Heavy Tail Robust Estimation and Inference for Average Treatment Effects

Saraswata Chaudhuri* and Jonathan B. Hill†
Dept. of Economics
University of North Carolina

May 23, 2014

Abstract

We study the probability tail properties of the Inverse Probability Weighting (IPW) estimators of the Average Treatment Effect (ATE) when there is limited overlap between the covariate distributions of the treatment and control groups. Under the ignorability assumption, such limited overlap is manifested in the propensity score for certain units being very close (but not equal) to 0 or 1. The existing IPW estimators: (i) are either based on the assumption of strict overlap, i.e., the propensity score is bounded away from 0 and 1, (ii) or they trim a fixed or shrinking portion the random variable Z that identifies ATE by $E[Z] = \text{ATE}$ when the propensity score is close to 0 or 1: equivalently, under certain circumstances, when the covariates are small or large. We show that such IPW estimators perform poorly under limited overlap. We propose a new tail-trimmed IPW estimator of ATE whose performance, unlike that of the existing IPW estimators, is robust to limited overlap more generally. This new estimator negligibly trims Z adaptively by its large values and thus sidesteps dimensionality, bias and poor correspondence properties associated with trimming by the covariates. We use higher order asymptotics to determine a reasonable trimming policy. The estimator is asymptotically normal and unbiased whether there is limited or strict overlap. We characterize the probability tail decay of Z using a threshold crossing model for treatment assignment. We show when Z exhibits power law tail decay due to limited overlap, and when it has an in finite variance, in which case existing IPW estimators do not have a Gaussian distribution limit. Monte Carlo experiments corroborate our theoretical results and demonstrate that in finite samples our estimator has significantly lower bias and mean-squared error, and is closer to normal than the existing IPW estimators of ATE.

JEL Classification: C12; C13; C30.

*Dept. of Economics, University of North Carolina at Chapel Hill, Saraswata_Chaudhuri@unc.edu.

†Corresponding author. Dept. of Economics, University of North Carolina at Chapel Hill, www.unc.edu/∼jbhill, jbhill@email.unc.edu.

AMS Classification: 62F12; 62F35.
Keywords: average treatment effect; limited overlap; tail trimming; robust estimation

1 Introduction

The strong ignorability assumption of Rosenbaum and Rubin (1983) can (nonparametrically) point identify the average treatment effect (ATE) in observational studies. Strong ignorability requires the existence of a set of observed covariates X satisfying two conditions: *unconfoundedness* of the treatment assignment conditional on the observed covariates, and *strict overlap* in the distribution of the observed covariates for the treatment and the control groups. We maintain throughout the assumption of perfect compliance, that is the treatment is taken *if and only if* it is assigned.

In this paper we focus on the *strict overlap* or common support condition, which requires the propensity score, which is defined as the probability of taking the treatment conditional on the observed covariates X, to be bounded away from zero and one. This is hard to satisfy in practice, in which case ATE for the population of interest is generally not point identified and the corresponding estimators do not have their intended meaning. See, amongst others, Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (1999), Frolich (2004), Lechner (2008), Busso, DiNardo, and McCrary (2009), and Crump, Hotz, Imbens, and Mitnik (2009).

We relax the strict overlap assumption partially by allowing the propensity score to be arbitrarily close to zero or one. Following Khan and Tamer (2010a) we refer to this relaxation of strict overlap as the *limited overlap* condition.[1] Limited overlap is useful since it accommodates conventional models where taking the treatment depends on a latent variable crossing some threshold (e.g. Busso, DiNardo, and McCrary (2009)). While limited overlap still allows for point identification of ATE, in the terminology of Khan and Tamer (2010a) this is actually *irregular* identification. Consequently the tails of the Inverse Probability Weighting (IPW) estimators of ATE may get thicker causing instability in estimation and inference and breakdown of the standard asymptotic properties such as $\sqrt{n}$ -convergence and asymptotic normality. The problem as we discuss below is the variable Z that point identifies ATE may have a Paretian distribution tail and therefore may have an infinite variance: identification is irregular precisely because Z may not belong to the domain of attraction of a normal law. Hence conventional estimators can have non-Gaussian limits when properly scaled (cf. Ibragimov and Linnik, 1971) and robust estimators can have a slower than $\sqrt{n}$ convergence rate (Khan and Tamer, 2010a). The sensitivity of location estimators to heavy tailed data in general is well known, dating to Bahadur (1960). See Jureckova (1981).

Our paper makes three contributions to the literature. First, under the framework of Khan and Tamer (2010a) that treats the propensity score as known, we extend their work and provide a detailed

[1] Crump, Hotz, Imbens, and Mitnik (2009) use limited overlap in a broader empirical sense, in particular "parts of the covariate space with limited numbers of observations for either the treatment or control group". See p. 188.

characterization of the effect of the relative tail behavior of the covariates X and the unobserved errors on subsequent estimation and inference based on IPW estimators. In the conventional threshold crossing models for treatment assignment, we characterize when the variable that point identifies ATE has a power law distribution tail, and possibly an infinite variance. Although an infinite variance does not guarantee a standard ATE estimator will have a non-Gaussian limit,[2] this nevertheless suggests the need for a heavy tail robust estimator that ensures standard inference. In a completely general environment, distribution tails have a complicated form that may or may not decay according to a power law. Thus, unless the practitioner knows error and covariate distributions, and the data generating process of the treatment assignment, an estimator that is robust to heavy tails and at the same time is also first order asymptotically equivalent to a non-robust estimator when tails are thin is imperative.

We therefore propose as the second contribution a new tail-trimmed IPW estimator of ATE that, unlike existing estimators, is *robust* in the sense that it is consistent, asymptotically unbiased and normally distributed even under limited overlap, irrespective of heavy tails. Further, it is first order asymptotically equivalent to a conventional IPW estimator when limited overlap is not substantial enough to render heavy tails. We negligibly trim the variable Z that point identifies the ATE adaptively by its large values (the sample portion of extremes trimmed vanishes to zero asymptotically).

In the ATE literature trimming and truncation is based on one of the following: i) removing (trimming) units from treated and control groups for which there are no comparable units in the opposite group (e.g. Heckman, Ichimura, Smith, and Todd, 1998; Frolich, 2004; Crump, Hotz, Imbens, and Mitnik, 2009; Busso, DiNardo, and McCrary, 2009), in which case the ATE thus estimated may not correspond to the population of interest; ii) capping (truncating) the propensity score (e.g. Frolich, 2004; Lee, Lessler, and Stuart, 2011; Chaudhuri and Min, 2012); or iii) trimming Z based on sample extremes of specific covariates (e.g. Khan and Tamer, 2010a). All three variously use X to determine the data transformation. Trimming by sample extremes of X poses a potentially severe dimensionality problem, and trimming by X or propensity score or capping the propensity score can cause asymptotic bias. Moreover, there may only be a weak correspondence between Z and the X, hence the above strategies using X need not impact the resulting ATE estimator in some samples. Indeed, all existing estimators that incorporate trimming to estimate ATE are either not robust to limited overlap and therefore heavy tails, or are asymptotically biased, and we show by simulation that trimming by the covariates can lead to a severe drop in efficiency in order to render an approximately normal estimator. By trimming *adaptively* by Z we sidestep dimensionality and poor correspondence issues. By trimming *negligibly* we ensure asymptotic normality in general, and we can use extreme value theory to approximate and estimate the bias for an asymptotically unbiased estimator. See Section 2.2 for further discussion and references.

As a third contribution, negligible trimming ensures asymptotic unbiasedness. However, by first

[2]See Chapter 9 in Feller (1971), and recently Chritsopeit and Werner (2001).

order asymptotics the bias-corrected estimator is unbiased and trimming more observations per sample leads asymptotically to greater efficiency. Obviously trimming more incurs greater small sample bias, even with bias estimation. We use higher order asymptotics to show how the trimming choice impacts the various components of our bias-corrected estimator. The higher order bias is always smaller when trimming is less. Technical requirements for our theory to hold plus higher order asymptotics together shed much light on how to choose the number of trimmed sample extremes of Z. See Section 3.5. Evidently this is first such use of higher order asymptotics for an optimal bias-corrected negligibly trimmed sample mean. See also Peng (2001) and Hill (2013).

It is important to recognize that the goal of our paper is fundamentally different from that of the conventional use of trimming in the ATE literature. In the latter, the focus is either to put bounds on the ATE (e.g. Lechner (2008)) or to locate a suitable region of common support to point identify the ATE for a subpopulation (that may or may not be the population of interest) defined by the common support and achieve internal validity of the ATE estimator. See Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (1999), Crump, Hotz, Imbens, and Mitnik (2009), Lee, Lessler, and Stuart (2011), and Traskin and Small (2012). In contrast, the ATE is already point identified under limited overlap. Our tail-trimmed IPW estimator is designed to overcome the problems of the existing IPW estimators that are associated with irregular identification as noted by Khan and Tamer (2010a) and discussed in Sections 2 and 4 below.

Assuming the propensity is score is known is clearly a shortcoming. Our primary objectives are explaining *why* identification is irregular, and *how* to render standard asymptotics with unbiasedness in a way that efficiently uses sample information and leads to a non-ad hoc choice for the trimming amount. Thus, we are not primarily concerned with estimating the propensity score. Introducing a plug-in estimator for the propensity score involves a choice of estimator and therefore a more involved limit theory. It will also affect the heavy tailedness of the variable that identifies the ATE. Including such a plug-in is certainly feasible, and suggests the next step in the robust estimation of ATE, but is beyond this paper's scope.

The rest of the paper is organized as follows. In Section 2 we motivate our estimator by describing the framework, discussing the problem of ATE estimation under limited overlap, and detailing existing methods to deal with it. We then introduce our new tail-trimmed estimator in Section 3 and present its asymptotic properties under a general set of high level assumptions. In Section 4 we specialize these assumptions to the conventional latent variable threshold crossing framework with separable error and covariate for treatment assignment. In this setting we characterize the distribution tails of the variable that identifies ATE. This framework is widely used (see Vytlacil (2002)) and hence is beneficial for appreciating why, where and how our estimator is robust to limited overlap. In Section 5 we consider other proposals of tail-trimming and present an improved estimator based on ideas developed in Khan and Tamer (2010a). Finally, we perform Monte Carlo experiments in Section 6 in order to demonstrate the degree of heavy tailedness in the variable that point identifies ATE, and

compare our robust unbiased estimator with existing estimators. Our estimator performs best overall within our simulation design: it exhibits the least bias and is closest to normal in a variety of cases associated with limited overlap. The other estimators considered in the experiment either exhibit bias when the limited overlap is strong and are far from normal which leads to poor inference, or require substantial trimming and therefore a severe drop in efficiency to be competitive in some cases.

We use the following notation. $a_n \sim b_n$ implies $a_n/b_n \to 1$ as $n \to \infty$. K is a positive finite constant, the value of which may change from line to line. $[z]$ is the integer part of z. $a \wedge b = \min\{a, b\}$.

2 Framework and literature review

2.1 IPW estimators under limited overlap

Let D be a binary variable such that $D = 1$ if the treatment is taken and $D = 0$ otherwise. Let $Y_1 := Y(D = 1)$ and $Y_0 := Y(D = 0)$ denote the potential outcomes. See Rubin (1974). Our object of interest is the population ATE:

$$\theta := E[Y_1 - Y_0]. \tag{1}$$

Y_1 and Y_0 cannot be simultaneously observed for the same unit: we only observe the realized outcome

$$Y = DY_1 + (1 - D)Y_0.$$

This causes a problem in observational studies where the treatment assignment is non-random, because the difference in the expected realized outcome for the treatment and the control groups $E[Y|D = 1] - E[Y|D = 0]$ cannot identify θ in general.

Identification of θ can, however, be achieved by the following *strong ignorability* assumption. Assume there exists a set of observed covariates X such that the following two conditions are satisfied (throughout $\perp$ expresses independence):

A1. Unconfoundedness: $Y_1, Y_0 \perp D|X$

A2. Strict Overlap: $0 < p_* \leq p(X) := P(D = 1|X) \leq 1 - p_* < 1$ *a.s.* for a constant p_*.

Then from A1 and A2:

$$E\left[\frac{D}{p(X)}Y - \frac{1-D}{1-p(X)}Y\right] = E\left[E\left[\frac{D}{p(X)}Y_1 - \frac{1-D}{1-p(X)}Y_0\middle| X\right]\right] = E\left(E[Y_1|X] \ - E[Y_0|X]\right) = \theta. \tag{2}$$

The IPW estimator of ATE proposed by Hirano, Imbens, and Ridder (2003) rests on this result and is defined as the sample analog of the left-hand-side of (2). They use a nonparametric estimator of the propensity score $p(X)$. Under A1 and A2 and standard regularity conditions, their estimator is

consistent, asymptotically unbiased and normally distributed. Under a suitable rate of convergence of the nonparametric estimator of $p(X)$ the asymptotic variance also attains the semiparametric efficiency bound (SEB)

$$E\left[\frac{V(Y_1|X)}{p(X)}+\frac{V(Y_0|X)}{1-p(X)}+\{(E[Y_1|X]-E[Y_0|X])-\theta\}^2\right]. \tag{3}$$

Alternatively, one could plug-in a correctly specified parametric estimator of $p(X)$ or, if known, even the true $p(X)$ to obtain IPW estimators that are consistent, asymptotically unbiased and normally distributed. Incidentally, the asymptotic variance for both are generally larger than the SEB in (3), the last one being the largest (Hahn (1998), Wooldridge (2007), Graham (2011)). It is straightforward to extend our discussion to the second case, while for the first it may be beneficial to directly work with the efficient influence function. Demonstration of power law tails in Section 4 will, however, be different and possibly complicated in these two cases. This is left for future research. In this paper, to fix ideas we simply focus on the last case where $p(X)$ is assumed to be known.

In the absence of strict or limited (as defined below) overlap in the support of the covariate (X) distributions for the treatment ($D=1$) and the control ($D=0$) groups, we have $p(X)=0$ or $p(X)=1$ with positive probability. This is a direct violation of A2 and is a well recognized problem with IPW estimators.[3] As is evident from (2), in the non-overlapping regions of the supports of the covariate distributions, identification and estimation of θ require extrapolation. But the validity of extrapolation cannot be guaranteed without imposing restrictions on the response surfaces $E[Y_1|X]$ and $E[Y_0|X]$. Thus, ATE θ is generally not (nonparametrically) point identified without further assumptions, and hence IPW estimators may be misleading.

In this paper we abstract from such severe, albeit realistic, non-overlap possibilities, and instead focus on the case of *limited overlap* that may indeed be difficult to rule out even after careful balancing of the covariates X by the analyst.[4] The terminology is borrowed from Khan and Tamer (2010a) and differs from its broader use by Crump, Hotz, Imbens, and Mitnik (2009).

A2′. Limited Overlap: $0<p(X):=P(D=1|X)<1$ *a.s.*

Although trivially A2′ nests strict overlap A2, the problem is far more subtle under A2. ATE θ is point identified but, as Khan and Tamer (2010a) showed, under A1 and A2′ the SEB is infinity, and not (3). This can lead to instability due to a slower than parametric rate of convergence of IPW estimators and large variance. For an intuitive understanding of the limited overlap problem,

[3]The nature of the problem is similar to that faced by the widely used matching estimators described in Heckman, Ichimura, and Todd (1998) and Dehejia and Wahba (1999).

[4]If the covariate distributions for the treatment and the control groups do not overlap for, say, $X\in\tilde{\mathcal{X}}\subset\mathcal{X}$, with $\mathcal{X}$ the support of X, it follows $p(X)=0$ or $p(X)=1$ *a.s.* for $X\in\tilde{\mathcal{X}}$. As long as $X\in\tilde{\mathcal{X}}$ has positive measure then $p(X)\in\{0,1\}$ occurs with positive probability. Limited overlap in the sense of A2′ rules out this situation.

consider $E[Y_1]$, and note

$$\begin{aligned}(i)\quad E\left[\tfrac{D}{p(X)}Y\right] &= E\left[\frac{D}{p(X)}Y_1\right] = E\left(E[Y_1|X]\right) = E[Y_1] \\ (ii)\quad V\left(\tfrac{D}{p(X)}Y\right) &= V\left(E[Y_1|X]\right) + E\left[\frac{1-p(X)}{p(X)}\left(E[Y_1|X]\right)^2 + \frac{1}{p(X)}V(Y_1|X)\right].\end{aligned}$$

Equation (i) represents the balancing property of the propensity score by which it assigns more weight to the observed outcome of a treated unit if, based on the observed covariates X, the unit was less likely to take the treatment (i.e., small $p(X)$). Thus it resembles the population of interest, obtains point identification and, consequently, an unbiased estimate of $E[Y_1]$. Equation (ii) represents the consequence of limited overlap where a *small* $p(X)$ leads to a *large* weight. Unless the proportion of treated units with *small* $p(X)$ is sufficiently low, estimation of $E[Y_1]$ is effectively determined by the realized outcomes of these units, as it should be by the IPW principle of Horvitz and Thompson (1952). However, this may also lead to instability and breakdown of the standard asymptotic properties of the IPW estimator of $E[Y_1]$, cf. Khan and Tamer (2010a).

Moreover, consider that $1/p(X)$ alone may be exceptionally heavy tailed. As a pathological example, if $p(X)$ is uniformly distributed then $1/p(X)$ is Pareto distributed with index 1 since $P(1/p(X) \geq c) = 1/c$, hence the mean of $1/p(X)$ is infinite. On the other hand, if the distribution of $p(X)$ concentrates probability near 0 then the tails of $1/p(X)$ are even heavier. The challenge, however, with characterizing $V(DY/p(X))$ is the obvious fact that $1/p(X)$, D and Y are all possibly dependent.

This is the limited overlap problem. A similar intuition applies to the IPW estimators of $E[Y_0]$ and the ATE $\theta := E[Y_1 - Y_0]$. Based on our reading, it seems that the consequences of the limited overlap problem even with the true $p(X)$ are less recognized in the literature.[5]

2.2 Existing methods to deal with the limited overlap problem

The above discussion hints at the possibility that in certain cases where the proportion of units with *small or large* $p(X)$ is not sufficiently low to prevent instability, but low enough to guarantee existence of θ, one could possibly remove some or all of these units and thus trim the tails of the distribution of the IPW estimator to restore the standard asymptotic properties. One strand of the literature attempts this by capping the weights, that is by truncating extreme observations of $p(X)$ by percentile cutpoints like 1 and 99 or by fixed cutpoints p_* and $1-p_*$, thereby mimicking the strict

[5]Note that the limited overlap problem is due to the tail behavior of the true propensity score $p(X)$, and is fundamentally different from the problem discussed in Kang and Schafer (2007). In their case, the problem with IPW-type estimators arises due to parametric mis-specification of the propensity score model that may result in the parametrically estimated $p(X)$ incorrectly being close to zero or one. Proposed solutions include re-weighted IPW (Lunceford and Davidian (2004), Busso, DiNardo, and McCrary (2009, 2011)), the generalized boosted model (McCaffrey, Ridgeway, and Morral (2004), Ridgeway and McCaffrey (2007)), enhanced projection (Cao, Tsiatis, and Davidian (2009)), and inverse probability tilting (Graham, Pinto, and Egel (2011)). Although important and useful for practical purposes, these are not applicable here.

overlap assumption (A2) in the sample. See, Lee, Lessler, and Stuart (2011) and Chaudhuri and Min (2012) respectively. Capping the weights trims the tails of the distribution of the IPW estimator and reduces instability, but this is ad hoc and can increase bias substantially. Nevertheless, Lee, Lessler, and Stuart (2011) give simulation evidence supporting percentile cutpoints, while Frolich (2004) finds capping works better than removing the concerned units altogether as is done by the conventional trimming rules with the IPW estimators. Potter (1993) explores different cutpoint selection methods based on minimizing a suitably chosen mean squared error function. These methods are still ad hoc and, to our knowledge, the asymptotic properties of the resulting estimators are not completely known.

