat{\gamma}) \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} S_i(\hat{\gamma}) \hat{k}_{id}^{IPW} \right]
$$

$$
\hat{b}_d = \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{k}_{id}^{IPW} S_i(\hat{\gamma}) \right] \left[ \frac{1}{n} \sum_{i=1}^{n} S_i(\hat{\gamma}) \right]^{-1}
$$

$$
\hat{k}_{id}^{IPW} = \frac{D_i(d)}{p_d(X_i, \hat{\gamma})} (Y_i - \hat{\mu}_d)
$$

$$
\hat{\mu}_d = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(d)Y_i}{p_d(X_i, \hat{\gamma})} / \left( \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(d)}{p_d(X_i, \hat{\gamma})} \right)
$$

for $d = 0, 1$. Note that alternatively, every term containing $\hat{k}_{id}^{IPW}$ could be replaced by the sum-normalized version. For DRI according to (3.2), the estimated variance/normalizing sum is given by

$$
\hat{V}_n[\hat{\theta}_{DR_n}] = \hat{V}_1 + \hat{V}_0 + \hat{V}_\Delta
$$

with

$$
\hat{V}_d = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i(d)(Y_i - \mu_d(X_i, \hat{\delta}_d))^2}{p_d(X_i, \hat{\gamma})^2}
$$

$$
\hat{V}_\Delta = \frac{1}{n} \sum_{i=1}^{n} (\mu_1(X_i, \hat{\delta}_1) - \mu_0(X_i, \hat{\delta}_0) - \hat{\theta}_n^{DR})^2
$$

for $d = 0, 1$.

B.3.2 Power Properties of m-out-of-n Bootstrap

As discussed in Section 4, convergence of the $m$-out-of-$n$ distribution to $T_{\alpha, \beta}$ is pointwise for all $(\alpha, \beta) \in (1, 2] \times [-1, 1]$. One might be worried that this can lead to poor approximations at the boundaries of the parameter space - especially for the tail index. In the following we briefly sketch the arguments why we believe this can be problematic for tail indices close to one.
Allowing the tail index $\alpha = \alpha_n$ to vary with $n$ in a local asymptotic sense would allow for local to uniform and local to hyperbola tails of propensity score density. On a more detailed note, a local to uniform tail corresponds to a sequence $\alpha_n \to 2$, i.e. an asymptotically regular problem. In that case, the limiting distribution in Theorem 4.1 converges smoothly to the Gaussian for a tail index approaching two. Comparing to the regular case, the distribution of the test statistic has similar quantiles for tail indices close to 2. Thus, we do not expect a uniformity problem arising from a discontinuity in pointwise asymptotics around $\alpha = 2$. The case of $\alpha_n \to 1$, i.e. of a local to hyperbola tail, is more complicated. We have no reason to assume that the current procedure controls the asymptotic size in this case. Thus it is likely that for any sample size there is a process with a tail index $\alpha_n > 1$ such that the bootstrap approximation error is non-negligible. Testing in this neighborhood is further complicated by the fact that the convergence rates for the treatment effect estimators become arbitrarily slow for these sequences, see Propositions 3.1 and 3.2.

To further investigate this issue, we provide power curves for the $m$-out-of-$n$ bootstrap based on different simulation designs with varying tail indices in the following. Due to computational constraints we treat the propensity scores as given and choose bootstrap sample size $m$ fixed according to $n^{2(\frac{1}{\alpha} - 1)}$ with lower bound $\log(n)^2 + 1$ as in Section 4. In smaller simulation exercises we did not find major differences to the case where the propensity scores are estimated. The IPW type estimator is chosen as $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} D_i(1) Y_i p_1(X_i, \gamma) - \frac{1}{n} \sum_{i=1}^{n} D_i(0) Y_i p_0(X_i, \gamma)$. The simulation setup can be summarized as follows:

1. Design: mean functions from (5.1). Parameter of interest: average treatment effect. Propensity scores logistic as in Section 5. The tail index of the IPW estimator $\alpha$ changes with the change in $\gamma$ in (5.1) according to $\alpha = 1 + 1/\gamma$.