A second and more conventional strand of the literature involves the removal of units from the treated and the control groups for which there is no comparable units in the opposite group. See, for example, Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (1999), Ho, Imai, King, and Stuart (2007), Crump, Hotz, Imbens, and Mitnik (2009), and Traskin and Small (2012). These trimming rules were predominantly designed in the context of matching estimators to obtain internal validity of the estimates, while Crump, Hotz, Imbens, and Mitnik (2009) applies generally. However, the resulting estimator may or may not identify ATE for the original population unless the treatment effect is homogeneous.[6] See Busso, DiNardo, and McCrary (2009), and see Frolich (2004) for simulation evidence.

A third strand of the literature, closest in spirit to the present study, is due to Khan and Tamer (2010a). They assume $D = I(\alpha + \beta X - U \geq 0)$ where X is a scalar covariate/index (or, in a different context, a "special regressor" as in Lewbel (1997)), and U is a random error independent of X. Khan and Tamer (2010b) show asymptotic normality is assured by removing units Z with $|X| > \gamma_n$ where γ_n is a positive number and n is the sample size. The proposed estimator based on the observed sample $\{Y_i, D_i, X_i\}_{i=1}^n$ trims by X_i:

$$\theta_n^{(tx)} := \frac{1}{n}\sum_{i=1}^{n} h(X_i) Y_i I\left(|X_i| \leq \gamma_n\right) \tag{4}$$

where $\gamma_n \to \infty$ as $n \to \infty$, $I(A)$ denotes an indicator variable for the event A, and

$$h(X_i) := \frac{D_i}{p(X_i)} - \frac{1 - D_i}{1 - p(X_i)} = \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))}.$$

Several features of their method are worth noting. First, Khan and Tamer (2010a) characterize the rate of convergence of $\theta_n^{(tx)}$ under the normalization $\beta = 1$ and assuming X_i and U_i are logistic or

[6] Ho, Imai, King, and Stuart (2007) and Traskin and Small (2012) remove units directly based on the covariates X, while the other papers use trimming rules based on $p(X)$. These two papers are less known to economists, so we point out the distinctive features. Ho, Imai, King, and Stuart (2007) avoid extrapolation by identifying a suitable convex hull of X and thereby on relying on interpolation. Traskin and Small (2012) gain interpretability of the resulting subpopulation in terms of the observed covariates X by using classification trees based on X. In the context of limited overlap such distinctions are moot because ATE θ for the original population is already point identified.

normal (which fixes $Var(U_i)$). In their setting the rate $(n/\ln(n))^{1/2}$ aligns identically with a sample mean of an iid random variable with power law distribution tails and a tail index of exactly 2 such that the variance of $h(X_i)Y_i$ is infinite, cf. Leadbetter, Lindgren, and Rootzen (1983) and Resnick (1987). But this fact is neither shown nor discussed in the literature.

Second, since our ATE estimator does not exploit the case of a "special regressor" as in Lewbel (1997) and subsequently in Khan and Tamer (2010a, Section 3 and implicitly in Section 4), we take an alternative but equivalent approach of normalization with the nuisance parameters β and $Var(U_i)$ that are not identified separately. In particular, we fix $Var(U_i) = 1$ and let $\beta > 0$ vary. We demonstrate in Section 4 that the tail decay of $h(X_i)Y_i$ is monotonic in β, with heavier tails and infinite variance occurring with $\beta \geq 1$. The converse is true if we fix $\beta = 1$ as in Lewbel (1997) and Khan and Tamer (2010a) and let $Var(U_i)$ vary: heavier tails align with larger $Var(X_i)/Var(U_i)$. This points to a natural signal-noise property: heavier tails align with a stronger signal (i.e. large β or large $Var(X_i)$) and smaller noise (i.e. small $Var(U_i)$). The variables $u_i := U_i/\beta$ and U_i have different dispersions when $\beta \neq 1$ which can have a dramatic impact on the tails of $h(X_i)Y_i$ and therefore on IPW estimators of ATE. To the best of our knowledge, a complete characterization of the rate of convergence or asymptotic distribution for $\theta_n^{(tx)}$ in this more general setting, where either β or $Var(U_i)$ is arbitrary, is not available.

Third, it is not clear how a covariate trimming rule should be modified when *multiple* covariates are required to ensure that assumption A1 holds. This limits the practical usefulness of $\theta_n^{(tx)}$ because, absent specific knowledge of the selection process, one would often condition on many pre-treatment covariates X_i to ensure A1. Possible solutions could be trimming based on $p(X_i)$, as in Crump, Hotz, Imbens, and Mitnik (2009) when $V(Y_1|X) = V(Y_0|X)$ is a constant X-*a.s.*, or based on the weight $h(X_i)$. Both are related to the literature on weight capping discussed above. However, $h(X_i)Y_i$ and not $h(X_i)$ identifies θ, hence if $E[h^2(X_i)Y_i^2] = \infty$ then *in general* trimming sufficiently many of the largest realizations of $|h(X_i)Y_i|$ will *guarantee* asymptotic normality irrespective of the relationship between covariate X, propensity score $p(X)$ and realized outcome Y, Csörgo, Horváth, and Mason (cf. 1986); Hahn, Weiner, and Mason (cf. 1991); Hill (cf. 2012, 2013). The only way to ensure sample extremes of $h(X_i)Y_i$ are removed is to use $h(X_i)Y_i$ itself as the trimming criterion, and not simply the covariate X_i. Indeed, the correspondence between large X_i and large $h(X_i)Y_i$ may be weak since trimming by X_i neglects information from the unobserved component U_i. See Section 5.

Fourth, tail-trimmed estimators like $\theta_n^{(tx)}$ may be asymptotically biased. See Csörgo, Horváth, and Mason (1986) and Peng (2001) for a general theory, and see Sections 3 and 4 in Khan and Tamer (2010a). Indeed and somewhat trivially, unless $\theta = 0$ and $h(X_i)Y_i$ has a symmetric distribution then in general $E[h(X_i)Y_iI(|X_i| \leq \gamma_n)] \neq \theta$. Hence $\theta_n^{(tx)}$ is in general biased when $h(X_i)Y_i$ has an asymmetric distribution or $\theta \neq 0$. Further, as noted above the use of X_i to determine which $h(X_i)Y_i$ are too large to support a stable, and asymptotically normal, estimator can lead to sub-optimal performance. If $\beta > 1$ then $\theta_n^{(tx)}$ can still be sensitive to limited overlap since $h(X_i)Y_i$ may have very heavy tails and

using X_i as the trimming criterion does not strictly lead to the trimming of extreme observations of $h(X_i)Y_i$. Thus $\theta_n^{(tx)}$ exhibits small sample bias due to influential extreme observations, even if it is asymptotically unbiased, and it is demonstrably non-normal in small samples. See our simulation study in Section 6 for verification of these claims.

Fourth, Khan and Tamer (2010a) use mean-squared-error minimization to select $\{\gamma_n\}$ for scalar X_i, when $\{U_i, X_i\}$ are logistic or or normal random variables. In practice, however, different covariates may have different and unknown distributions. Further, the use of pre-chosen and non-random trimming thresholds γ_n may result in an unstable estimator in small samples since large $h(X_i)Y_i$ are not guaranteed to be trimmed. See Lewbel (1997) and Andrews and Schafgans (1998) for use of similar trimming rules in slightly different contexts. In the literature there does not yet exist an adaptive method for deciding what γ_n should be.[7]

There is little guidance in the literature regarding the trimming level (see also Stuart, 2010, p. 10), and in the ATE literature a solution for asymptotic bias under trimming is not proposed. Our estimator proposed below seeks to fill these gaps. Our trimming rule is adaptive and is based on $h(X_i)Y_i$, as opposed to X_i or $p(X_i)$. We demonstrate by simulation that only trimming $h(X_i)Y_i$ when $h(X_i)Y_i$ is an extreme value leads to a sharp and approximately normal estimator in general. Since the very choice of γ_n can render a poorly performing $\theta_n^{(tx)}$, in Section 5 we propose an adaptive version of $\theta_n^{(tx)}$ by using an order statistic of X_i for the threshold. This new estimator has similar asymptotic properties as $\theta_n^{(tx)}$, but is far sharper and closer to normal in general as long as it is set to trim a large portion of the sample due to the potentially weak correspondence between sample extremes of X_i and $h(X_i)Y_i$.[8] By using an order statistic for the threshold we can control the amount of trimming, and by forcing heavy trimming we can ensure the largest $h(X_i)Y_i$ are removed. This is key to making estimators like $\theta_n^{(tx)}$ work in practice. Our estimator, by comparison, directly trims only extreme observations of $h(X_i)Y_i$, hence it performs the best of those estimators considered in this paper.

Finally, it should be remembered that the ATE is already identified under limited overlap and hence our focus is beyond internal stability. Therefore, the careful approach of Crump, Hotz, Imbens, and Mitnik (2009) of not involving the outcome Y_i in the trimming rule in order to avoid *deliberate bias with respect to the treatment effects being analyzed* is not necessary for our purpose.

[7] A recent paper by Klein, Shen, and Vella (2011) proposes data dependent trimming in the context of selection models with binary outcome and covariate. Their trimming rule is to enforce identification at infinity by ruling out observations for which probability of selection is not close to 1. See Heckman (1990) and Andrews and Schafgans (1998) for the related early literature.

[8] In our simulation experiment $\mathring{\theta}_n$ requires 43 of all 100 observations to be trimmed for approximate normality and therefore valid asymptotic inference, versus 9 of 100 for our estimator. See Section 6 where we also provide theory based details and a numerical demonstration that explain these simulation results.

3 Tail-Trimmed ATE Estimation

We present our core estimator and then discuss asymptotic bias. We then present a bias-corrected estimator, and finally an estimator of the asymptotic variance.

3.1 The Tail-Trimmed Estimator

Our goal is IPW estimation and inference of θ using the observed sample $\{Y_i, D_i, X_i\}_{i=1}^n$ on n units drawn at random from the population of interest. The fact that $h(X_i)Y_i$ under limited overlap has support $\mathbb{R}$ and therefore may have an infinite variance suggests we use a classic tail-trimmed mean for estimating θ. See, for example, Csörgo, Horváth, and Mason (1986), Hahn, Kuelbs, and Samur (1987), Hahn, Weiner, and Mason (1991) and Hill (2013), and see Section 4 below for details on the probability tail decay properties of $h(X_i)Y_i$.

Write

$$h_i := h(X_i) := \frac{D_i}{p(X_i)} - \frac{1-D_i}{1-p(X_i)} \quad \text{and} \quad Z_i := h_i Y_i.$$

Define sample order statistics of mean centered Z_i:

$$\hat{Z}_{n,i} := Z_i - \frac{1}{n}\sum_{j=1}^n Z_j, \;\; \hat{Z}_{n,i}^{(a)} := \left|\hat{Z}_{n,i}\right| \quad \text{and} \quad \hat{Z}_{n,(1)}^{(a)} \geq \hat{Z}_{n,(2)}^{(a)} \geq \cdots \geq \hat{Z}_{n,(n)}^{(a)},$$

and let $\{k_n\}$ be an intermediate order sequence: $k_n \in \{1, ..., n\}$, $k_n \to \infty$ and $k_n/n \to 0$. The tail-trimmed IPW estimator is

$$\hat{\theta}_n^{(tz)} := \frac{1}{n-k_n}\sum_{i=1}^n Z_i I\left(\left|Z_i - \frac{1}{n}\sum_{j=1}^n Z_j\right| < \hat{Z}_{n,(k_n)}^{(a)}\right). \tag{5}$$

There are several features of this tail-trimmed mean that demand clarification. Let $\{c_n\}$ be any k_n/n two-tailed quantile sequence for $Z_i - \theta$:

$$\frac{n}{k_n} P\left(|Z_i - \theta| \geq c_n\right) \to 1, \tag{6}$$

and notice since Z_i has support $\mathbb{R}$ it follows $c_n \to \infty$.

First, we scale by $n - k_n$ and use the mean-centered variable $Z_i - 1/n\sum_{j=1}^n Z_j$ as the trimming criterion in order to achieve an *asymptotically unbiased estimator* when Z_i is symmetrically distributed about θ. In order to see this, we show in the appendix that for asymptotic normality it suffices to work with

$$\frac{1}{n-k_n}\sum_{i=1}^n Z_i I\left(|Z_i - \theta| < c_n\right).$$

Then, under distribution symmetry:

$$E\left[\frac{1}{n-k_n}\sum_{i=1}^{n} Z_i I\left(|Z_i-\theta|<c_n\right)\right] = \frac{n}{n-k_n}E\left[Z_i I\left(|Z_i-\theta|<c_n\right)\right] = \theta\frac{n}{n-k_n}P\left(|Z_i-\theta|<c_n\right) = \theta,$$

irrespective of how $\{k_n\}$ is chosen and therefore irrespective of whether $c_n = O(1)$ or $c_n \to \infty$.

Second, $k_n \to \infty$ and $k_n/n \to \infty$ imply trimming matters for asymptotics, but is *negligible*. The threshold $\hat{Z}^{(a)}_{n,(k_n)}$ is an intermediate order statistic hence $\hat{Z}^{(a)}_{n,(k_n)} \xrightarrow{p} \infty$ since $1/n\sum_{i=1}^{n} Z_i \xrightarrow{p} E[Z_i]$ and Z_i has support $\mathbb{R}$, (cf. Leadbetter, Lindgren, and Rootzen, 1983; Galambos, 1987). Negligibility is required if Z_i may be asymmetrically distributed. In this case if $\{k_n\}$ is a central order sequence, $k_n \sim \lambda n$ where $\lambda \in (0,1)$, then $\hat{\theta}^{(tz)}_n$ will be asymptotically biased, and our bias-correction estimator is not valid since it relies on approximating tail moments by using Karamata theory for power law tails. Further, intermediate order trimming ensures asymptotic normality for $\hat{\theta}^{(tz)}_n$ for any $\kappa > 1$. Notice trimming cannot be too light, e.g. $k_n \sim k$ a fixed constant hence $\{k_n\}$ is an extreme order sequence, since then $\hat{\theta}^{(tz)}_n$ will not belong to the domain of attraction of a normal law when the tail index $\kappa < 2$ (cf. Ibragimov and Linnik, 1971; Csörgo, Horváth, and Mason, 1986).

Third, if Z_i is known to be symmetrically distributed then a central order sequence $k_n \sim \lambda n$ may also be used. This intrinsically matters since the estimator will then be asymptotically unbiased and our bias-correction is not required. Moreover, the rate of convergence of $\hat{\theta}^{(tz)}_n$ is $n^{1/2}$ when $k_n \sim \lambda n$, but only $o(n^{1/2})$ when trimming is negligible $k_n/n \to 0$ and tails are heavy $E[Z_i^2] = \infty$. See the main results below.

We require two additional assumptions. First, probability decays according to a power law.

A3. Power Law: The observations $(Y_i, D_i, X_i)'$ are iid. $Z_i := h_i Y_i$ has a non-degenerate continuous distribution with support $\mathbb{R}$ and

$$P\left(|Z_i-\theta|>c\right) = dc^{-\kappa}(1+o(1)) \text{ where } d \in (0,\infty) \text{ and } \kappa > 1. \tag{7}$$

Remark In Section 4 we exploit a latent variable model of treatment assignment in order to show when A3 holds under limited overlap.

Remark Distribution continuity helps simplify notation, but all subsequent results can be shown to hold without continuity.

Remark The tail index κ is identically the moment supremum of Z_i (Resnick, 1987):

$$\kappa := \arg\sup\left\{\alpha > 0 : E|Z_i|^{\alpha} < \infty\right\}. \tag{8}$$

Thus $E|Z_i|^p < \infty$ *if and only if* $p < \kappa$. Since ATE $\theta = E[Z_i]$ exists when Z_i is integrable $E|Z_i| < \infty$ we assume $\kappa > 1$. All subsequent results are valid if either $\kappa > 2$ hence $E[Z_i^2] < \infty$, or the distribution

of Z_i has power law tails with tail index $\kappa \in (1,2]$. Limit theory for tail-trimmed estimators requires computation of tail-trimmed moments, which by invoking Karamata's Theorem is simplified when tail decay is of the power law class (7).

In order to handle stochastic centering $Z_i - 1/n \sum_{j=1}^n Z_j$ in the order statistic $\hat{Z}^{(a)}_{n,(k_n)}$, we need to restrict the rate of trimming $k_n \to \infty$. If Z_i were known to have a finite variance then we only need $k_n = o(n)$, but in the infinite variance case $\kappa < 2$ the mean $1/n \sum_{i=1}^n Z_i$ has a rate of convergence $n^{1-1/\kappa} < n^{1/2}$ in view of independence and power law A3 (Ibragimov and Linnik, 1971). As long as this slower rate $n^{1-1/\kappa}$ is faster than the $k_n^{1/2}$ rate of convergence for $\hat{Z}^{(a)}_{n,(k_n)}$, then $1/n \sum_{i=1}^n Z_i$ does not affect asymptotics (cf. Hill, 2014). This in turn always holds when $k_n = O(\ln(n))$.

A4. Trimming Rate: $k_n = O(\ln(n))$.

Define the tail trimmed variance and a bias term:

$$\sigma_n^2 := E\left[\left(Z_i I\left(|Z_i - \theta| \le c_n\right) - E\left[Z_i I\left(|Z_i - \theta| \le c_n\right)\right]\right)^2\right] \tag{9}$$

$$\mathcal{B}_n = \frac{n}{n - k_n} E\left[\left(Z_i - \theta\right) I\left(|Z_i - \theta| \ge c_n\right)\right].$$

Notice by $c_n \to \infty$ and dominated convergence:

$$\sigma_n^2 = E\left[Z_i^2 I\left(|Z_i - \theta| \le c_n\right)\right] - \theta^2 \times (1 + o(1)) \sim E\left[\left(Z_i - \theta\right)^2 I\left(|Z_i - \theta| \le c_n\right)\right],$$

while $E[Z_i^2 I(|Z_i - \theta| \le c_n)] \to \infty$ if $E[Z_i^2] = \infty$, hence

$$\text{if } E\left[Z_i^2\right] = \infty \text{ then } \sigma_n^2 = E\left[Z_i^2 I\left(|Z_i - \theta| \le c_n\right)\right] \times (1 + o(1)). \tag{10}$$

Unless otherwise stated, all proofs are presented in the appendix, Section 8.

Theorem 3.1 *Under A1, A3 and A4,* $\hat{\theta}_n^{(tz)} \xrightarrow{p} \theta$ *and* $n^{1/2}\sigma_n^{-1}(\hat{\theta}_n^{(tz)} + \mathcal{B}_n - \theta) \xrightarrow{d} N(0,1)$.

By dominated convergence $\mathcal{B}_n \to 0$ because $E|Z_i| < \infty$ hence $E[\hat{\theta}_n^{(tz)}] \to \theta$. As a special case bias is $\mathcal{B}_n = 0$ if Z_i has a symmetric distribution. Otherwise $(n^{1/2}/\sigma_n)\mathcal{B}_n \to 0$ provided Z_i has a finite variance ($\kappa > 2$) or a hairline infinite variance ($\kappa = 2$) and we increase the amount of trimming $k_n \to \infty$ as $n \to \infty$ sufficiently slowly such that the threshold c_n is large and therefore $\mathcal{B}_n$ is small. Finally, if Z_i has an asymmetric distribution and is heavy tailed $\kappa < 2$ then $\hat{\theta}_n^{(tz)}$ is *asymptotically biased* with respect to its limit distribution.

Theorem 3.2 *Let A1, A3 and A4 hold. If* Z_i *has a symmetric distribution, or* $\kappa \ge 2$ *with* $k_n \to \infty$ *and* $k_n/\ln(n) \to 0$, *then* $(n^{1/2}/\sigma_n)(\hat{\theta}_n^{(tz)} - \theta) \xrightarrow{d} N(0,1)$. *Otherwise, if* Z_i *has an asymmetric distribution and* $\kappa < 2$ *then* $(n^{1/2}/\sigma_n)\mathcal{B}_n \to \infty$ *for any intermediate order* $\{k_n\}$.

The rate of convergence $n^{1/2}/\sigma_n$ when Z_i has an infinite variance requires a characterization of the tail-trimmed variance σ_n^2. Under A3 and by definition (6) we can always choose the thresholds to be

$$c_n = d^{1/\kappa} (n/k_n)^{1/\kappa}. \tag{11}$$

Under power law A3 and variance property (10), we need only invoke Karamata's Theorem to deduce (see Theorem 0.6 in Resnick (1987))[9]

$$\text{if } \kappa = 2 \text{ then } E\left[(Z_i-\theta)^2 I(|Z_i-\theta| \le c_n)\right] \sim d\ln(n) \tag{12}$$

$$\text{if } \kappa < 2 \text{ then } E\left[(Z_i-\theta)^2 I(|Z_i-\theta| \le c_n)\right] \sim \frac{\kappa}{2-\kappa} c_n^2 P(|Z_i-\theta| > c_n) = \frac{\kappa}{2-\kappa} d^{2/\kappa} \left(\frac{n}{k_n}\right)^{2/\kappa-1}.$$

Combine Theorem 3.1 with (10) and (12) to obtain the following.

Corollary 3.3 *Let A1, A3 and A4 hold. If $\kappa > 2$ then $n^{1/2}(\hat{\theta}_n^{(tz)} + \mathcal{B}_n - \theta) \xrightarrow{d} N(0, E[(Z_i-\theta)^2])$; if $\kappa = 2$ then $(n/\ln(n))^{1/2}(\hat{\theta}_n^{(tz)} + \mathcal{B}_n - \theta) \xrightarrow{d} N(0,d)$; and if $\kappa \in (1,2)$ then*

$$\frac{n^{1/2}}{(n/k_n)^{1/\kappa-1/2}} \left(\hat{\theta}_n^{(tz)} + \mathcal{B}_n - \theta\right) \xrightarrow{d} N\left(0, \frac{\kappa}{2-\kappa} d^{2/\kappa}\right).$$

Remark Tail trimming has no impact on efficiency if $E[Z_i^2] < \infty$. However, when $\kappa < 2$ the rate of convergence can be increased by increasing the rate of trimming $k_n \to \infty$: the removal of more sample examples when $\kappa < 2$ improves sharpness in this sense.

3.2 First Order Asymptotics: MSE Minimization

Let $\theta = 0$ to reduce notation. It can be shown that $\hat{\theta}_n^{(tz)}$ is asymptotically a linear function of $Z_i I(|Z_i| \le c_n)$ and $\mathcal{B}_n$ by a first order asymptotic approximation:

$$\begin{aligned} \hat{\theta}_n^{(tz)} - \theta &= \frac{1}{n-k_n} \sum_{i=1}^{n} \{Z_i I(|Z_i| \le c_n) - E[Z_i I(|Z_i| \le c_n)]\} \\ &\quad - \frac{n}{n-k_n} E[Z_i I(|Z_i| > c_n)] + o_p(1) \\ &= \mathcal{Z}_n - \mathcal{B}_n + o_p(1), \end{aligned} \tag{13}$$

[9] If $\kappa = 2$ then for any finite $a > 0$ and some $K(a) > 0$ we have $E[Z_i^2 I(|Z_i| \le c_n)] = K(a) + \int_a^{c_n^2} P(|Z_i| \le u^{1/2}) du$ $\sim K(a) + d\int_a^{c_n^2} u^{-1} du = K(a) + d(\ln(c_n^2) - \ln(a))$. Now use $c_n^2 = d(n/k_n)$ and $k_n = o(n)$ to deduce $E[Z_i^2 I(|Z_i| \le c_n)]$ $\sim d(\ln(n)$.

say. See the proof of Theorem 3.2 and see Appendix A of Hill (2013). By construction:

$$E\left(\mathcal{Z}_n-\mathcal{B}_n\right)^2=\left(\frac{n}{n-k_n}\right)^2\frac{\sigma_n^2}{n}+\mathcal{B}_n^2=\left(\frac{n}{n-k_n}\right)^2\left(\frac{\sigma_n^2}{n}+\left(E\left[Z_iI\left(|Z_i|>c_n\right)\right]\right)^2\right).$$

We need to approximate σ_n^2 and $E[(Z_i-\theta)I(|Z_i-\theta|>c_n)]$ in order to see how trimming impacts the mean-squared-error.

A characterization of bias $\mathcal{B}_n$ requires a sharpened power law assumption A3:

$$P\left(Z_i\leq -c\right)\sim d_1c^{-\kappa}\text{ and }P\left(Z_i\geq c\right)\sim d_2c^{-\kappa}\text{ as }c\to\infty\text{ where }\kappa>1\text{ and }d_1,d_2\in(0,\infty).\qquad(14)$$

This diminishes generality since conceivably the tails may have different indices, e.g. $d_1c^{-\kappa_1}$ and $d_2c^{-\kappa_2}$ for $\kappa_1\gtreqless\kappa_2$. We aim to analyze bias in a simple context for the sake of brevity, while all that follows can be easily extended to the general case $\kappa_1\gtreqless\kappa_2$. Define the two-tailed Paretian tail scale

$$d=d_1+d_2,$$

and define left and right tail quantile functions:

$$Q_1(u):=\inf\left\{c\geq 0:P\left(Z_i\leq -c\right)\geq u\right\}\text{ and }Q_2(u):=\inf\left\{c\geq 0:P\left(Z_i\geq c\right)\geq u\right\},$$

where $0\leq u\leq 1$. Under power law A3 notice $Q_i(u)=d_i^{1/\kappa}u^{-1/\kappa}$ as $u\to 0$. Use threshold construction (11) to deduce bias under power law A3 is

$$\begin{aligned}
\mathcal{B}_n &= \frac{n}{n-k_n}E\left[\left(Z_i-\theta\right)I\left(|Z_i-\theta|>c_n\right)\right]=\frac{n}{n-k_n}\left(\int_0^{k_n/n}Q_2(u)du-\int_0^{k_n/n}Q_1(u)du\right)\qquad(15)\\
&\sim \frac{n}{n-k_n}\int_0^{k_n/n}d_2^{1/\kappa}u^{-1/\kappa}du-\frac{n}{n-k_n}\int_0^{k_n/n}d_1^{1/\kappa}u^{-1/\kappa}du\\
&= d_2^{1/\kappa}\left(\frac{\kappa}{\kappa-1}\right)\frac{n}{n-k_n}\left(\frac{k_n}{n}\right)^{1-1/\kappa}-d_1^{1/\kappa}\left(\frac{\kappa}{\kappa-1}\right)\frac{n}{n-k_n}\left(\frac{k_n}{n}\right)^{1-1/\kappa}\\
&= \frac{n}{n-k_n}\times\frac{\left(d_2^{1/\kappa}-d_1^{1/\kappa}\right)}{d^{1/\kappa}}\left(\frac{\kappa}{\kappa-1}\right)\left(\frac{k_n}{n}\right)c_n\\
&= \frac{n}{n-k_n}\times\left(d_2^{1/\kappa}-d_1^{1/\kappa}\right)\left(\frac{\kappa}{\kappa-1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa}=:\frac{n}{n-k_n}B_n.
\end{aligned}$$

The variance σ_n^2 depends on the tail index in view of Karamata theory (12) when $\kappa\leq 2$, and an argument similar to (15) when $\kappa>2$.

Lemma 3.4 *Under power law A3:*

$$\kappa > 2 \,:\, \sigma_n^2 \sim E\left[(Z_i-\theta)^2\right] - d^{1/\kappa}\left(\frac{\kappa}{\kappa-2}\right)\left(\frac{k_n}{n}\right)^{1-2/\kappa} := s_n^2$$
$$\kappa = 2 \,:\, \sigma_n^2 \sim d\ln(n) := s_n^2$$
$$\kappa < 2 \,:\, \sigma_n^2 \sim \frac{\kappa}{2-\kappa} d^{2/\kappa}\left(\frac{n}{n-k_n}\right)^2\left(\frac{n}{k_n}\right)^{2/\kappa-1} := s_n^2.$$

Define the first order asymptotic mse as

$$\text{MSE}\left(\hat{\theta}_n^{(tz)}\right) = \left(\frac{n}{n-k_n}\right)^2\left(\frac{1}{n}s_n^2 + B_n^2\right).$$

The outer $n/(n-k_n)$ arises from the scale correction to ensure asymptotic unbiasedness when Z_i has a symmetric distribution, and smaller k_n implies a smaller $n/(n-k_n)$. Bias $|B_n|$ is monotonically increasing in k_n. The variance s_n^2 is monotonically decreasing in k_n when $\kappa \neq 2$. The case $\kappa = 2$ is therefore the simplest to handle: $\text{MSE}\left(\hat{\theta}_n^{(tz)}\right)$ is monotonically smaller for smaller k_n.

Otherwise, if $\kappa \neq 2$ then there is a trade-off: more trimming increases $|B_n|$, but dampens the noise effect of extreme values in sample means and therefore decreases s_n^2 and increases the rate of convergence. If $\kappa < 2$ then B_n^2 dominates because $B_n^2 \propto (k_n/n)^{2(1-1/\kappa)} = (k_n/n)(n/k_n)^{2/\kappa-1} > (1/n)(n/k_n)^{2/\kappa-1} \propto s_n^2/n$, hence we should trim less in order to reduce the mean-squared-error as n increases. Conversely, with the A4 property $k_n = O(\ln(n))$ if $\kappa > 2$ then dispersion s_n^2/n dominates since $B_n^2 \propto (k_n/n)^{2(1-1/\kappa)} < 1/n \propto s_n^2/n$ for large enough n, hence trimming should be higher. This proves the following.

Theorem 3.5 *Under A3 and A4 the first order asymptotic mse of $\hat{\theta}_n^{(tz)}$ is monotonically increasing in k_n if $\kappa \leq 2$ and monotonically decreasing in k_n if $\kappa > 2$.*

This reveals three challenges. First, choosing k_n based on the first order asymptotic mse requires knowledge of κ. In the latent variable model this implies at least knowledge of the distributions of $\{U, X\}$. Second, if $\kappa > 2$ were known then we would not need tail trimming for asymptotic Gaussian inference. Third, if $\kappa > 2$ were known and a large k_n where chosen to reduce the mse, then first order asymptotic bias and small sample bias would be elevated (small sample bias is discussed below), which can lead to poor inference. If accurate asymptotic inference is desired, then minimizing first order bias is clearly better than minimizing the mse, hence we should trim less. In order to ensure A4 $k_n = O(\ln(n))$ holds we may therefore use, for example:

$$k_n \sim \lambda\left(\ln(n)\right)^a \text{ for } \lambda > 0 \text{ and } a \in (0, 1]. \tag{16}$$

3.3 Bias-Corrected Tail-Trimmed Estimation

By Theorem 3.2, there may be asymptotic bias in the limit distribution of $\hat{\theta}_n^{(tz)}$ if Z_i is not symmetrically distributed about zero and $\kappa \in (1,2)$. In this subsection we characterize and estimate bias for a new ATE estimator. We then provide consistent estimators of the asymptotic variances of our estimators.

3.3.1 Bias-Correction Estimator

Consider the simple Paretian tail form (14) to simplify arguments. See Section 3.4 below for discussion on the case where tail decay is faster than a power law, a case that may arise under strict overlap A2.

Bias $\mathcal{B}_n$ characterized in (15) is easily estimated since it is only a function of tail exponents and the fractile k_n. Define

$$\hat{Z}_{n,i} := Z_i - \frac{1}{n}\sum_{j=1}^{n} Z_j \text{ and } \hat{Z}_{n,i}^{(a)} := \left|\hat{Z}_{n,i}\right| \text{ and } \hat{Z}_{n,i}^{(-)} := -\hat{Z}_{n,i} I\left(\hat{Z}_{n,i} < 0\right) \text{ and } \hat{Z}_{n,i}^{(+)} := \hat{Z}_{n,i} I(\hat{Z}_{n,i} > 0).$$

Denote the order statistics $\hat{Z}_{n,(1)}^{(\cdot)} \geq \hat{Z}_{n,(2)}^{(\cdot)} \geq \cdots \geq \hat{Z}_{n,(n)}^{(\cdot)}$. Let $\{m_n\}$ be an intermediate order sequence: $1 \leq m_n < n$, $m_n \to \infty$ and $m_n = o(n)$. The threshold c_n is estimated with $\hat{Z}_{n,(k_n)}^{(a)}$, and we estimate κ with Hill (1975)'s seminal tail index estimator:[10]

$$\hat{\kappa}_{m_n}^{-1} = \frac{1}{m_n - 1}\sum_{j=1}^{m_n - 1} \ln\left(\hat{Z}_{n,(i)}^{(a)} / \hat{Z}_{n,(m_n)}^{(a)}\right).$$

Hall (1982) proposes estimators of the scales d_1 , d_2, and d:

$$\hat{d}_{m_n,1} := \frac{m_n}{n}\left(\hat{Z}_{n,(m_n)}^{(-)}\right)^{\hat{\kappa}_{m_n}}, \quad \frac{m_n}{n}\left(\hat{Z}_{n,(m_n)}^{(+)}\right)^{\hat{\kappa}_{m_n}} \text{ and } \hat{d}_{m_n} := \frac{m_n}{n}\left(\hat{Z}_{n,(m_n)}^{(a)}\right)^{\hat{\kappa}_{m_n}}.$$

A sharper tail estimator $\hat{\kappa}_{m_n}$ arises when many tail observations are used because $\hat{\kappa}_{m_n} = \kappa + O(1/m_n^{1/2})$ provided a second order tail decay property holds (cf. Hall (1982)). Further, if more observations are used for tail component estimation in the sense

$$m_n / k_n \to \infty$$

then the tail estimators do not affect the limit distribution of our ATE estimator hence asymptotics

[10] Many alternative estimators of κ are available: see Hill (2010) for references.

are greatly simplified. See Hill (2013). Therefore, as in Hill (2013), we estimate bias with[11]

$$\begin{aligned}\hat{\mathcal{B}}_n &= \frac{n}{n-k_n}\left(\frac{\hat{d}^{1/\hat{\kappa}_{m_n}}_{m_n,2}-\hat{d}^{1/\hat{\kappa}_{m_n}}_{m_n,1}}{\hat{d}^{1/\hat{\kappa}_{m_n}}_{m_n}}\right)\frac{\hat{\kappa}_{m_n}}{\hat{\kappa}_{m_n}-1}\frac{k_n}{n}\hat{Z}^{(a)}_{n,(k_n)} \\ &= \frac{n}{n-k_n}\left(\frac{\hat{Z}^{(+)}_{n,(m_n)}-\hat{Z}^{(-)}_{n,(m_n)}}{\hat{Z}^{(a)}_{n,(m_n)}}\right)\frac{\hat{\kappa}_{m_n}}{\hat{\kappa}_{m_n}-1}\frac{k_n}{n}\hat{Z}^{(a)}_{n,(k_n)}\end{aligned} \tag{17}$$

and the bias-corrected tail-trimmed ATE estimator is therefore

$$\hat{\theta}^{(tz:bc)}_n = \hat{\theta}^{(tz)}_n + \hat{\mathcal{B}}_n.$$

3.3.2 Optimal Bias-Correction

A shortcoming of $\hat{\theta}^{(tz:bc)}_n$ is it uses one fractile m_n for tail exponent estimation. In practice, however, $\hat{\mathcal{B}}_n$ is well defined only when $\hat{\kappa}_{m_n} > 1$,[12] and when m_n is not greater than the number of negative or positive $Z_i's$. [13] Further, it seems desirable to choose m_n such that $\hat{\theta}^{(tz:bc)}_n$ is close to an unbiased estimator, cf. Hill (2013). Consider $m_n(\phi) = [\phi m_n]$ where $\phi \in \Phi^* = [\underline{\phi}, \bar{\phi}]$ for some chosen $0 < \underline{\phi} < \bar{\phi}$, and let $\hat{\mathcal{B}}_n(\phi)$ be the bias estimator computed with $m_n(\phi)$. The new bias-corrected estimator is then

$$\hat{\theta}^{(tz:bc*)}_n = \hat{\theta}^{(tz:bc)}_n(\phi^*_n) := \frac{1}{n-k_n}\sum_{i=1}^n Z_i I\left(\left|Z_i - \frac{1}{n}\sum_{j=1}^n Z_j\right| < \hat{Z}^{(a)}_{n,(k_n)}\right) + \hat{\mathcal{B}}_n(\phi^*_n)$$

where

$$\phi^*_n = \underset{\phi\in\Phi^*}{\arg\min}\left|\hat{\theta}^{(tz:bc)}_n(\phi^*_n) - \frac{1}{n}\sum_{i=1}^n Z_i\right|$$

and

$$\Phi^* = \left\{\phi \in \left[\underline{\phi}, \bar{\phi}\right] : \hat{\kappa}_{m_n(\phi)} > 1 \text{ and } m_n(\phi) > \min\left\{\sum_{i=1}^n I\left(Z_i < 0\right), \sum_{i=1}^n I\left(Z_i > 0\right)\right\}\right\}.$$

Notice $\hat{\theta}^{(tz:bc)}_n$ merely fixes $\phi = 1$. In view of the multiplicative form $m_n(\phi) = [\phi m_n]$ with $\phi > 0$, and $m_n/k_n \to \infty$, the estimator $\hat{\theta}^{(tz:bc^*)}_n$ has the same limit distribution as $\hat{\theta}^{(tz:bc)}_n$ (see Hill, 2013, Theorem 2.2).

Sampling error can render $\hat{\theta}^{(tz:bc^*)}_n$ farther from the untrimmed $1/n\sum_{i=1}^n Z_i$ than the non-bias-

[11] In principle different order sequences $\{m_n, m_{1,n}, m_{2,n}\}$ can used to estimate d, d_1, and d_2, but in practice there will not be a convenient way to determine all four sequences $\{k_n, m_n, m_{1,n}, m_{2,n}\}$. For practical simplicity we therefore use one sequence $\{m_n\}$ for all tail estimators. Our simulations suggest this does not hinder the performance of our estimator.

[12] The problem of selecting the tail fractile m_n in extreme value theory is well known. If only $\hat{\kappa}_{m_n}$ is desired then minimum mean-squared-error, plotting and regression methods exist when the data are iid. See Huisman, Koedijk, Kool, and Palm (2001) and Hill (2010) for references.

[13] If, for example, m_n is greater than the number of negative $Z_i's$ then $Z^{(-)}_{(m_n)} = 0$ which does not provide useful information about the left tail.

corrected $\hat{\theta}_n^{(tz)}$. In practice we use whichever estimator is closest to the unbiased estimator:

$$\begin{aligned}\hat{\theta}_n^{(tz:o)} \quad &: = \hat{\theta}_n^{(tz:bc^*)} I\left(\left|\hat{\theta}_n^{(tz:bc^*)} - \frac{1}{n}\sum_{i=1}^n Z_i\right| < \left|\hat{\theta}_n^{(tz)} - \frac{1}{n}\sum_{i=1}^n Z_i\right|\right) \\ &\quad + \hat{\theta}_n^{(tz)} I\left(\left|\hat{\theta}_n^{(tz:bc^*)} - \frac{1}{n}\sum_{i=1}^n Z_i\right| \geq \left|\hat{\theta}_n^{(tz)} - \frac{1}{n}\sum_{i=1}^n Z_i\right|\right).\end{aligned}$$

As long as $\hat{\theta}_n^{(tz)}$ is biased asymptotically in its limit distribution, then $\hat{\theta}_n^{(tz:bc^*)}$ will be chosen with probability approaching one (cf. Hill and Prokhorov, 2014, Proof of Theorem 7.2).

3.3.3 Large Sample Properties

In order to ensure $\hat{\kappa}_{m_n}$ is $m_n^{1/2}$-convergent we must impose a second order tail decay property and restrict the rate of increase of the fractiles m_n. The following second order power law property is a popular way to do this (Hall, 1982), but other tail forms are also viable. See Haeusler and Teugels (1985) and Goldie and Smith (1987), and see Hsing (1991) and Hill (2010).