2. Design: mean functions from (5.1). Parameter of interest: average treatment effect. Propensity scores beta as in Section 2. The tail index of the IPW estimator $\alpha$ changes with the change in $a$ in (2.9) according to $\alpha = 1 + a$.

3. Design: linear mean function as in the description of Figure 2.2. Parameter of interest: potential outcome mean $E[Y_i(1)]$. Propensity scores logistic as in Section 5. The tail index of the IPW estimator $\alpha$ changes with the change in $\gamma$ in (5.1) according to $\alpha = 1 + 1/\gamma$.

4. Design: linear mean function as in the description of Figure 2.2. Parameter of interest: potential outcome mean $E[Y_i(1)]$. Propensity scores beta as in Section 2. The tail index of the IPW estimator $\alpha$ changes with the change in $a$ in (2.9) according to $\alpha = 1 + a$.

In general, the results suggest moderate size distortions in finite samples as tail indices get closer and closer to one. The power properties deteriorate smoothly with decreasing tail indices but are generally solid. One can see the clear reverse dependency between tail index and $n$: For small $\alpha$ designs at a large $n$ there is always a design with smaller $\alpha$ that delivers a similar power curve for a smaller $n$. For example consider Figure 1: It depicts the relationship between the degree of irregularity and the size and power properties of
the $m$-out-of-$n$ bootstrap for the IPW estimator in the first design. One can see that for $n = 5000$ the test starts to be distorted and to lose power around $\alpha = 1.3$, with power for this setting comparable to that at $\alpha = 1.5$ and sample size $n = 500$. From an empirical perspective, a tail-index of 1.3 is already very extreme and corresponds to a convergence rate slower than $n^{-1/4}$. It seems that $\alpha$ poses as a practical restriction for valid statistical inference similar to the dimensionality in the context of nonparametric function estimation.
The panels depict the power curves for the $m$-out-of-$n$ bootstrap based on $MC = 48000$ simulations and the choice of $m = \max\{n^2(\frac{\alpha - 1}{\alpha}), \log(n)^2 + 1\}$. The x-axis is reported in terms of the distance to the population average treatment effect. On each panel different curves correspond to different sample sizes.

Figure 1: Power Properties of the $m$-out-of-$n$ Bootstrap: Design 1

Figure 2: Power Properties of the $m$-out-of-$n$ Bootstrap: Design 2
The panels depict the power curves for the $m$-out-of-$n$ bootstrap based on MC = 48000 simulations and the choice of $m = \max\{n^2(\frac{\alpha - 1}{\alpha}), \log(n)^2 + 1\}$. The x-axis is reported in terms of the distance to the population average potential outcome. On each panel different curves correspond to different sample sizes.

Figure 3: Power Properties of the $m$-out-of-$n$ Bootstrap: Design 3

Figure 4: Power Properties of the $m$-out-of-$n$ Bootstrap: Design 4

The panels depict the power curves for the $m$-out-of-$n$ bootstrap based on MC = 48000 simulations and the choice of $m = \max\{n^2(\frac{\alpha - 1}{\alpha}), \log(n)^2 + 1\}$. The x-axis is reported in terms of the distance to the population average potential outcome. On each panel different curves correspond to different sample sizes.
B.3.3 \textit{m-choice} Alternatives

In the following we outline the explicit inference procedure for the IPWII estimator for the average treatment effect for the simulation design given in Section 5. Here step 3 of the \textit{m}-selection procedure is based on the estimated propensity scores from step 1. In general, step 3 can include the propensity score re-estimation. Potential differences are discussed below. As the simulation study is based on the bounded potential outcome mean functions (5.1), the \textit{m}-selection step 3 will be based on the simplified statistic:

\[
\hat{\theta}_{m}^{D/p} = \frac{1}{m} \sum_{i=1}^{m} \frac{D_{i}(1) - p(X_{i}, \hat{\gamma})}{p(X_{i}, \hat{\gamma})(1 - p(X_{i}, \hat{\gamma}))}.
\]  