A3′. Second Order Power Law: The observations $(Y_i, D_i, X_i)'$ are iid. $Z_i := h_i Y_i$ has a non-degenerate continuous distribution with support $\mathbb{R}$ and a power law distribution tail: for some $d_1, d_2 > 0$ and $\kappa > 1$

$$P\left(Z_i - \theta < -c\right) = d_1 c^{-\kappa}\left(1 + O(c^{-\xi})\right) \quad \text{and} \quad P\left(Z_i - \theta > c\right) = d_2 c^{-\kappa}\left(1 + O(c^{-\xi})\right). \tag{18}$$

Further, $m_n \to \infty$, $m_n = o(n^{2\xi/(2\xi+\kappa)})$ and $m_n/k_n \to \infty$.

Remark The fractile bound $m_n = o(n^{2\xi/(2\xi+\kappa)})$ reflects the need to use observations strictly from the tails of the distribution when Z_i deviates from a Pareto law. In the Pareto case $\xi = \infty$ hence we need only bound $m_n = o(n)$.

Remark In practice ξ and κ are not be known. The A3′ and A4 requirements $m_n = o(n^{2\xi/(2\xi+\kappa)})$, $m_n/k_n \to \infty$, and $k_n = O(\ln(n))$ are, however, always satisfied when $k_n = [\lambda_k(\ln(n))^{1-2\iota}]$ and $m_n = [\lambda_m(\ln(n))^{1+a-\iota}]$ for tiny $\iota > 0$ and some $a \geq 0$ and $\lambda_k, \lambda_m > 0$.

The limit distribution of $\hat{\theta}_n^{(tz:bc)}$ is based on the joint limiting properties of the non-tail component $Z_i I(|Z_i - \theta| < c_n)$ and tail components that govern $\hat{Z}_{n,(k_n)}^{(\cdot)}$ and $\hat{\kappa}_{m_n}$. The key underlying process governing Hill (1975) and Hall (1982) estimators is the tail indicator $I(|Z_i - \theta| > c_n)$, cf. Hsing (1991, Section2). Hence stack non-tail and tail variables and construct their covariance matrix:

$$\mathcal{W}_{n,i} := \begin{bmatrix} Z_i I\left(|Z_i - \theta| \leq c_n\right) - E\left[Z_i I\left(|Z_i - \theta| \leq c_n\right)\right] \\ \left(\frac{n}{k_n}\right)^{1/2} \left\{I\left(|Z_i - \theta| > c_n\right) - P\left(|Z_i - \theta| > c_n\right)\right\} \end{bmatrix}$$

$$\Sigma_n := E\left[\mathcal{W}_{n,i}\mathcal{W}_{n,i}'\right] \text{ and } \mathcal{V}_n^2 := \mathcal{K}_n'\Sigma_n\mathcal{K}_n \text{ where } \mathcal{K}_n := \left[\begin{array}{cc} 1, & -\frac{1}{\kappa-1}\left(\frac{k_n}{n}\right)^{1/2} c_n \end{array}\right]$$

$$\mathfrak{V}_n^2 := \mathcal{V}_n^2 I_{n,i}^* + \sigma_n^2\left(1 - I_{n,i}^*\right) \text{ where } I_{n,t}^* := I\left(\left|\hat{\theta}_n^{(tz:bc^*)} - \frac{1}{n}\sum_{i=1}^n Z_i\right| < \left|\hat{\theta}_n^{(tz)} - \frac{1}{n}\sum_{i=1}^n Z_i\right|\right).$$

Theorem 3.6 *Under A1, A3′ and A4 we have* $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz:bc)} - \theta) \xrightarrow{d} N(0,1)$ *and* $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz:bc^*)} - \theta) \xrightarrow{d} N(0,1)$ *where* $\mathcal{V}_n^2 = \sigma_n^2(1+o(1))$ *if* $\kappa < 2$ *and* $\mathcal{V}_n^2 = \sigma_n^2(1+O(1))$ *if* $\kappa \geq 2$. *Furthermore* $n^{1/2}\mathfrak{V}_n^{-1}(\hat{\theta}_n^{(tz:o)} - \theta) \xrightarrow{d} N(0,1)$ *where* $\mathfrak{V}_n^2 = \mathcal{V}_n^2 + o_p(1)$ *if* $\kappa < 2$ *and* $\mathfrak{V}_n^2 = \sigma_n^2 + o_p(1)$ *if* $\kappa \geq 2$.

Remark The order of $\mathcal{V}_n^2$ is the same as that of σ_n^2 defined in (9) since the tail component of $\mathcal{W}_{n,i}$ is dominated by the tail-trimmed component.

Remark $\hat{\theta}_n^{(tz:o)} = \hat{\theta}_n^{(tz:bc^*)}$ hence $\mathfrak{V}_n^2 = \mathcal{V}_n^2 + o_p(1)$ when $\kappa < 2$, because $\hat{\theta}_n^{(tz)}$ is then asymptotically biased in its limit distribution $(n^{1/2}/\sigma_n)|\mathcal{B}_n| \to \infty$. Conversely, $\hat{\theta}_n^{(tz:o)} = \hat{\theta}_n^{(tz)} + o_p(1)$ hence $\mathfrak{V}_n^2 = \sigma_n^2 + o_p(1)$ when $\kappa \geq 2$, because $\hat{\theta}_n^{(tz)}$ is asymptotically unbiased $(n^{1/2}/\sigma_n)|\mathcal{B}_n| \to 0$.

3.3.4 Inference

Consistent estimators of the scale σ_n^2 for $\hat{\theta}_n^{(tz)}$ and $\mathcal{V}_n^2$ for $\{\hat{\theta}_n^{(tz:bc)}, \hat{\theta}_n^{(tz:bc^*)}, \hat{\theta}_n^{(tz:o)}\}$ are easily constructed. Recall $\hat{Z}_{n,i} := Z_i - 1/n\sum_{j=1}^n Z_j$, define components

$$\hat{\mathcal{K}}_n := \left[\begin{array}{cc} 1, & -\frac{1}{\kappa-1}\left(\frac{k_n}{n}\right)^{1/2} \hat{Z}_{n,(k_n)}^{(a)} \end{array}\right]$$

$$\widehat{\mathcal{W}}_{n,i} := \left[\begin{array}{c} Z_i I\left(\left|Z_i - \frac{1}{n}\sum_{j=1}^n Z_j\right| < \hat{Z}_{n,(k_n)}^{(a)}\right) - \frac{1}{n}\sum_{i=1}^n Z_i I\left(\left|Z_i - \frac{1}{n}\sum_{j=1}^n Z_j\right| < \hat{Z}_{n,(k_n)}^{(a)}\right) \\ \left(\frac{n}{k_n}\right)^{1/2}\left\{I\left(\left|Z_i - \frac{1}{n}\sum_{j=1}^n Z_j\right| < \hat{Z}_{n,(k_n)}^{(a)}\right) - \frac{k_n}{n}\right\} \end{array}\right],$$

and define scales $\hat{\Sigma}_n := 1/n\sum_{i=1}^n \widehat{\mathcal{W}}_{n,i}\widehat{\mathcal{W}}_{n,i}'$ and:

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n\left\{Z_i I\left(\left|Z_i - \frac{1}{n}\sum_{j=1}^n Z_j\right| < \hat{Z}_{n,(k_n)}^{(a)}\right) - \frac{1}{n}\sum_{i=1}^n Z_i\right\}^2$$

$$\hat{\mathcal{V}}_n^2 := \hat{\mathcal{K}}_n'\Sigma_n\hat{\mathcal{K}}_n$$

$$\hat{\mathfrak{V}}_n^2 := \hat{\mathcal{V}}_n^2 I_{n,i}^* + \hat{\sigma}_n^2\left(1 - I_{n,i}^*\right) \text{ where } I_{n,t}^* := I\left(\left|\hat{\theta}_n^{(tz:bc^*)} - \frac{1}{n}\sum_{i=1}^n Z_i\right| < \left|\hat{\theta}_n^{(tz)} - \frac{1}{n}\sum_{i=1}^n Z_i\right|\right).$$

Theorem 3.7 *Under the conditions of Theorem 3.6* $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$, $\hat{\mathcal{V}}_n^2/\mathcal{V}_n^2 \xrightarrow{p} 1$ *and* $\hat{\mathfrak{V}}_n^2/\mathfrak{V}_n^2 \xrightarrow{p} 1$.

3.4 Bias-Correction: Thin Tails

In practice it may not be obvious that limited overlap is occurring, or if it is that the probability tails of Z_i decay according to a power law. In Section 4 we show that limited overlap can lead to power law tail decay in Z_i in a latent variable framework for treatment assignment, when the treatment outcome Y_j and covariates X have distribution tails that decay exponentially fast. In principle, however, Z_i may have very thin tails. For example, limited overlap A2′ nests strict overlap A2, and under the latter Z_i may have exponential tails or simply a bounded support.

We now characterize the large sample properties of the bias-corrected estimator $\hat{\theta}_n^{(tz:bc)}$ when (18) fails to hold. As a convenient benchmark we assume Z_i has an exponential distribution tail

$$P\left(|Z_i - \theta| \geq c\right) = \xi \exp\left\{-\varpi c^{\delta}\right\} \text{ where } \xi, \varpi, \delta > 0. \tag{19}$$

The following extends easily to the case of a bounded support.

Clearly (19) implies by dominated convergence $\lim_{n\to\infty} \sigma_n^2 = E[(Z_i - \theta)^2] < \infty$. In this case the expression for the bias term in (15) is meaningless since it is based on Paretian tail parameters, hence the limiting properties of $\hat{\mathcal{B}}_n$ are yet unknown. Since $E[Z_i^2] < \infty$, then for $k_n \to \infty$ slow enough we need only show $n^{1/2}\hat{\mathcal{B}}_n \xrightarrow{p} 0$ in order to verify by Theorem 3.2

$$n^{1/2}\left(\hat{\theta}_n^{(tz:bc)} - \theta\right) = n^{1/2}\left(\hat{\theta}_n^{(tz)} - \theta\right) + o_p(1) \xrightarrow{d} N\left(0, E\left[(Z_i - \theta)^2\right]\right).$$

Intuitively, since $|Z_i|$ has exponential tails it has all moments and therefore its moment supremum $\arg\sup\{\alpha > 0 : E|Z_i|^{\alpha} < \infty\} = \infty$. This suggests Hill (1975)'s estimator should simply diverge in probability hence $\hat{\kappa}_{m_n}/(\hat{\kappa}_{m_n} - 1) \xrightarrow{p} 1$. Further, limit theory for order statistics covers a broad range of tail decay cases. The next result shows when $n^{1/2}\hat{\mathcal{B}}_n \xrightarrow{p} 0$. The proof is similar to arguments in Hill (2013, Theorem 2.3), and therefore relegated to the supplemental material Chaudhuri and Hill (2013).

Theorem 3.8 *Let Z_i have exponential tails (19), and assume $k_n = O(\ln(n))$, $m_n \to \infty$ and $m_n = o(n)$. Then $n^{1/2}\hat{\mathcal{B}}_n \xrightarrow{p} 0$ hence $n^{1/2}(\hat{\theta}_n^{(tz:bc)} - \theta) \xrightarrow{d} N(0, E[(Z_i - \theta)^2])$.*

Remark Notice we no longer restrict $m_n = o(n)$ nor require $m_n/k_n \to \infty$ since these merely expedite consistency of tail exponent estimators and a simple asymptotic variance formula under power law (18). We still require $k_n = O(\ln(n))$ in view of the plug-in $1/n\sum_{i=1}^n Z_i$ used for mean-centering in the trimming indicator.

Remark In general the tail decay of Z_i will be unknown. Exactly as discussed in Remark 3.3.3 following A3′, it is always viable to use $k_n = [\lambda_k(\ln(n))^{1-2\iota}]$ and $m_n = [\lambda_m(\ln(n))^{1+a-\iota}]$ for tiny $\iota > 0$ and some $a \geq 0$ and $\lambda_k, \lambda_m > 0$.

3.5 Higher Order Asymptotics and Fractile Choice

If Section 3.2 we show the non-bias-corrected estimator $\hat{\theta}_n^{(tz)}$ is easily and entirely characterized by first order asymptotics. The bias-corrected estimator $\hat{\theta}_n^{(tz:bc)} = \hat{\theta}_n^{(tz)} + \hat{\mathcal{B}}_n$, however, exhibits bias due to tail-trimming in $\hat{\theta}_n^{(tz)}$ *and* due to the bias estimator $\hat{\mathcal{B}}_n$. The latter arises because $\hat{Z}_{n,(k_n)}^{(a)}$, $\hat{Z}_{n,(m_n)}^{(+)}$ and $\hat{\kappa}_{m_n}$ are first order equivalent to sums of tail indicators, and the expected values of these tail indicators are tail probabilities with higher order properties.

We now derive a higher order bias characterization. Higher order bias follows from a higher order expansion for $\hat{\theta}_n^{(tz:bc)} - \theta$ under a second order power law assumption A3″, below. Because the expansion is quite complicated we do not present a corresponding mean-squared-error expression. Further, since mean-centering $Z_i - 1/n \sum_{j=1}^n Z_j$ in the trimming indicator does not affect our final conclusion about how k_n impacts bias, without reducing generality we assume $\theta = 0$ and therefore expand the estimator:

$$\hat{\theta}_n^{(tz:bc)} = \hat{\theta}_n^{(tz)} + \hat{\mathcal{B}}_n = \frac{1}{n}\sum_{i=1}^n Z_i I\left(|Z_i| < Z_{(k_n)}^{(a)}\right) + \frac{n}{n-k_n}\left(\frac{Z_{(m_n)}^{(+)} - Z_{(m_n)}^{(-)}}{Z_{(m_n)}^{(a)}}\right)\frac{\hat{\kappa}_{m_n}}{\hat{\kappa}_{m_n} - 1}\frac{k_n}{n}Z_{(k_n)}^{(a)}.$$

In order to glean the most information possible, but without reducing too much generality, we sharpen power law A3′ by being precise about the second order term, and bounding $m_n \to \infty$ and $k_n \to \infty$ in order to deduce which bias terms dominate.

A3″. Second Order Power Law: The observations $(Y_i, D_i, X_i)'$ are iid. $Z_i := h_i Y_i$ has a non-degenerate continuous distribution with support $\mathbb{R}$ and a power law distribution tail: for some $d_1, d_2 > 0$ and $\kappa > 1$

$$P\left(Z_i < -c\right) = d_1 c^{-\kappa}\left(1 + \varpi c^{-\xi}\right) \quad \text{and} \quad P\left(Z_i > c\right) = d_2 c^{-\kappa}\left(1 + \varpi c^{-\xi}\right) \tag{20}$$

Further, $m_n \to \infty$, $m_n = O(\ln(n))$, $k_n = o(m_n)$ and $k_n / m_n^{1/2} \to \infty$.

Remark A3″ implies A4 since $m_n = O(\ln(n))$ and $k_n = o(m_n)$ imply $k_n = o(\ln(n))$.

The higher order expansion and bias are complicated by the numerous terms that arise from the bias estimator itself, and the second order power law. We therefore only give the following conclusions here and save the technical details for the supplemental appendix Chaudhuri and Hill (2013). We provide there an expansion of $\hat{\theta}_n^{(tz:bc)} - \theta$ and bias expression. The higher order expansion is based on expanding the following first order expansion.

Lemma 3.9 *Under A3″* $\hat{\theta}_n^{(tz:bc)} - \theta = \mathfrak{B}_n + o_p(\mathcal{V}_n / n^{1/2})$ *for some random variable* $\mathfrak{B}_n$.

The term $o_p(\mathcal{V}_n / n^{1/2})$ arises solely from first order asymptotics by substituting c_n for $\hat{Z}_{n,(k_n)}^{(a)}$ in the trimming indicator. By Karamata theory $\mathcal{V}_n$ is unaffected by k_n when $\kappa = 2$ or is smaller when

k_n is larger when $\kappa \neq 2$. Since $o_p(\mathcal{V}_n/n^{1/2})$ merely captures the first order efficiency result for tail-trimmed sums with or without bias-correction (see, e.g., Hill, 2013), and since it can be verified that $(n^{1/2}/\mathcal{V}_n)|E[\mathfrak{B}_n]| \to \infty$ when $\kappa < 2$ such that $\mathfrak{B}_n$ dominates any bias discussion, we drop $o_p(\mathcal{V}_n/n^{1/2})$ and define higher order bias as

$$\mathcal{B}_n(\hat{\theta}_n^{(tz:bc)}) := E\left[\mathfrak{B}_n\right].$$

When $m_n = O(\ln(n))$, $k_n = o(m_n)$ and $k_n/m_n^{1/2} \to \infty$, bias is dominated by few terms as n increases. There are four possible cases. In the following $o(1)$ and $O(1)$ terms are functions of n. Write $\mathcal{B}_n = \mathcal{B}_n(\hat{\theta}_n^{(tz:bc)})$.

Theorem 3.10 *Let A3″ hold. Then* $\mathcal{B}_n = o(\mathcal{V}_n/n^{1/2})$ *if* $\kappa \geq 2$ *and otherwise* $(n^{1/2}/\mathcal{V}_n)|\mathcal{B}_n| \to \infty$. *Furthermore:*

(a). If tails are symmetric $d_1 = d_2$ *and* Pareto $\xi = 0$ *then* $\mathcal{B}_n(\hat{\theta}_n^{(tz:bc)}) \sim (k_n/n)^{1-1/\kappa} m_n^{-1} d^{1/\kappa} \kappa(\kappa - 1)^{-1} \times O(1)$.

(b). If distribution tails are symmetric $d_1 = d_2$ *and* not exactly Pareto $\xi < \infty$ *then*

$$\mathcal{B}_n\left(\hat{\theta}_n^{(tz:bc)}\right) \sim \left(\frac{k_n}{n}\right)^{1-1/\kappa} \frac{1}{m_n} d^{1/\kappa} \left\{ \frac{\kappa}{\kappa - 1} + \frac{\kappa}{(\kappa-1)^2} \frac{\xi}{\kappa + \xi} \frac{\varpi\,(m_n/n)^{\xi/\kappa}}{d^{\xi/\kappa} + \varpi\,(m_n/n)^{\xi/\kappa}} \right\} \times O(1).$$

(c). If tails are asymmetric $d_1 \neq d_2$ *and* Pareto $\xi = \infty$ *then* $\mathcal{B}_n(\hat{\theta}_n^{(tz:bc)}) \sim (k_n/n)^{1-1/\kappa}(d_2^{1/\kappa} - d_1^{1/\kappa}) m_n^{-1/2} \times o(1)$.

(d). If tails are asymmetric $d_1 \neq d_2$ *and* not exactly Pareto $\xi < \infty$ *then*

$$\mathcal{B}_n\left(\hat{\theta}_n^{(tz:bc)}\right) = -\left(\frac{k_n}{n}\right)^{1-1/\kappa} \frac{1}{m_n^{1/2}} \left(d_2^{1/\kappa} - d_1^{1/\kappa}\right) \left\{ \frac{\kappa^2}{(\kappa-1)^2} \left(\tilde{\mathcal{R}}_n^{(a)} - \tilde{r}_n^{(a)}\right) \frac{1}{m_n^{1/2}} + o(1) \right\}$$

where $0 < \tilde{\mathcal{R}}_n^{(a)} = O(1)$ *and* $\tilde{r}_n^{(a)} = o(1)$ *depend on* m_n.

Remark We impose $m_n = O(\ln(n))$ and $k_n/m_n^{1/2} \to \infty$ in order to determine which bias terms dominate. Otherwise, meaningful inference from the bias expansion leads to numerous and in general less illuminating special cases. Consult Lemmas A.9 and A.10 in Chaudhuri and Hill (2013). The requirement $m_n = O(\ln(n))$ is quite practical since the A3′ bound $m_n = o(n^{2\xi/(2\xi+\kappa)})$ otherwise requires knowledge of the tail exponents ξ and κ. The requirement $k_n/m_n^{1/2} \to \infty$ is not too restrictive since we can always satisfy A3″ with $m_n \sim \lambda_m(\ln(n))^{\delta_m}$ and $k_n \sim \lambda_k(\ln(n))^{\delta_k}$ for $\lambda_k, \lambda_m > 0$ and $\delta_m/2 < \delta_k < \delta_m$.

Remark In each case as tails become heavier $\kappa \searrow 1$ higher order bias increases. Further, $\mathcal{B}_n(\hat{\theta}_n^{(tz:bc)}) \sim K(k_n/n)^{1-1/\kappa}/m_n$ under tail symmetry and otherwise $\mathcal{B}_n(\hat{\theta}_n^{(tz:bc)}) \sim K(k_n/n)^{1-1/\kappa}/m_n^{1/2}$, hence bias is inherently smaller under symmetric tails. This is logical since trimming is symmetric hence bias from trimming itself vanishes under tail symmetry.

Each case reveals trimming fewer observations (small k_n) leads to smaller bias.

Corollary 3.11 *Under A3″ it follows* $E[\hat{\theta}_n^{(tz:bc)}] - \theta$ *is closer to zero for* smaller k_n.

The general bias expression in Chaudhuri and Hill (2013, Lemma A.3) reveals the dual role of the number of tail observations m_n used for tail exponent estimation. A larger m_n implies a faster rate of convergence for tail estimators hence smaller higher bias for $\hat{\theta}_n^{(tz:bc)}$. A smaller m_n implies observations are taken from the extreme tails which are more closely described by a Pareto law, hence approximations for order statistics are sharper. In each case a larger m_n leads to smaller bias, *except* for an asymmetric non-exact Pareto law. This arises primarily from the two-tailed tail index estimator $\hat{\kappa}_{m_n}$: asymmetry and a deviation from exact Pareto decay exacerbates the bias of $\hat{\kappa}_{m_n}$, which overwhelms the otherwise diminishing impact of large m_n on bias. This bias is monotonically smaller for larger ξ and smaller $|d_2^{1/\kappa} - d_1^{1/\kappa}|$: the closer to a Pareto law, or tail symmetry, the smaller is $E[\hat{\theta}_n^{(tz:bc)}] - \theta$.

Corollary 3.12 *Assume A3″ holds. Then* $E[\hat{\theta}_n^{(tz:bc)}] - \theta$ *is closer to zero for* larger m_n*, except when* $d_2 \neq d_1$ *and* $\xi \in (0, \infty)$*, i.e.* Z_i *has asymmetric tails that are non-exactly Pareto. In the latter case, bias is smaller for larger* ξ *and smaller* $|d_2^{1/\kappa} - d_1^{1/\kappa}|$.

4 Probability Tail Decay - Threshold Crossing Latent Variable Model for Treatment Assignment

In this section we characterize the probability tail decay of $Z := hY$, the variable that, under A1 and A2′, point-identifies the average treatment effect $\theta = E[Z]$. We repeatedly use the following terminology concerning tail decay. Consult Resnick (1987) for details. A *regularly varying* function $\mathbb{R}(c)$ satisfies by Karamata's theorem $\mathbb{R}(c) = c^{-\xi}\mathcal{L}(c)$ where $\xi \neq 0$ is finite and $\mathcal{L}(c)$ is *slowly varying*: $\mathcal{L}(c)/\mathcal{L}(\lambda c) \to 1$ as $c \to \infty$ $\forall \lambda > 0$. Slow variation covers constants and powers of the natural log. A *power law* or regularly varying distribution tail has the form $P(|Z| \geq c) = c^{-\kappa}\mathcal{L}(c)$ with a positive *tail index* $\kappa > 0$. A *Paretian* tail $P(|Z| \geq c) = dc^{-\kappa}(1 + o(1))$ is a special case with slowly varying component $\mathcal{L}(c) = d(1 + o(1))$.

4.1 General Characterization

Our first result provides a general characterization of probability tails. Denote by E_{Y_i} expectations with respect to the measure induced by Y_i. As in (8), let κ denote the moment supremum of Z. Recall $a \wedge b = \min\{a, b\}$.

A5. The distributions of $DY/p(X)$ and $(1-D)Y/(1-p(X))$ are absolutely continuous on their support, and $p(X)|Y_1$ and $p(X)|Y_0$ have absolutely continuous distributions with Borel measurable density functions $f_{p(X)|Y_1}$ and $f_{p(X)|Y_0}$ for each $p(x) \in (0, 1)$ and Y_1, Y_0-*a.s.*

Theorem 4.1 *Let $c > 1$ be arbitrary. Under A1 and A5:*

$$P\left(|Z| > c\right) = E_{Y_1}\left[\int_0^{\frac{|Y_1|}{c}\wedge 1} r f_{p(X)|Y_1}(r)dr\right] + E_{Y_0}\left[\int_{\left(1-\frac{|Y_0|}{c}\right)\vee 0}^{1} (1-r) f_{p(X)|Y_0}(r)dr\right] \quad (21)$$

$$\frac{\partial}{\partial c}P\left(|Z| > c\right) = -\frac{1}{c^3}E_{Y_1}\left[I\left(|Y_1| \leq c\right) Y_1^2 f_{p(X)|Y_1}\left(\frac{|Y_1|}{c}\right)\right] \quad (22)$$
$$-\frac{1}{c^3}E_{Y_0}\left[I\left(|Y_0| \leq c\right) Y_0^2 f_{p(X)|Y_0}\left(1-\frac{|Y_0|}{c}\right)\right] = -\frac{1}{c^3}d\left(c\right).$$

If Z had a Paretian tail then $P(|Z| > c) = dc^{-\kappa}(1 + o(1))$ as $c \to \infty$ hence $(\partial/\partial c)P(|Z| > c) \sim -\kappa d c^{-\kappa-1}$. Property (22) therefore suggests that Z has a tail structure similar to a power law with index $\kappa = 2$, but with a multiplicative scale $d(c)$ governed by the threshold c and the distributions of $p(X), Y_0$ and Y_1. The fact that the scale itself is a function of c substantially complicates demonstrating the power law property since it is possible that $d(c) \to \infty$ so slow that $d\left(c\right)/c^3 \to 0$ at an exponential rate, or so fast that Z has a tail index $\kappa < 2$. In its generality, therefore, Theorem 4.1 is not particularly illuminating.

In order to characterize the scale $d(c)$ we need the conditional density $f_{p(X)|Y_j}(r)$. We therefore consider the popular latent variable framework for treatment assignment:

$$D = I\left(\alpha + \beta X - U \geq 0\right).$$

Obviously in practice β cannot be identified, hence $\beta = 1$ is the standard assumption. In the common standardized form $D = I(X - u \geq 0)$ with $u = U/\beta$ this is synonymous to $\beta = 1$ and X and U having different variances gauged by β. We allow $\beta \gtrless 1$ for ease, but everything that following is synonyms to fixing $\beta = 1$ and inspecting the relative probability tails of $\{U, X\}$.

We assume for simplicity U is independent of X, Y_1 and Y_0. Also, take $E[U] = 0$ and $Var(U) = 1$ as normalization in the rest of the paper. The assumption that the error U is additively separable and independent of X has implications on the treatment assignment (cf. Vytlacil (2002)). Generality is also lost due to the specific index structure $\alpha + \beta X$, but these help to abstract from issues peripheral to the demonstration of the power law tail decay. Without loss of further generality take $\beta > 0$.[14]

A6. U has an absolutely continuous distribution with density function f_U . X has support $\mathcal{X}$. Further $X|Y_1$ and $X|Y_0$ have absolutely continuous distributions with Borel measurable density functions $f_{X|Y_1}(x)$ and $f_{X|Y_0}(x)$ for each $x \in \mathcal{X}$ and *a.s.* with respect to Y_1, Y_0.

[14]Note that $\beta = 0$ implies $p(X) = F_U(\alpha) = p$ (constant) and as a result, under assumptions A1 and A2, $\theta = E[Y|D = 1] - E[Y|D = 0]$ meaning that there is no need for an IPW estimator. While its variance will increase with the proximity of p to 0 or 1, the IPW estimator does not, however, suffer from the limited overlap problem asymptotically as long as the constant $p \in (0, 1)$.

By independence of U and X:

$$p(X) = P(D = 1|X) = P(\alpha + \beta X \geq U) = F_U(\alpha + \beta X)$$

hence under A6 it follows for $j = 0, 1$:

$$f_{p(X)|Y_j}(r) = f_{X|Y_j}\left(\frac{F_U^{-1}(r) - \alpha}{\beta}\right)\frac{1}{\beta}\frac{1}{f_U\left(F_U^{-1}(r)\right)} \text{ where } r \in (0, 1).$$

Therefore, the result in (22) can be written as

$$\frac{\partial}{\partial c}P(|Z| > c) = -\frac{1}{c^3}\frac{1}{\beta}\mathcal{F}(\alpha, \beta, c) \text{ where } \mathcal{F}(\alpha, \beta, c) := \mathcal{F}_1(\alpha, \beta, c) + \mathcal{F}_0(\alpha, \beta, c), \tag{23}$$

and

$$\mathcal{F}_1(\alpha, \beta, c) := E_{Y_1}\left[Y_1^2 I(|Y_1| \leq c) f_{X|Y_1}\left(\frac{F_U^{-1}\left(\frac{|Y_1|}{c}\right) - \alpha}{\beta}\right)\frac{1}{f_U\left(F_U^{-1}\left(\frac{|Y_1|}{c}\right)\right)}\right]$$

$$\mathcal{F}_0(\alpha, \beta, c) := E_{Y_0}\left[Y_0^2 I(|Y_0| \leq c) f_{X|Y_0}\left(\frac{F_U^{-1}\left(1 - \frac{|Y_0|}{c}\right) - \alpha}{\beta}\right)\frac{1}{f_U\left(F_U^{-1}\left(1 - \frac{|Y_0|}{c}\right)\right)}\right].$$

It remains to deduce power law properties as a consequence of the behavior of $\mathcal{F}(\alpha, \beta, c)$ as $c \to \infty$.

The behavior of the ratios $f_{X|Y_j}((q_1 - \alpha)/\beta)/f_U(q_j)$ and therefore the relative tail decay of $X|Y_j$ and U plays a key roll, where for $j = 0, 1$ the $q_j's$ are quantiles

$$q_0 := F_U^{-1}(1 - |Y_0|/c) \text{ and } q_1 := F_U^{-1}(|Y_1|/c) \text{ for } |Y_j|/c \leq 1.$$

We demonstrate below by example how these two ratios influence the tail behavior of Z. Given the simplicity of the setup and a similar setting in Busso, DiNardo, and McCrary (2009) and Khan and Tamer (2010a), we focus on the cases where $Y_j \perp X, U$, and either $\{U, X\}$ are either identically distributed, or normally or Laplace distributed. Further, to avoid notational clutter we assume $\alpha = 0$, hence

$$D = I(\beta X - U \geq 0). \tag{24}$$

In practice a more general setting will clearly be desired. The following derivations serve as a basic groundwork for showing under limited overlap why heavy tails arise, and how sensitive they are to β.

4.2 Example: $\{U, X\}$ are iid

A brief example sheds some light on how the covariate slope β and the relative tail behavior of X and U affects the tail behavior of Z . In Khan and Tamer (2010a), following Lewbel (1997), the latent variable case treated is the standardization $\beta = 1$. Then $f_{X|Y_j}((q_1-\alpha)/\beta)/f_U(q_j) = f_{X|Y_j}(q_j)/f_U(q_j)$, and since $Y_j \perp X$ this further reduces to $f_X(q_j)/f_U(q_j)$. Thus, if X and U have the same densities *a.s.* Y_j then $\mathcal{F}_j(0,1,c) := E_{Y_j}[Y_j^2 I(|Y_j| \leq c)]$, and if Y_j has a finite variance then by dominated convergence $\lim_{c\to\infty} \mathcal{F}_j(0,1,c) = E[Y_j^2]$. This implies by (23) that $(\partial/\partial c)P(|Z| > c) \sim -c^{-3}(E[Y_0^2] + E[Y_1^2])$ hence Z has a Paretian tail with index 2. This proves the following.

Theorem 4.2 *Let $Y_j \perp X, U$, and let $\{U, X\}$ be iid. Then $P(|Z| > c) = dc^{-2}(1 + o(1))$ where $d = (1/2)(E[Y_0^2] + E[Y_1^2])$.*

Remark By dominated convergence the same conclusion follows when $f_X(r)/f_U(r) \to (0, \infty)$ as $|r| \to \infty$. Hence, the tail index is identically 2 when X and U have the same rate of distribution tail decay.

Two simple lessons are (i) when $Y_j \perp X, U$, and X and U have the same impact on $\mathcal{F}_j(\alpha, \beta, c)$ for $j = 0, 1$, then Z is heavy tailed with a hairline infinite variance; and (ii) lighter or heavier tails are driven by tail differences in X and U, and $\beta \gtreqless 1$, an issue largely ignored in the literature on IPW estimators for ATE. Notice (i) explains Khan and Tamer (2010a, Section 4) finding that their tail-trimmed ATE estimator has a $o(n^{1/2})$ rate of convergence when all variables are identically logit distributed: Z has an infinite variance hence negligible trimming results in sub-$n^{1/2}$ convergence (Csörgo, Horváth, and Mason, 1986; Hahn, Weiner, and Mason, 1991; Hill, 2013).

4.3 Example: Laplace Errors and Covariates

Let $(\epsilon_1, \epsilon_0, X, U)$ be independently distributed Laplace with mean 0 and variance 1. The cdf is

$$F(r) = \frac{1}{2}e^{\sqrt{2}r} \text{ if } r \leq 0 \text{ and } F(r) = 1 - \frac{1}{2}e^{-\sqrt{2}r} \text{ if } r > 0, \tag{25}$$

hence $f(r) = (1/\sqrt{2})e^{-\sqrt{2}|r|}$. Each variable is therefore symmetrically distributed about zero and ATE $\theta = 0$. In order to align with Khan and Tamer (2010a)'s setup of trimming based on X, and for simplicity, assume $Y_j \perp X, U$.

Theorem 4.3 *Under treatment assignment (24) and $Y_j \perp X, U$ it follows Z is symmetrically distributed about zero, and $P(|Z| > c) = dc^{-(1+1/\beta)}(1 + O(e^{-c/4}))$ for some d that depends on β.*

Remark As β increases the probability tail of Z becomes monotonically heavier, but the tail index $1 + 1/\beta$ is always above one, hence the ATE exists. The distribution is symmetric due to the treatment assignment location $\alpha = 0$, independence $Y_j \perp X, U$, and symmetry about zero for the distributions of

all variables $(\epsilon_1, \epsilon_0, X, U)$. We only show the impact of β on the tail index and not on the scale d in order to conserve space, and because only the tail index matters for large sample theory.

Remark Khan and Tamer (2010a: Section 4.1) consider $\beta = 1$ with $\{U, X\}$ that have a common logit distribution and therefore identical tails. By the simple example discussed in Section 4.1 this implies Z has a Paretian tail with index 2. However, in the heavier tail ($\beta > 1$) and thinner tail ($\beta < 1$) cases the trim-by-X estimator $\theta_n^{(tx)}$ behaves differently. Ultimately this is because Khan and Tamer (2010a) both standardize $\beta = 1$ *and* treat the distributions of U and X as identical. In a general framework, however, with treatment assignment $D = I(\beta X - U \geq 0)$ and $\beta > 0$, if we standardize $X - u \geq 0$ it follows $u := U/\beta$ has a variance that depends on β. See Sections 5.1 and 5.2 below for details on the trim-by-X estimator $\theta_n^{(tx)}$ and its general behavior.

Remark The second order term $O(e^{-c/4})$ is $O(c^{-\xi})$ for any $\xi > 0$. This implies power law assumption A3′ holds, and since $\xi > 0$ is arbitrary then any fractile $m_n \to \infty$ and $m_n = o(n)$ can be used for tail exponent estimation in the bias-corrected ATE estimator.

4.4 Example: Normal Errors and Covariates

Repeat the setup above, except now assume (Y_1, Y_0, X, U) are independently distributed $N(0, 1)$. Again ATE θ is identically 0. We obtain

$$\begin{aligned}\mathcal{F}_j(0, \beta, c) &= E_{Y_j}\left[Y_j^2 I\left(|Y_j| \leq c\right) \frac{f_{X|Y_j}\left(q_j/\beta\right)}{f_U\left(q_j\right)}\right] = E_{Y_j}\left[Y_j^2 I\left(|Y_j| \leq c\right) \frac{f_X\left(q_j/\beta\right)}{f_U\left(q_j\right)}\right] \\ &= \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^c y^2 \exp\left\{-\frac{y^2}{2}\right\} \exp\left\{-\frac{1-\beta^2}{2\beta^2} q_j^2\right\} dy.\end{aligned}$$

We only compute $\mathcal{F}_1(0, \beta, c)$ since $\mathcal{F}_0(0, \beta, c)$ is similar. Let $\Phi(z)$ denote the standard normal cdf. In this simplified setting it follows by a change of variables $z = y/c$ that

$$\mathcal{F}_1(0, \beta, c) = \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^1 c^3 z^2 \exp\left\{-\frac{c^2 z^2}{2}\right\} \exp\left\{-\frac{1-\beta^2}{2\beta^2} (\Phi^{-1}(z))^2\right\} dz. \tag{26}$$

Although exponential tails in general will lead to results similar to the Laplace case above, there are concrete differences worth noting. In particular, in the present case Z has a Paretian tail with index $\kappa = 1 + 1/\beta^2$ which is more sensitive to changes in β than in the Laplace case when $\beta \in (0, 2)$.

Theorem 4.4 *Under treatment assignment (24) and $Y_j \perp X, U$ it follows Z is symmetrically distributed about zero, and $P(|Z| > c) = dc^{-(1+1/\beta^2))}(1 + o(e^{-c/2}))$ for some d that depends on β.*

Remark The higher order term $o(e^{-c/2})$ is $O(c^{-\xi})$ for any $\xi > 0$, hence A3′ holds and any $m_n \to \infty$ and $m_n = o(n)$ is valid.

4.5 Example: Heterogeneous Distributions

The preceding examples exclude for simplicity the case where the errors and covariates have different distributions. Consider the case $Y_j \perp X, U$, as above, hence $\mathcal{F}_j(0,\beta,c) = E_{Y_j}[Y_j^2 I(|Y_j| \leq c) f_X(q_j/\beta)/f_U(q_j)]$. Then, for a given $\beta > 0$, a relatively heavier (thinner) tailed error U is associated with thinner (heavier) tails in Z. For example, if (Y_0, Y_1, X) are Laplace and U is normal then Z is heavier tailed than if all (Y_0, Y_1, X, U) are Laplace, and additionally in this case if $\beta = 1$ then $\kappa < 2$.

Conversely, if (Y_0, Y_1, X) are normal and U is Laplace or has a power law distribution tail, then Z is thinner tailed than if all (Y_0, Y_1, X, U) are normal. A similar outcome arises if (Y_0, Y_1, X) and U belong to the same distribution class but have different variances. For example, if (Y_0, Y_1, X) are iid normal, U is normal and $V[X] > V[U]$ then Z is heavier tailed than if all are iid normal.

5 Other Tail-Trimmed Estimators: Trimming by X

In this section we study the properties of the trim-by-X estimator:

$$\theta_n^{(tx)} = \frac{1}{n}\sum_{i=1}^{n} Z_i I\left(|X_i| \leq \gamma_n\right),$$

where $\{\gamma_n\}$ is a sequence of positive numbers, $\gamma_n \to \infty$. In order to simplify derivations we work in the following latent variable framework with Laplace distributed variables, in which case ATE $\theta = 0$.

A7. Assume the latent variable model (24) for treatment assignment holds; $Y_j \perp X, U$; and (Y_0, Y_1, X, U) are iid Laplace distributed with cdf (25).