(B.7)

The corresponding asymptotic variance, or normalizing sum under irregular identification, is given below in (B.8). Using \(\hat{\theta}_{m}^{D/p}\) in the \textit{m}-selection is beneficial compared to using \(\hat{\theta}_{n}\), as the conditional mean error of the potential outcomes is eliminated, which results in a faster stabilization of the Kolmogorov distance over \(B_{1}\). Thus the inference procedure in this special case reduces to:

1. Given sample \((Y_{i}, D_{i}, X_{i}')\), \(i = 1, ..., n\)
   i) Estimate propensity scores \(p(X_{i}, \hat{\gamma}) = \frac{1}{1 + \exp(-X_{i}'\hat{\kappa})}\).
   ii) Compute \(\hat{T}_{n} = \sqrt{\frac{n}{m_{n}} \hat{\theta}_{n} - \hat{\theta}_{0}} / \hat{V}[\hat{\theta}_{n}]\), where \(\hat{\theta}_{n}\) comes from (3.1) and \(\hat{V}[\hat{\theta}_{n}]\) is defined in Appendix B.3.1.

2. Choose \(m\)-grid of the form \(m_{j} = [0.75^n, j = 0, 1, ..., 40]\).
   Restrict \(m = \log(n)^2\).

3. for \text{\textbf{s}} from 1 to \(S=100\)
   Select \(m^{[s]}\) according to:
   i) For \(b_{1} = 1, ..., B_{1} = 10000\) and each element \(m_{j}\) in the \(m\)-grid, draw samples of size \(m_{j}\) with replacement from \((Y_{i}, D_{i}, p(X_{i}, \hat{\gamma}))\) and calculate \(T_{m_{j}}^{[b_{1}]} = \sqrt{m_{j}} \frac{\hat{\theta}_{m_{j}}^{D/p}[b_{1}]}{\hat{V}^{(b_{1})}[\hat{\theta}_{m_{j}}^{D/p}]}\) according to (B.7) and (B.8).
   ii) Choose \(m^{[s]} = \arg \min_{m_{j}} KS(T_{m_{j}}^{[s]}, T_{[m_{j+1}^{[s]}]}),\) with \(KS\) being an estimate of Kolmogorov sup-distance.

end loop

4. Calculate the optimal \(\hat{m} = \frac{1}{S} \sum_{s=1}^{S} m^{[s]}\)

5. For \(b_{2} = 1, ..., B_{2}\), draw \(\hat{m}\) observations with replacement from \((Y_{i}, D_{i}, X_{i}')\) and calculate \(T_{\hat{m}, n}^{[b_{2}]} = \sqrt{\frac{m}{\hat{m}} \frac{\hat{\theta}_{\hat{m}, n} - \hat{\theta}_{n}}{\hat{V}^{(\hat{m})}[\hat{\theta}_{\hat{m}, n}]}}\), with \(\hat{\theta}_{\hat{m}, n}\) and \(\hat{V}^{(\hat{m})}[\hat{\theta}_{\hat{m}, n}]\) being equivalents of \(\hat{\theta}_{n}\) and \(\hat{V}[\hat{\theta}_{n}]\), but estimated on a bootstrap sample of length \(\hat{m}\).

6. Calculate the p-value according to \(\frac{1}{B_{2}} \sum_{b_{2}=1}^{B_{2}} \mathbb{I}(|T_{\hat{m}, n}^{[b_{2}]}| > |T_{n}|)\).