Under A7 $\theta_n^{(tx)}$ is unbiased since $E[Z_i I(|X_i| > \gamma_n)] = E[(D_i Y_{1,i} + (1 - D_i)Y_{0,i})h_i I(|X_i| > \gamma_n)] = 0$. We abstract from the possibility of bias in order to focus on the convergence rate. Note that Khan and Tamer's (2010: Section 4.1) characterization of bias for $\theta_n^{(tx)}$ is presumedly under the assumption $\alpha \neq 0$ (see their footnote 7). Define the variance

$$\mathcal{S}_n^2 = E\left[\left(Z_i I\left(|X_i| \leq \gamma_n\right) - E\left[Z_i I\left(|X_i| \leq \gamma_n\right)\right]\right)^2\right] = E\left[Z_i^2 I\left(|X_i| \leq \gamma_n\right)\right] - \theta^2 \times (1 + o(1)), \quad (27)$$

where the second equality follows from $\gamma_n \to \infty$ and dominated convergence.

Khan and Tamer (2010b,a) study this estimator's convergence rate of bias under A7 with $\beta = 1$, and with other distributions. We characterize the limit distribution and rate of convergence $n^{1/2}/\mathcal{S}_n$ of $\theta_n^{(tx)}$, and again use $\beta \gtrless 1$ to mimic the setting where $\beta = 1$ and $\{U, X\}$ may have different distribution tails.

We then give details on precisely which observations Z_i are trimmed and show there may be only a weak correspondence between extremes in X_i and in Z_i . This will shed light on the inability of

$\theta_n^{(tx)}$ to control for heavy tails in small samples unless the sample portion of trimmed Z_i is very large, based on our controlled experiments in Section 6. We also compare $\theta_n^{(tx)}$ to the trim-by-Z estimator $\hat{\theta}_n^{(tz)}$ across these criteria. Finally, we present an improved version of $\theta_n^{(tx)}$ that uses a stochastic threshold and discuss how to set the trimming fractile such that it compares closely with $\hat{\theta}_n^{(tz)}$.

5.1 Properties of $\theta_n^{(tx)}$

We first characterize the variance $\mathcal{S}_n^2$. Under A7, $E[Y_j^2] = 1$ and hence by independence $E[Z_i^2 I(|X_i| \leq \gamma_n)] = E[h_i^2 I(|X_i| \leq \gamma_n)]$. Now apply dominated convergence and $\theta = 0$ to deduce $\mathcal{S}_n^2 \sim E[h_i^2 I(|X_i| \leq \gamma_n)]$, while

$$E\left[h_i^2 I\left(|X_i| \leq \gamma_n\right)\right] = E\left[\frac{1}{F_u(\beta X_i)} I\left(|X_i| \leq \gamma_n\right)\right] + E\left[\frac{1}{1 - F_u(\beta X_i)} I\left(|X_i| \leq \gamma_n\right)\right]. \tag{28}$$

By the Laplace assumption the first term in (28) is

$$\begin{aligned} E\left[\frac{1}{F_u(\beta X_i)} I\left(|X_i| \leq \gamma_n\right)\right] &= \int_{-\gamma_n}^{\gamma_n} \frac{1}{F(\beta x)} \frac{\partial}{\partial x} F\left(x\right) dx \qquad (29) \\ &= \sqrt{2}\left[\int_{-\gamma_n}^{0} e^{\sqrt{2}x(\beta-1)} dx + \int_{0}^{\gamma_n} \frac{e^{-\sqrt{2}x}}{2 - e^{-\sqrt{2}\beta x}} dx\right] \\ &= \int_{0}^{\sqrt{2}\gamma_n} e^{x(\beta-1)} dx + \int_{0}^{\sqrt{2}\gamma_n} \frac{e^{(\beta-1)x}}{2e^{\beta x} - 1} dx = \int_{0}^{\sqrt{2}\gamma_n} e^{x(\beta-1)} dx \times \left(1 + o(1)\right). \end{aligned}$$

The same derivation applies to the second term in (28).

If $\beta < 1$ then the tail index of Z_i is $\kappa > 2$ by Theorem 4.3, and $\int_0^{\sqrt{2}\gamma_n} e^{x(\beta-1)} dx = O(1)$ hence $E[h_i^2 I(|X_i| \leq \gamma_n)] \sim 2\int_0^{\infty} e^{-x(1-\beta)} dx = 2/(1-\beta) = E[h_i^2]$. The case studied in Khan and Tamer (2010a) is $\beta = 1$ which aligns with a tail index $\kappa = 2$ by Theorem 4.3, and $E[h_i^2 I(|X_i| \leq \gamma_n)] \sim \sqrt{2}\gamma_n \to \infty$ by (29). Finally, if $\beta > 1$ then the tail index $\kappa < 2$ by Theorem 4.3 , and $\int_0^{\sqrt{2}\gamma_n} e^{x(\beta-1)} dx = (\beta - 1)^{-1}\left(\exp\{\sqrt{2}\gamma_n(\beta-1)\} - 1\right)$ hence $E[h_i^2 I(|X_i| \leq \gamma_n)] \sim 2(\beta-1)^{-1}\left(\exp\{\sqrt{2}\gamma_n(\beta-1)\} - 1\right) \to \infty$.

This proves $\mathcal{S}_n^2$ is finite for each n and any β, and monotonically increasing in β when $\beta \geq 1$. Khan and Tamer (2010b, Theorem 4.1) assume the Lindeberg condition holds in order to prove asymptotic normality in a general environment. Using arguments in Khan and Tamer (2010b, Section 3), however, the condition is straightforward to verify here, hence we omit a proof.

Theorem 5.1 *Let A7 hold. If* $\beta < 1$ *then* $n^{1/2}(\theta_n^{(tx)} - \theta) \xrightarrow{d} N(0, 2/(1-\beta))$ *; if* $\beta = 1$ *then* $(n^{1/2}/\gamma_n)(\theta_n^{(tx)} - \theta) \xrightarrow{d} N(0,2)$*; and if* $\beta > 1$ *then* $(n^{1/2}/e^{\sqrt{2}\gamma_n(\beta-1)})(\theta_n^{(tx)} - \theta) \xrightarrow{d} N(0, 2/(\beta-1))$.

Theorem 5.1 reveals substantial differences in estimator behavior for the full range of $\beta > 0$. Small $\beta \in (0,1)$ implies Z_i has a finite variance hence $\theta_n^{(tx)}$ is $n^{1/2}$-convergent with asymptotic variance $2/(1$

$-\beta$) identical to the untrimmed estimator $1/n\sum_{i=1}^{n} Z_i$. Unity $\beta = 1$ aligns with a hairline infinite variance and convergence rate $n^{1/2}/\gamma_n = o(n^{1/2})$ with an asymptotic variance that depends on γ_n. Since $\theta_n^{(tx)}$ is unbiased when $Y_j \perp X$ and $\{Y_0, Y_1, U, X\}$ have symmetric distributions, a small γ_n such that $\gamma_n \to \infty$ slowly aligns with a smaller estimator variance and therefore mean-squared-error, and higher rate of convergence. Examples include $\gamma_n \sim \lambda(\ln(n))^{\delta}$ or $\gamma_n \sim \lambda \ln(\ln(n))$ for small $\delta, \lambda > 0$, in which case the asymptotic variance is λ^2. Recall that a slow $\gamma_n \to \infty$ implies a relatively greater degree of trimming occurs.

Finally, $\beta > 1$ aligns with heavy tails, in particular by Theorem 4.3 Z_i has a Paretian tail with index $\kappa = 1 + 1/\beta < 2$. In this case for a chosen sequence $\{\gamma_n\}$ the rate of convergence is exponentially slower. For example, if we use $\gamma_n = \lambda(\ln(n))^{\delta}$ for $\delta \in (0, 1]$ as in Khan and Tamer (2010b) for cases where the errors and regressors have exponential tails, then $\theta_n^{(tx)}$ has a convergence rate $n^{1/2}/(\ln(n))^{\delta}$ when $\beta = 1$ but only $n^{1/2}/e^{\sqrt{2}(\beta-1)\lambda(\ln(n))^{\delta}}$ when $\beta > 1$. Consider if $\gamma_n = \lambda \ln(n)$ and $\beta > 1$ then the rate is just $n^{1/2-\sqrt{2}\lambda(\beta-1)}$. We therefore require information on β in order to set λ small enough just to ensure $n^{1/2-\sqrt{2}\lambda(\beta-1)} \to \infty$. The choice of $\gamma_n = \lambda \ln(\ln(n))$, however, is always valid since $n^{1/2}/e^{\gamma_n(\beta-1)} = n^{1/2}/(\ln(n))^{\sqrt{2}\lambda(\beta-1)} \to \infty$.

5.2 Comparison of Estimators

We now compare the trim-by-X estimator $\theta_n^{(tx)}$ and the trim-by-Z estimator $\hat{\theta}_n^{(tz)}$ based on their rates of convergence and the ability to remove extreme observations of Z_i.

5.2.1 Rates of Convergence

We first derive the limit distribution of $\hat{\theta}_n^{(tz)}$. Combine the above derivation for $E[Z_i^2]$ in the case $\beta < 1$, with Corollary 3.3 for rates of convergence, and Theorem 3.2 for the power law property, to deduce the following.

Theorem 5.2 *Let A7 hold. If* $\beta < 1$ *then* $n^{1/2}(\hat{\theta}_n^{(tz)} - \theta) \xrightarrow{d} N(0, 2/(1-\beta))$ *, if* $\beta = 1$ *then* $(n/\ln(n))^{1/2}(\hat{\theta}_n^{(tz)} - \theta) \xrightarrow{d} N(0, d)$*, and if* $\beta > 1$ *then* $n^{1/2}/((n/k_n)^{\beta/(\beta+1)-1/2})(\hat{\theta}_n^{(tz)} - \theta) \xrightarrow{d} N(0, d^{2\beta/(\beta+1)}(\beta + 1)/(\beta - 1))$.

A direct comparison of the convergence rates of $\theta_n^{(tx)}$ and $\hat{\theta}_n^{(tz)}$ when $\beta \geq 1$ is complicated by the presence of the threshold γ_n in $\theta_n^{(tx)}$ and fractile k_n (with associated threshold c_n) in $\hat{\theta}_n^{(tz)}$. As a starting point, Khan and Tamer (2010a) suggest $\gamma_n = \lambda \ln(n)$ for some $\lambda > 0$ for the logit case with $\beta = 1$. Since Laplace and logit distributions will lead to the same essential results, consider $\gamma_n = \lambda \ln(n)$. Then $\theta_n^{(tx)}$ and $\hat{\theta}_n^{(tz)}$ have the same rates of convergence when $\beta \leq 1$ by Theorems 5.1 and 5.2.

However, if $\beta > 1$ then $e^{\gamma_n(\beta-1)} = n^{\sqrt{2}\lambda(\beta-1)}$ hence $\theta_n^{(tx)}$ has rate $n^{1/2-\sqrt{2}\lambda(\beta-1)} \to \infty$ *only provided* $\beta < 1 + 1/(2^{3/2}\lambda)$. Conversely, $\hat{\theta}_n^{(tz)}$ has a rate $n^{1/2}/(n/k_n)^{\beta/(\beta+1)-1/2} \to \infty$ for any value $\beta > 1$. Now, Paretian tail decay and the threshold construction imply $c_n = d^{1/(1+1/\beta)}(n/k_n)^{1/(1+1/\beta)}$. If

the fractile k_n implies the thresholds of $\hat{\theta}_n^{(tz)}$ satisfy $c_n \sim \lambda \ln(n)$, similar to γ_n, then we must have a number of trimmed $Z_i's$ equal to $k_n \sim Kn/(\ln(n))^{1+1/\beta}$. In this case the rate of convergence for $\hat{\theta}_n^{(tz)}$ is $n^{1/2}/(\ln(n))^{1-(\beta+1)/(2\beta)}$ which is faster than the rate $n^{1/2-\sqrt{2}\lambda(\beta-1)}$ for $\theta_n^{(tx)}$ with threshold $\gamma_n = \lambda \ln(n)$.

The preceding discussion suggests that the trim-by-Z estimator $\hat{\theta}_n^{(tz)}$ has a faster rate of convergence than the trim-by-X estimator $\theta_n^{(tx)}$ in the heavy tail case $\beta > 1$ when the same type of thresholds are used. Although we only treat the Laplace case here, in general this follows from the fact that limited overlap and therefore heavy tails imply potentially many large values of Z_i are present, while this slows down the convergence rate. The estimator $\hat{\theta}_n^{(tz)}$ removes extreme $Z_i's$ by construction, while for a given threshold sequence $\theta_n^{(tx)}$ is more likely to leave extremes present, which leads to its slower rate. See details below.

5.2.2 Ability to Remove Extreme Observations

By construction $\theta_n^{(tx)}$ removes Z_i only when X_i is large. We demonstrate the correspondence between extreme values of X_i and Z_i is weak by simulating $P(|Z_i| > c_z \mid |X_i| > c_x)$, the conditional probability that Z_i is large when X_i is large, for various $\{c_x, c_z\}$.

We use $D = I(\beta X - U \geq 0)$, as in Section 4, for choices $\beta \in \{.25, 1, 2\}$. Each (Y_0, Y_1, X, U) is iid standard normal or Laplace with cdf (25). We draw $R = 10,000$ samples $\{Z_i\}_{i=1}^n$ of size $n = 1,000,000$, and compute

$$P_{n,r}(c_z, c_x) := \frac{1/n \sum_{i=1}^n I(|Z_i| > c_z) I(|x_i| > c_x)}{1/n \sum_{i=1}^n I(|x_i| > c_x)}$$

for each r^{th} sample and $\{c_x, c_z\} \in [1, 10]$ with increments of 1: By the law of large numbers $P_{n,r}(c_z, c_x)$ will be very close to $P(|Z_i| > c_z||X_i| > c_x)$ with high probability.

Plots of $1/R \sum_{r=1}^R P_{n,r}(c_z, c_x)$ are contained in Figure 1. Always $P_{n,r}(c_z, c_x) \leq .6$, and $P_{n,r}(c_z, c_x) \leq .05$ when both $c_x, c_z \geq 4$. The event $|X_i| > c_x$ for large c_x is a very weak predictor of $|Z_i| > c_z$ for large c_z. Furthermore, the probability is smaller when tails are heavier: $P_{n,r}(c_z, c_x) \leq .3$ and $.4$ for Laplace and Normal, when $\beta = 2$ hence $\kappa < 2$. However, when $c_x \leq 2$ then $P_{n,r}(c_z, c_x)$ remains close to its maximum for all c_z. This is precisely what we find in our simulation experiments below: we must use small c_x to ensure as close a correspondance between X_i and Z_i extremes as possible, hence we must trim a large number of observations to ensure an adaptive version of $\theta_n^{(tx)}$ has small bias and is close to normally distributed.

5.3 Adaptive Threshold

In practice, for any chosen threshold γ_n a fixed number of $Z_i's$ will be trimmed. Further, a comprehensive strategy for choosing γ_n does not yet exist, and for some samples a chosen γ_n may result in

no trimming at all, or very few observations trimmed, and therefore estimator instability. A simple improvement for $\theta_n^{(tx)}$ bases trimming on an order statistic of X_i.

Under the assumption that there is only one covariate X, define $X_i^{(a)} := |X_i|$, denote the order statistics $X_{(1)}^{(a)} \geq X_{(2)}^{(a)} \cdots$, and let $\{k_n^{(x)}\}$ be an intermediate order sequence: $k_n^{(x)} \to \infty$ as $k_n^{(x)}/n \to 0$. Then an adaptive version of $\theta_n^{(tx)}$ is

$$\hat{\theta}_n^{(tx)} = \frac{1}{n}\sum_{i=1}^{n} Z_i I\left(|X_i| \leq X_{(k_n^{(x)})}^{(a)}\right),$$

in which case $P(|X_i| > \gamma_n) = k_n^{(x)}/n$. As discussed above, and demonstrated by simulation below, in practice a large $k_n^{(x)}$ should be used to increase the likelihood large $Z_i's$ are in fact trimmed considering that there is a weak correspondence between extreme X_i and Z_i.

Lemma A.2 in Hill (2013) can be extended to $\hat{\theta}_n^{(tx)}$ to verify $(n^{1/2}/\mathcal{S}_n)(\hat{\theta}_n^{(tx)} - \theta_n^{(tx)}) \xrightarrow{p} 0$. This proves the next claim.

Theorem 5.3 *Under A7 $\hat{\theta}_n^{(tx)}$ satisfies Theorem 5.1.*

6 Monte Carlo Study

We present a Monte Carlo experiment in order to study IPW estimators of θ. See the supplemental appendix Chaudhuri and Hill (2013) for a second experiment in which we demonstrate the power law property of $Z = Yh$.

6.1 Design

We use the latent variable treatment assignment setup from Section 4. We use $D = I(\beta X - U \geq 0)$ for choices $\beta \in \{.25, 1, 2\}$, and $Y_j \perp X, U$.

Initially we draw all variables from the same distribution. Each (Y_0, Y_1, X, U) is iid standard normal, or Laplace with cdf (25). We then draw $(Y_0, Y_1, X) \sim$ Laplace with $U \sim$ normal, and $(Y_0, Y_1, X) \sim$ normal with $U \sim$ Laplace. Under distribution symmetry and $Y_j \perp X, U$ in all cases the ATE $\theta = 0$ and $\hat{\theta}_n^{(tz)}$, $\theta_n^{(tx)}$ and $\hat{\theta}_n^{(tx)}$ are asymptotically unbiased. We compute $\hat{\theta}_n^{(tz)}$ and bias $\hat{\mathcal{B}}_n$ both with sample mean-centering and using the true $\theta = 0$ for centering, where in the latter case:

$$\hat{\theta}_n^{(tz)} = \frac{1}{n-k_n}\sum_{i=1}^{n} Z_i I\left(|Z_i| < Z_{(k_n)}^{(a)}\right) \text{ and } \hat{\mathcal{B}}_n = \frac{n}{n-k_n}\left(\frac{Z_{(m_n)}^{(+)} - Z_{(m_n)}^{(-)}}{Z_{(m_n)}^{(a)}}\right)\frac{\hat{\kappa}_{m_n}}{\hat{\kappa}_{m_n}-1}\frac{k_n}{n}Z_{(k_n)}^{(a)}.$$

The sample size is $n \in \{100, 10000\}$, where the latter proxies for the asymptotic case.

Define $\tilde{\theta}_n := 1/n\sum_{i=1}^{n} Z_i$. We compute the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and the optimal bias-corrected version $\hat{\theta}_n^{(tz:o)} = \hat{\theta}_n^{(tz:bc^*)} I(|\hat{\theta}_n^{(tz:bc^*)} - \tilde{\theta}_n| < |\hat{\theta}_n^{(tz)} - \tilde{\theta}_n|) + \hat{\theta}_n^{(tz)} I(|\hat{\theta}_n^{(tz:bc^*)} - \tilde{\theta}_n| \geq |\hat{\theta}_n^{(tz)} -$

$\tilde{\theta}_n|$) where $\hat{\theta}_n^{(tz:bc^*)} = \hat{\theta}_n^{(tz:bc^*)}(\phi_n^*) = \hat{\theta}_n^{(tz)} + \hat{\mathcal{B}}_n(\phi_n^*)$. We use fractiles $k_n = [2(\ln(n))^{1-\iota}]$ and $m_n(\phi_n^*)$ $= [\phi_n^* \ln(n)]$ where $\iota = 10^{-10}$ and ϕ_n^* minimizes $|\hat{\theta}_n^{(tz)} + \hat{\mathcal{B}}_n(\phi_n^*) - 1/n \sum_{i=1}^n Z_i|$ over $\phi \in [2, 8]$ subject to $\hat{\kappa}_{m_n(\phi_n^*)} > 1$, $Z^{(-)}_{(m_n(\phi_n^*))} < 0$ and $Z^{(-)}_{(m_n(\phi_n^*))} > 0$ (i.e. all sample tail components are well defined). First, this choice is justified by Theorem 3.6 since for Laplace or Normal $\{U, X\}$ under Theorems 4.3 and 4.4 Z_i in each case has a second order tail form $P(|Z_i| > c) = dc^{-\kappa}(1 + O(c^{-\xi}))$ with ξ $\geq \kappa$. In all cases $m_n = O(\ln(n))$ with $m_n/k_n \to \infty$ is valid. Second, in this study we trim $k_n \approx [2\ln(n)] = 9$ and 18 observations when $n = 100$ and $n = 10,000$ respectively. These fractiles work exceptionally well for heavy tail robustness, but work less well for estimating tail exponents required for bias-correction. We therefore allow the use of larger values for m_n, in particular up to $8k_n$.