The denominator for \(T_{m_{j}}^{[s]}\) is step 3 is given by:

\[
\hat{V}_{m}[\hat{\theta}_{m}^{D/p}] = \hat{\kappa}_{1} + \hat{\kappa}_{0} - 2\hat{\omega}
\]  

(B.8)
with
\[
\hat{k}_d = \frac{1}{m} \sum_{i=1}^{m} (\hat{k}_{id}^{D/p} - \hat{b}_d S_i(\hat{\gamma}))^2,
\hat{\omega} = \left[ \frac{1}{m} \sum_{i=1}^{m} \hat{k}_{id}^{D/p} S_i(\hat{\gamma})' \right]^{-1} \left[ \frac{1}{m} \sum_{i=1}^{m} S_i(\hat{\gamma}) S_i(\hat{\gamma})' \right]^{-1} \left[ \frac{1}{m} \sum_{i=1}^{m} S_i(\hat{\gamma}) \hat{k}_{id}^{D/p} \right]
\]

\[
\hat{k}_{id}^{D/p} = \frac{D_i(d) - p_d(X_i, \hat{\gamma})}{p_d(X_i, \hat{\gamma})}, \quad \hat{b}_d = \left[ \frac{1}{m} \sum_{i=1}^{m} \hat{k}_{id}^{D/p} S_i(\hat{\gamma})' \right]^{-1} \left[ \frac{1}{m} \sum_{i=1}^{m} S_i(\hat{\gamma}) S_i(\hat{\gamma})' \right]^{-1}
\]

Note, that the difference in the denominator (estimated asymptotic variance under regular identification) enters only in \( \hat{k}_d \) compared to \( T_n \) in Appendix B.3.1. In general, we recommend to choose the test statistic for the inference procedure to take into account the peculiarities of the problem at hand. Table B.2 summarizes different versions of the test statistic, which can be used as a test statistic in the step 3 of the \( m \)-selection algorithm.
of the test statistic, one needs to assume that either potential outcomes or their conditional mean errors converge to a non-zero bound for extreme propensity scores.

The table contains the different versions of the test statistic, which can be used in the re-estimation algorithm for the counterfactual expectation of the nuisance parameters. The table provides explicit expressions for the denominator of the test statistic. The first column provides options for re-estimation of the nuisance parameters: DR, IPW, and weighted residuals only. The second column provides explicit expressions for the denominator of the test statistic. The third column provides options for re-estimation of the nuisance parameters: DR, IPW, and weighted residuals only. The second column provides explicit expressions for the denominator of the test statistic.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Bounded</th>
<th>Pivotal</th>
<th>Re-estimated</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.2: Selection of Overviews
**B.3.4 Monte Carlo Distribution of $T_n$ for Doubly Robust Estimation**

Figure 5: QQ-plot of $T_n$ for the Doubly Robust Estimator under Irregular Identification.

QQ plots over 9000 simulation draws of the estimated test statistic as in (4.1) for the DR estimator for irregular identification design with $\alpha = 5/3$. Different panels correspond to sample sizes $n$. The simulation design is described in Section 5. On each panel the empirical quantiles of the test statistic are plotted against the standard normal quantiles.

Figure 5 depicts the QQ plots of the simulated distribution of the test statistic for the DR estimator as in (4.1) under irregular simulation design. Although the test statistic does not exactly match the normal distribution, the extreme quantiles which are used for inference are close to the standard normal ones by coincidence.
B.3.5 Monte Carlo Distribution of $\hat{m}$

Figure 6: Distribution of $\hat{m}$ in the $m$-out-of-$n$ Bootstrap

Histograms over 9000 simulation draws of the block length $\hat{m}$ for the DR estimator chosen by a bootstrap aggregation algorithm discussed in B.3. $\bar{m}$ denotes the average block length chosen over all simulation draws. The simulation design is described in Section 5. Left column corresponds to the regular identified case with $\gamma = 0.5$ with the tail index $\alpha = 2$ and the right column corresponds to the irregular identified case with $\gamma = 1.5$ with the tail index $\alpha = 5/3$. Rows correspond to the sample sizes. Histograms are normalized to probability.

Figure 6 depicts the $\hat{m}$-choices distribution over 9000 Monte Carlo iterations for the inference procedure described in Section 4 and B.3. The simulation design corresponds to one described in Section 5. One can see that for the regularly identified cases the algorithm chooses larger $m$'s and for irregular cases the probability mass on the histograms shifts towards smaller values (relative to $n$) with increasing $n$. Thus the proposed $m$-selection procedure shows some adaptation to the type of identification.
References Supplementary Material