We compare $\hat{\theta}_n^{(tz)}$ and $\hat{\theta}_n^{(tz:o)}$ to the untrimmed estimator $\tilde{\theta}_n = 1/n \sum_{i=1}^n Z_i$, the trim-by-$X$ estimator $\theta_n^{(tx)} = 1/n \sum_{i=1}^n Z_i I(|x_i| \leq \gamma_n)$ with threshold $\gamma_n = \ln(\ln(n))$, and the adaptive version $\hat{\theta}_n^{(tx)}$ $= 1/n \sum_{i=1}^n Z_i I(|x_i| \leq x^{(a)}_{(k_n^{(x)})})$ based on the order statistics of $x_i^{(a)} := |x_i|$ with $k_n^{(x)} = [2n/\ln(n)]$. Our choice of threshold γ_n for $\theta_n^{(tx)}$ is based on the fact that by design $\theta_n^{(tx)}$ is unbiased, while a small and slow $\gamma_n \to \infty$ implies heavier trimming which augments the convergence rate when $\beta > 1$, and $\gamma_n = \ln(n)$ is not always valid in practice. See Section 5. Further, with $\gamma_n = \ln(\ln(n))$ about 12% of observations are trimmed for $\theta_n^{(tx)}$ when $n = 100$, close to the 9% for $\hat{\theta}_n^{(tz)}$ and $\hat{\theta}_n^{(tz:o)}$. The fractile choice $k_n^{(x)}$ for $\hat{\theta}_n^{(tx)}$ implies heavy trimming, while $k_n^{(x)}$ is much larger than k_n to ensure extreme $Z_i's$ are trimmed. As a control we also use the much smaller $k_n^{(x)} = k_n$.

Let $\check{\theta}_{n,r}$ be the r^{th} sample's value of any estimator above over $r = 1, ..., R$ samples, $R = 10,000$. In Tables 1-2 we present the simulation mean $1/R \sum_{r=1}^R \check{\theta}_{n,r}$, median, mean squared error $s_n^2 := 1/R \sum_{r=1}^R \check{\theta}_{n,r}^2$, and the percent of observations that are trimmed on average per sample. We also use the simulation standardized ratio $\check{\theta}_{n,r}/s_n$ to test for normality by the Kolmogorov-Smirnov test. We report the KS statistic divided by its 5% critical value (denoted $\text{cv}_{.05}$) or $\text{KS}/\text{cv}_{.05}$: values above one imply rejection of standard normality at the 5% level. In Table 3 we report rejection frequencies for an asymptotic test of $\theta = 0$ against $\theta \neq 0$ at the {1%, 5%, 10%} levels based on the statistic $\check{\theta}_{n,r}/s_n$ and critical values from a standard normal distribution. We also compute kernel densities for each standardized estimator $\check{\theta}_{n,r}/s_n$, along with an iid standard normal random variable with sample size $10,000$.

6.2 Results

In the following we primarily discuss results for the realistic case $n = 100$, and omit all density plots. The tail-trimmed estimator with sample mean-centering is the one that would be used in practice, so we only report those results. Indeed, since $1/n \sum_{i=1}^n Z_i$ characterizes the sample center better than the true $\theta = 0$, centering with $1/n \sum_{i=1}^n Z_i$ for tail-trimming in general leads to better small sample bias for our estimators than if we use $\theta = 0$.[15] See Tables T.1-T.6 in Chaudhuri and Hill (2013) for

[15]Even if Z_i is symmetrically distributed about 0, in heavy tail cases there may be few large $Z_i > 0$ that render large $1/n \sum_{i=1}^n Z_i > 0$ (the same story applies for large $Z_i < 0$). The present study verifies such small sample bias in

all omitted results (sample mean-centering with $n = 10,000$, true mean-centering for each n), and see Figures F.1-F.4 there for all density plots.

Overall the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ is best: it has a small empirical bias and mean-squared-error [mse], it is roughly normally distributed even when $n = 100$, and empirical size in the test of $\theta = 0$ is near the nominal levels. The bias-corrected $\hat{\theta}_n^{(tz:o)}$ works nearly as well, but recall $\hat{\theta}_n^{(tz)}$ is already asymptotically unbiased. The structure of $\hat{\theta}_n^{(tz:o)}$ generates additional sampling error, but overall this estimator works quite well.

The untrimmed estimator $\tilde{\theta}_n$ is very sensitive to the limited overlap case $\beta > 1$. The presence of very large values influences the estimator's sign, giving the appearance of bias in small and large samples.

The trim-by-X estimator $\theta_n^{(tx)}$ compares closely to the untrimmed $\tilde{\theta}_n$: it is biased in finite samples when $\beta > 1$ suggesting sensitivity to limited overlap; it has a large mse; and is non-normal even when n is very large. This provides strong evidence of the weak link between extremes of X and Z. If, however, we utilize an order statistic of X for the threshold and set the amount of trimming $k_n^{(x)}$ to be large, then the adaptive version $\hat{\theta}_n^{(tx)}$ performs on par with $\hat{\theta}_n^{(tz)}$ in terms of bias and approximate normality.

Our experiment demonstrates that $\hat{\theta}_n^{(tx)}$ must have a large number of trimmed observations to perform well. This is clearly seen since the estimator performs so poorly when we use the same small fractile k_n as for $\hat{\theta}_n^{(tz)}$. Consider $n = 100$ for the case where all variables are normally distributed and $\beta > 1$. If we trim $k_n = [2(\ln(n))^{1-2\iota}] = 9$ observations then on average $\widehat{\mathring{\theta}}_n = .0223$ with mse .8403 and KS/cv$_{.05}$ is 7.37, but if we trim $k_n^{(x)} = [2n/\ln(n)] = 43$ then the mean is .0003, mse is .0428 and KS/cv$_{.05}$ is 1.054, a dramatic improvement. By comparison, for $\hat{\theta}_n^{(tz:o)}$ the mean is .004, mse is .0299, and KS/cv$_{.05}$ is .6913. Thus our estimator is closer to normal, and has half the mse due to a smaller dispersion. In simulations not reported here we trimmed 20, 25, 30 or 35 observations and found $\widehat{\mathring{\theta}}_n$ performs almost as bad with 20 (KS/cv$_{.05} = 6.11$), and far better with 30 (KS/cv$_{.05} = 1.61$). We need to trim at least 35 to gain results similar to $\hat{\theta}_n^{(tz)}$. The potential distribution deviation from normality by $\hat{\theta}_n^{(tx)}$ is clearly shown in the density plots in Chaudhuri and Hill (2013).

One difference between $\hat{\theta}_n^{(tx)}$ with large $k_n^{(x)}$ and $\hat{\theta}_n^{(tz)}$is worth noting: $\hat{\theta}_n^{(tx)}$ has a smaller mean squared error than $\hat{\theta}_n^{(tz)}$ when $\beta < 1$ and $\{U, X\}$ are iid. This arises because βX is thinner tailed than U when $\beta < 1$, while the correspondence between extremes of βX and Z is increased when βX and (therefore) Z are thinner tailed.

$1/n \sum_{i=1}^n Z_i$. However, if the sample size is small than removal of one or two large $Z_i > 0$ can render large $1/n \sum_{i=1}^n Z_i < 0$, hence trimming one or several large $Z_i < 0$ can decrease small sample bias. In this example, by using $|Z_i - 1/n \sum_{j=1}^n Z_j|$ for trimming, we elevate the center and therefore increase the likelihood that negative observations will be recognized as sample extremes and therefore be trimmed. Such centering improves our estimator's overall performance.

7 Conclusion

Under assumptions of unconfoundedness and limited overlap, the Average Treatment Effect can be point identified as the mean of a random variable Z that depends on the realized outcome and the propensity score for each sample unit. We give a formal treatment of the probability tail behavior of Z in a latent variable framework for treatment assignment, and show that greater degrees of limited overlap align with power law tail decay for Z, and possibly an infinite variance such that standard asymptotics for conventional estimators breaks down. We also study a version of a "robust" estimator of ATE obtained by tail trimming Z based on the covariates. In general such estimators are biased, and there may be a poor correspondence between the covariate's and Z''s extreme values so that trimming a very large number of observations may be required to ensure large $Z's$ are removed, a very inefficient way to robustify an IPW estimator. Small and even large sample performance of robust estimators crucially depend on the number of extreme observations of Z that are trimmed, hence we use information from Z itself to determine when to trim, we correct for bias, and use higher order asymptotics to show trimming few observations aligns with smaller higher order bias. Finally, we show in a controlled experiment the superior performance of our estimator.

The theory presented here provides a new groundwork for understanding and robustifying ATE estimators that are sensitive to limited overlap. We anticipate that our main results concerning the power law property of Z, and robustness of our estimator, will carry over to the case when the propensity score is estimated, but this must be treated elsewhere.

8 Appendix: Proofs

We require technical results proved in the supplemental appendix Chaudhuri and Hill (2013). Define

$$\tilde{Z}_i := Z_i - \theta, \;\; \tilde{Z}_i^{(a)} := \left|\tilde{Z}_i\right|, \;\text{ and } \;\; \tilde{Z}_{(1)}^{(a)} \geq \tilde{Z}_{(2)}^{(a)} \geq \cdots \geq \tilde{Z}_{(n)}^{(a)}$$

$$\hat{Z}_{n,i} := Z_i - \frac{1}{n}\sum_{j=1}^{n} Z_j, \;\; \hat{Z}_{n,i}^{(a)} := \left|\hat{Z}_{n,i}\right|, \;\text{ and } \;\; \hat{Z}_{n,(1)}^{(a)} \geq \hat{Z}_{n,(2)}^{(a)} \geq \cdots \geq \hat{Z}_{n,(n)}^{(a)}.$$

Lemma 8.1 *Under A1 and A3* $1/n\sum_{i=1}^{n} \tilde{Z}_i\{I(|\tilde{Z}_i| < \hat{Z}_{n,(k_n)}^{(a)}) - I(|\tilde{Z}_i| < \tilde{Z}_{(k_n)}^{(a)})\} = o_p(\sigma_n/n^{1/2})$.

Lemma 8.2 *Under A1 and A3* $1/n\sum_{i=1}^{n} \tilde{Z}_i\{I(|\hat{Z}_{n,i}| < \hat{Z}_{n,(k_n)}^{(a)}) - I(|\tilde{Z}_i| < \hat{Z}_{n,(k_n)}^{(a)})\} = o_p(\sigma_n/n^{1/2})$.

PROOF OF THEOREM 3.1. We will show:

$$\frac{n^{1/2}}{\sigma_n}\left(\hat{\theta}_n^{(tz)} - \theta\right) = \frac{n^{1/2}}{\sigma_n}\frac{1}{n-k_n}\sum_{i=1}^{n} \tilde{Z}_i I\left(\left|\tilde{Z}_i\right| < \tilde{Z}_{(k_n)}^{(a)}\right) = o_p\left(1\right). \tag{30}$$

The claim then follows from $E[\tilde{Z}_i] = 0$ and Theorem 2.2 in Hill (2013).

Use $\sum_{i=1}^n I(|\hat{Z}_{n,i}| < \hat{Z}^{(a)}_{n,(k_n)}) = n - k_n \; a.s.$ by construction and distribution continuity to deduce:

$$\hat{\theta}^{(tz)}_n - \theta - \frac{1}{n-k_n}\sum_{i=1}^n \tilde{Z}_i I\left(\left|\tilde{Z}_i\right| < \tilde{Z}^{(a)}_{(k_n)}\right) = \frac{1}{n-k_n}\sum_{i=1}^n \tilde{Z}_i \left\{ I\left(\left|\hat{Z}_{n,i}\right| < \hat{Z}^{(a)}_{n,(k_n)}\right) - I\left(\left|\tilde{Z}_i\right| < \tilde{Z}^{(a)}_{n,(k_n)}\right)\right\}.$$

The last summand is identically:

$$\frac{1}{n-k_n}\sum_{i=1}^n \tilde{Z}_i \left\{ I\left(\left|\tilde{Z}_i\right| < \hat{Z}^{(a)}_{n,(k_n)}\right) - I\left(\left|\tilde{Z}_i\right| < \tilde{Z}^{(a)}_{(k_n)}\right)\right\}$$
$$+ \frac{1}{n-k_n}\sum_{i=1}^n \tilde{Z}_i \left\{ I\left(\left|\hat{Z}_{n,i}\right| < \hat{Z}^{(a)}_{n,(k_n)}\right) - I\left(\left|\tilde{Z}_i\right| < \hat{Z}^{(a)}_{n,(k_n)}\right)\right\}.$$

The claim now follows from Lemmas 8.1 and 8.2. $\mathcal{QED}$.

PROOF OF THEOREM 3.2. If Z_i is symmetric about zero then $\mathcal{B}_n = 0$, so let Z_i be asymmetric. The claim then follows from properties of σ_n^2 detailed in (12) in the infinite variance case, the construction of c_n in (11) and bias formula (15). Together, we have the following. If $\kappa > 2$ then $\sigma_n^2 \to (0, \infty)$ hence $(n^{1/2}/\sigma_n)\mathcal{B}_n \sim Kn^{1/2}(k_n/n)c_n = Kn^{1/2}(k_n/n)^{1-1/\kappa} = Kk_n^{1-1/\kappa}/n^{1/2-1/\kappa}$. Therefore as long as $k_n/\ln(n) \to 0$ then $k_n/n^{(\kappa-2)/(2(\kappa-1))} \to 0$ for any $\kappa > 2$, hence $n^{1/2}\mathcal{B}_n = Kk_n^{1-1/\kappa}/n^{1/2-1/\kappa} \to 0$. Similarly, if $\kappa = 2$ then $\sigma_n^2 \sim K\ln(n)$ hence $(n^{1/2}/\sigma_n)\mathcal{B}_n \sim k_n^{1-1/2}/((\ln(n))^{1/2}n^{1/2-1/2}) = (k_n/\ln(n))^{1/2} \to 0$. Finally, if $\kappa < 2$ then $\sigma_n^2 \sim Kc_n^2(k_n/n)$ hence $(n^{1/2}/\sigma_n)\mathcal{B}_n \sim Kn^{1/2}(k_n/n)c_n/(c_n^2(k_n/n))^{1/2} = Kk_n^{1/2} \to \infty$. $\mathcal{QED}$.

PROOF OF THEOREM 3.6. We will prove $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}^{(tz:bc)}_n - \theta) \xrightarrow{d} N(0,1)$. Then $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}^{(tz:bc*)}_n - \theta) \xrightarrow{d} N(0,1)$ follows from arguments in Hill (2013, Theorems 2.1 and 2.2). The claim $n^{1/2}\mathfrak{V}_n^{-1}(\hat{\theta}^{(tz:o)}_n - \theta) \xrightarrow{d} N(0,1)$ with $\mathfrak{V}_n^2 = \mathcal{V}_n^2 + o_p(1)$ if $\kappa < 2$ and $\mathfrak{V}_n^2 = \sigma_n^2 + o_p(1)$ if $\kappa \geq 2$ follows from the above theory for $\{\hat{\theta}^{(tz)}_n, \hat{\theta}^{(tz:bc)}_n\}$, and the proof of Theorem 7.2.b,c in Hill and Prokhorov (2014).

In view of (30) we need only prove

$$\frac{n^{1/2}}{\mathcal{V}_n}\left\{\frac{1}{n-k_n}\sum_{i=1}^n \tilde{Z}_i I\left(\left|\tilde{Z}_i\right| < \tilde{Z}^{(a)}_{(k_n)}\right) + \hat{\mathcal{B}}_n\right\} \xrightarrow{d} N(0,1). \tag{31}$$

See Hill (2013, Theorem 2.1) for a proof that $\mathcal{V}_n^2 = \sigma_n^2(1 + o(1))$ if $\kappa < 2$ and $\mathcal{V}_n^2 = \sigma_n^2(1 + O(1))$ if $\kappa \geq 2$. We will show

$$\left(\frac{\hat{Z}^{(+)}_{n,(m_n)} - \hat{Z}^{(-)}_{n,(m_n)}}{\hat{Z}^{(a)}_{n,(m_n)}}\right)\frac{k_n}{n}\hat{Z}^{(a)}_{n,(k_n)} - \left(\frac{\tilde{Z}^{(+)}_{(m_n)} - \tilde{Z}^{(-)}_{(m_n)}}{\tilde{Z}^{(a)}_{(m_n)}}\right)\frac{k_n}{n}\tilde{Z}^{(a)}_{(k_n)} = o_p\left(\frac{\sigma_n}{n^{1/2}}\right).$$

The claim then follows from (31) and Theorem 2.1 in Hill (2013).

Recall $P(|Z_i| > c_n) = k_n/n$, and define sequences of positive non-random numbers $\{\tilde{c}_n^{(-)}, \tilde{c}_n^{(+)}, \tilde{c}_n\}$ by the relations

$$P(Z_i < -\tilde{c}_n^{(-)}) = P(Z_i > \tilde{c}_n^{(+)}) = P\left(|Z_i| > \tilde{c}_n\right) = \frac{m_n}{n}. \tag{32}$$

By power law A3′ $\tilde{c}_n^{(-)}/\tilde{c}_n = d_1^{1/\kappa}/d^{1/\kappa}$ and $\tilde{c}_n^{(a)}/\tilde{c}_n = d_2^{1/\kappa}/d^{1/\kappa}$.

By Lemma A.1 in Chaudhuri and Hill (2013), $\hat{Z}^{(\cdot)}_{n,(m_n)}/\tilde{c}_n^{(\cdot)} = 1 + O_p(1/m_n^{1/2})$, $\hat{Z}^{(a)}_{n,(m_n)}/\tilde{c}_n^{(a)} = 1 + O_p(1/m_n^{1/2})$ and $\hat{Z}^{(a)}_{n,(k_n)}/c_n = 1 + O_p(1/k_n^{1/2})$, and by Lemma 3 in Hill (2010), $\tilde{Z}^{(\cdot)}_{(m_n)}/\tilde{c}_n^{(\cdot)} = 1 + O_p(1/m_n^{1/2})$, $\tilde{Z}^{(a)}_{(m_n)}/\tilde{c}_n = 1 + O_p(1/m_n^{1/2})$ and $\tilde{Z}^{(\cdot)}_{(k_n)}/c_n = 1 + O_p(1/k_n^{1/2})$. Therefore:

$$\begin{aligned}
&\left(\frac{\hat{Z}^{(+)}_{n,(m_n)} - \hat{Z}^{(-)}_{n,(m_n)}}{\hat{Z}^{(a)}_{n,(m_n)}}\right)\frac{k_n}{n}\hat{Z}^{(a)}_{n,(k_n)} - \left(\frac{\tilde{Z}^{(+)}_{(m_n)} - \tilde{Z}^{(-)}_{(m_n)}}{\tilde{Z}^{(a)}_{(m_n)}}\right)\frac{k_n}{n}\tilde{Z}^{(a)}_{(k_n)} \\
&\quad = \frac{k_n}{n}\left\{O_p\left(1/m_n^{1/2}\right) \times c_n \times \left(1 + O_p\left(1/k_n^{1/2}\right)\right)\right. \\
&\qquad\qquad \left. + O_p\left(1/k_n^{1/2}\right) \times \frac{d_2^{1/\kappa} - d_1^{1/\kappa}}{d^{1/\kappa}} \times \left(1 + O_p\left(1/m_n^{1/2}\right)\right)\right\} \\
&\quad = O_p\left(\left(\frac{k_n}{n}\right)^{1-1/\kappa}\frac{1}{m_n^{1/2}}\right) + O_p\left(\frac{k_n^{1/2}}{n}\right) = o_p\left(\left(\frac{k_n}{n}\right)^{1-1/\kappa}\frac{1}{k_n^{1/2}}\right) + O_p\left(\frac{k_n^{1/2}}{n}\right) =: r_n,
\end{aligned}$$

say, where the last equality uses $k_n/m_n = o(1)$ under A3′. If $\kappa > 2$ then $\sigma_n \sim K$ and $(k_n/n)^{1/2-1/\kappa} \to 0$ hence $r_n = o_p(1/n^{1/2}) = o_p(\sigma_n/n^{1/2})$. If $\kappa = 2$ then $\sigma_n \sim K\ln(n)$ and again $r_n = o_p(\sigma_n/n^{1/2})$. If $\kappa < 2$ then $\sigma_n \sim K(n/k_n)^{1/\kappa-1/2}$, and since $k_n^{1/2}/n^{1/2} = o((n/k_n)^{1/\kappa-1/2})$ and $(k_n/n)^{1-1/\kappa}/k_n^{1/2} \sim K\sigma_n/n^{1/2}$ it again follows $r_n = o_p(\sigma_n/n^{1/2})$. $\mathcal{QED}$.

PROOF OF THEOREM 3.7. See Theorem 3.3 in Hill (2013) for a proof of $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$ and $\hat{\mathcal{V}}_n^2/\mathcal{V}_n^2 \xrightarrow{p} 1$. Then $\hat{\mathfrak{V}}_n^2/\mathfrak{V}_n^2 \xrightarrow{p} 1$ follows by construction. $\mathcal{QED}$.

PROOF OF THEOREM 4.1. By mutual exclusivity of the events $\{D = 1\}$ and $\{D = 0\}$ it follows

$$P\left(|hY| > c\right) = P\left(\left|\frac{DY_1}{p(X)} - \frac{(1-D)Y_0}{1-p(X)}\right| > c\right) = P\left(\left|\frac{DY_1}{p(X)}\right| > c\right) + P\left(\left|\frac{(1-D)Y_0}{1-p(X)}\right| > c\right). \tag{33}$$

Observe

$$\begin{aligned}
P\left(\frac{DY_1}{p(X)} > c\right) &= E_{Y_1}\left[I\left(\frac{Y_1}{p(X)} > c\right)p(X)|Y_1\right] \\
&= E_{Y_1}\left(E\left[p(X)I\left(p(X) < \frac{Y_1}{c} \wedge 1\right)|Y_1\right]\right) = E_{Y_1}\left[\int_0^{\frac{Y_1}{c}\wedge 1} r f_{p(X)|Y_1}\left(r|y\right)dr\right]
\end{aligned}$$

and

$$P\left(\frac{DY_1}{p(X)} < -c\right) = E_{Y_1}\left[\int_0^{\frac{-Y_1}{c}\wedge 1} r f_{p(X)|Y_1}(r|y)\, dr\right]$$

hence

$$P\left(\left|\frac{DY_1}{p(X)}\right| > c\right) = E_{Y_1}\left[\int_0^{\frac{|Y_1|}{c}\wedge 1} r f_{p(X)|y}(r|y)\, dr\right]. \tag{34}$$

By the same argument

$$P\left(\left|\frac{(1-D)Y_0}{1-p(X)}\right| > c\right) = E_{Y_0}\left[\int_{\left(1-\frac{|Y_0|}{c}\right)\vee 0}^1 (1-r) f_{p(X)|Y_0}(r) dr\right]. \tag{35}$$

Differentiate both sides of (34) and (35) with respect to c to deduce:

$$\begin{aligned}
&\frac{\partial}{\partial c} P\left(\left|\frac{DY_1}{p(X)}\right| > c\right) \qquad (36)\\
&= \frac{\partial}{\partial c}\int_{|Y_1|>c}\left\{\int_0^1 r f_{p(X)|Y_1}(r)dr\right\} f_{Y_1}(y)dy\\
&\qquad + \frac{\partial}{\partial c}\int_{|Y_1|\le c}\left\{\int_0^{\frac{|Y_1|}{c}} r f_{p(X)|Y_1}(r)dr\right\} f_{Y_1}(y)dy\\
&= \frac{\partial}{\partial c}\int_{-\infty}^{-c}\left\{\int_0^1 r f_{p(X)|Y_1}(r)dr\right\} f_{Y_1}(y)dy + \frac{\partial}{\partial c}\int_{-c}^{\infty}\left\{\int_0^1 r f_{p(X)|Y_1}(r)dr\right\} f_{Y_1}(y)dy\\
&\qquad + \frac{\partial}{\partial c}\int_{-c}^{c}\left\{\int_0^{\frac{|Y_1|}{c}} r f_{p(X)|Y_1}(r)dr\right\} f_{Y_1}(y)dy\\
&= -E[p(X)|Y_1=-c]f_{Y_1}(-c) - E[p(X)|Y_1=c]f_{Y_1}(c) + E[p(X)|Y_1=-c]f_{Y_1}(-c)\\
&\qquad + E[p(X)|Y_1=c]f_{Y_1}(c) + \int_{-c}^{c}\frac{\partial}{\partial c}\left\{\frac{|Y_1|}{c}\right\}\frac{|Y_1|}{c} f_{p(X)|Y_1}\left(\frac{|Y_1|}{c}\right) f_{Y_1}(y)dy\\
&= -\frac{1}{c^3}E_{Y_1}\left[Y_1^2 f_{p(X)|Y_1}\left(\frac{|Y_1|}{c}\right)1(|Y_1|\le c)\right],
\end{aligned}$$

and

$$\frac{\partial}{\partial c} P\left(\left|\frac{(1-D)Y_0}{1-p(X)}\right| > c\right) = -\frac{1}{c^3}E_Y\left[Y_0^2 f_{p(X)|Y_0}\left(1-\frac{|Y_0|}{c}\right)1(|Y_0|\le c)\right]. \tag{37}$$

Now combine (33)-(37) to prove the claims. $\mathcal{QED}$.

PROOF OF THEOREM 4.3. We only characterize $\mathcal{F}_1(\alpha,\beta,c)$ in (23) since $\mathcal{F}_0(\alpha,\beta,c)$ is

similar. Define $q_1 := F_U^{-1}(|Y_1|/c)$. By the Laplace definition it follows

$$q_1 = \frac{1}{\sqrt{2}}\left\{\ln 2 + \ln\left(\frac{|Y_1|}{c}\right)\right\} < 0 \text{ if } \frac{|Y_1|}{c} \leq 1/2 \text{ and } q_1 = -\frac{1}{\sqrt{2}}\left\{\ln 2 + \ln\left(1 - \frac{|Y_1|}{c}\right)\right\} > 0 \text{ if } \frac{|Y_1|}{c} > 1/2.$$

Use $Y_j \perp X, U$ and substitute $y = |Y_1|$ to deduce

$$\begin{aligned}
&\mathcal{F}_1(\alpha, \beta, c) \\
&= E_{Y_1}\left[Y_1^2 I\left(|Y_1| \leq c\right) \frac{f_X\left(q_1/\beta\right)}{f_U\left(q_1\right)}\right] \\
&= \sqrt{2}\int_0^{c/2} y^2 \exp\left\{-\sqrt{2}y\right\} \times \exp\left\{\left(\ln 2 + \ln\left(y/c\right)\right)\left(1/\beta - 1\right)\right\} dy \\
&\quad + \sqrt{2}\int_{c/2}^{c} y^2 \exp\left\{-\sqrt{2}y\right\} \times \exp\left\{\left(\ln 2 + \ln\left(1 - y/c\right)\right)\left(1/\beta - 1\right)\right\} dy \\
&= 2^{\frac{2-\beta}{2\beta}}\int_0^{c/2} y^2 \exp\left\{-\sqrt{2}y\right\} \times \left(y/c\right)^{1/\beta - 1} dy + 2^{\frac{2-\beta}{2\beta}}\int_{c/2}^{c} y^2 \exp\left\{-\sqrt{2}y\right\} \times \left(1 - y/c\right)^{1/\beta - 1} dy \\
&= 2^{\frac{1-2\beta}{2\beta}} c^{-(1/\beta - 1)}\left[\int_0^{c/\sqrt{2}} \exp\left\{-y\right\} \times y^{1+1/\beta} dy + \int_{c/\sqrt{2}}^{\sqrt{2}c} y^2 \exp\left\{-y\right\} \times \left(\sqrt{2}c - y\right)^{1/\beta - 1} dy\right] \\
&= 2^{1/(2\beta)-1} c^{-(1/\beta - 1)}\left(\mathcal{I}_1(c) + \mathcal{I}_2(c)\right).
\end{aligned}$$

It suffices to show each $\mathcal{I}_i(c) = K + O(e^{-c/4})$ and at least one $\lim_{c\to\infty} \mathcal{I}_i(c) > 0$. It then follows by (23) that $(\partial/\partial c)P(|Z| > c) = -Kc^{-2-1/\beta}(1 + O(e^{-c/4}))$, hence by dominated convergence $P(|Z| > c) = Kc^{-(1+1/\beta)}(1 + O(e^{-c/4}))$ as claimed.

If $\beta = 1$ then $\mathcal{I}_1(c) + \mathcal{I}_2(c) = \int_0^{\sqrt{2}c} y^2 \exp\{-y\} dy = 2 + o(e^{-c/4})$, and if $\beta \neq 1$ then $\lim_{c\to\infty} \mathcal{I}_1(c) \in (0, \infty)$ in view of the exponential term $\exp\{-y\}$. It remains to bound $\mathcal{I}_2(c)$. If $\beta < 1$ then

$$\begin{aligned}
\mathcal{I}_2\left(c\right) &= \int_{c/\sqrt{2}}^{\sqrt{2}c} y^2 \exp\left\{-y\right\} \times \left(\sqrt{2}c - y\right)^{1/\beta - 1} dy \leq 2^{(1-\beta)/2\beta} c^{1/\beta - 1} \int_{c/\sqrt{2}}^{\sqrt{2}c} y^2 \exp\left\{-y\right\} dy \\
&\leq 2^{(1+\beta)/2\beta} \frac{c^{1/\beta+1}}{\exp\left\{\sqrt{2}c\right\}} = o\left(e^{-c/4}\right).
\end{aligned}$$

Finally, if $\beta > 1$ then $e^{c/4} y^2 \exp\{-y\} \times (\sqrt{(}2)c - y)^{1/\beta - 1} dy \leq K y^{-(1+\delta)}$ for all $y \in [c/\sqrt{2}, \sqrt{2}c - \iota]$, tiny $\iota > 0$, some tiny $\delta > 0$ and some large $K > 0$. Therefore $\int_{c/\sqrt{2}}^{\sqrt{2}c-\iota} y^2 \exp\{-y\} \times (\sqrt{2}c - y)^{1/\beta - 1} dy$ $= o(e^{-c/4})$ for any tiny $\iota > 0$, hence $\mathcal{I}_2(c) = K + O(e^{-c/4})$. $\mathcal{QED}$.

PROOF OF THEOREM 4.4. Symmetry follows from $\alpha = 0$, independence $Y_j \perp X, U$, and distribution symmetry for all (Y_0, Y_1, X, U).

Now let $\Phi(w)$ and $\phi(w)$ be the normal cdf and pdf. In order to characterize the standard normal quantile $\Phi^{-1}(u/c)$ for $u \in [0,c]$, we use the expansion $1-\Phi(w) = (1+O\left(1/w^2\right)) \times \phi(w)/w$ to solve $u/c = \phi(w(c))/w(c)$ for some $w(c)$ as $c \to \infty$ hence as $w(c) \to \infty$. See Gray and Wang (1991), cf. Lew (1981) and Hawkes (1982). Rudimentary algebra reveals $w(c)$ satisfies

$$w(c) = 2^{1/2}\left(\ln\left(c\right)\right)^{1/2}\left(1-\frac{\ln(u)}{\ln\left(c\right)}-\frac{\ln\left(2\pi\right)}{\ln\left(c\right)}\right)^{1/2}\left(1+O\left(1/\ln\left(c\right)\right)\right).$$

Since $|\Phi^{-1}(u/c)| = w(c)$ use formula (26) to deduce $\mathcal{F}_1(0,\beta,c)$ is identically:

$$\begin{aligned}
&\left(\frac{2}{\pi}\right)^{1/2}\int_0^c \frac{u^2}{\exp\left\{u^2/2\right\}}\exp\left\{\frac{\beta^2-1}{\beta^2}\ln\left(c\right)\left(1-\frac{\ln(u)}{\ln\left(c\right)}-\frac{\ln\left(2\pi\right)}{\ln\left(c\right)}\right)\left(1+O\left(1/\ln\left(c\right)\right)\right)\right\}du \\
&\qquad = \left(\frac{2}{\pi}\right)^{1/2}\int_0^c u^2\exp\left\{-u^2/2\right\}c^{(\beta^2-1)\beta^{-2}(1-\ln(u)/\ln(c)-\ln(2\pi)/\ln(c))(1+O(1/\ln(c)))}du \\
&\qquad = \left(\frac{2}{\pi}\right)^{1/2}c^{(\beta^2-1)\beta^{-2}(1+O(1/\ln(c)))}\int_0^c u^2\exp\left\{-u^2/2\right\}c^{-(\beta^2-1)\beta^{-2}O(1/\ln(c))(\ln(u)+\ln(2\pi))}du
\end{aligned}$$

If $\beta = 1$ then $\lim_{c\to\infty}\mathcal{F}_1(0,1,c) = 1$, in particular $\mathcal{F}_1(0,1,c) = 1 + o(e^{-c/2})$ is easily verified given the normal density.

Now assume $\beta \neq 1$ and let $d\left(\beta\right)$ be a positive finite function of β that may change from line to line. Observe

$$\ln\left(c^{-(\beta^2-1)\beta^{-2}O(1/\ln(c))(\ln(u)+\ln(2\pi))}\right) = -O\left(1\right)\times(\ln(2\pi u)),$$

hence by the monotonicity of the natural log $c^{-(\beta^2-1)\beta^{-2}O(1/\ln(c))(\ln(u)+\ln(2\pi))} = Ku^{-O(1)}$. Similarly $c^{(\beta^2-1)\beta^{-2}O(1/\ln(c))} = O(1)\times c^{(\beta^2-1)\beta^{-2}}$. Therefore:

$$\mathcal{F}_1(0,\beta,c) = d(\beta)\times O\left(1\right)\times c^{(\beta^2-1)\beta^{-2}}\int_0^c u^2\exp\left\{-u^2/2\right\}Ku^{-O(1)}du = d(\beta)\times O\left(1+o(e^{-c/2})\right)\times c^{(\beta^2-1)\beta^{-2}}.$$

Finally, use the formula for $(\partial/\partial c)P(|Z|>c)$ in (23) to deduce

$$\frac{\partial}{\partial c}P\left(|Z|>c\right) = -d(\beta)c^{-3}c^{(\beta^2-1)\beta^{-2}}O\left(1+o(e^{-c/2})\right) = -d(\beta)c^{-2-\beta^{-2}}\left(1+o(e^{-c/2})\right),$$

hence by dominated convergence $P(|Z|>c) = d(\beta)c^{-(1+\beta^{-2})}(1+o(e^{-c/2}))$. $\mathcal{QED}$.

References

Andrews, D., and M. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497–517.

Bahadur, R. (1960): "Asymptotic Efficiency of Tests and Estimates," *Sankhya*, 22, 229–252.

Busso, M., J. DiNardo, and J. McCrary (2009): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," Mimeo.

——— (2011): "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," Mimeo.

Cao, W., A. Tsiatis, and M. Davidian (2009): "Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data," *Biometrika*, 96, 723–734.

Chaudhuri, S., and J. Hill (2013): "Supplemental Appendix for Robust Estimation for Average Treatment Effects," mimeo.

Chaudhuri, S., and H. Min (2012): "Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing Data," Mimeo.

Chritsopeit, N., and H. Werner (2001): "A Necessary and Sufficient Condition of a Sequence of Random Variables Converging to a Normal Distribution," *Econometric Theory*, 17, 278–281.

Crump, R., V. Hotz, G. Imbens, and O. Mitnik (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199.

Csörgo, S., L. Horváth, and D. Mason (1986): "What Portion of the Sample Makes a Partial Sum Asymptotically Stable or Normal?," *Probability Theory and Related Fields*, 72, 1–16.

Dehejia, R., and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of American Statistical Association*, 94, 1053–1062.

Feller, W. (1971): *An Introduction to Probability Theory and Its Applications (Vol. II)*. Wiley, New York.

Frolich, M. (2004): "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86, 77–90.

Galambos, J. (1987): *The Asymptotic Theory of Extreme Order Statistics*. Krieger: Malabar.

Goldie, C., and R. Smith (1987): "Slow Variation with Remainder: Theory and Applications," *Quarterly Journal of Mathematics*, 38, 45–71.

Graham, B. S. (2011): "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions," *Econometrica*, 79, 437 – 452.

Graham, B. S., C. Pinto, and D. Egel (2011): "Inverse Probability Tilting for Moment Condition Models with Missing Data," Mimeo.

Gray, H., and S. Wang (1991): "A General Method for Approximating Tail Probabilities," *Journal of American Statistical Association*, 86, 159–166.

Haeusler, E., and J. Teugels (1985): "On Asymptotic Normality of Hill's Estimator for the Exponent of Regular Variation," *Annals of Statistics*, 13, 743–756.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Hahn, M., J. Kuelbs, and J. Samur (1987): "Asymptotic Normality of Trimmed Sums of ϕ-Mixing Random Variables," *Annals of Probability*, 15, 1395–1418.

Hahn, M., D. Weiner, and D. Mason (1991): *Sums, Trimmed Sums and Extremes*. Birkhäuser: Berlin.

Hall, P. (1982): "On Some Simple Estimates of an Exponent of Regular Variation," *Journal of the Royal Statistical Societ Series B*, 44, 37–42.

Hawkes, A. (1982): "Approximating the Normal Tail," *Journal of the Royal Statistical Societ Series D*, 31, 231–236.

Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

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

Hill, B. M. (1975): "A Simple General Approach to Inference about the Tail of a Distribution," *Annals of Statistics*, 3(5), 1163–1174.

Hill, J. B. (2010): "On Tail Index Estimation for Dependent, Heterogeneous Data," *Econometric Theory*, 26, 1398–1436.

——— (2012): "Heavy-Tail and Plug-In Robust Consistent Conditional Moment Tests of Functional Form," in *Festschrift in Honor of Hal White*, ed. by X. Chen, and N. Swanson, pp. 241–274. Springer: New York.

——— (2013): "Expected Shortfall Estimation and Gaussian Inference for Infinite Variance Time Series," *Journal of Financial Econometrics*, forthcoming.

——— (2014): "Tail Index Estimation for a Filtered Dependent Time Series," *Statistica Sinica*, p. forthcoming.

Hill, J. B., and A. Prokhorov (2014): "GEL Estimation for Heavy-Tailed GARCH Models with Robust Empirical Likelihood Inference," *Working Paper, Dept. of Economics, University of North Carolina - Chapel Hill*.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores," *Econometrica*, 71, 1161–1189.

Ho, D., K. Imai, G. King, and E. Stuart (2007): "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 15, 199–236.

Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of American Statistical Association*, 47, 663–685.

Hsing, T. (1991): "On Tail Index Estimation Using Dependent Data," *Annals of Statistics*, 19, 1547–1569.

Huisman, R., K. Koedijk, C. Kool, and F. Palm (2001): "Tail-Index Estimates in Small Samples," *Journal of Business and Economic Statistics*, 19, 208–216.

Ibragimov, I., and I. Linnik (1971): *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff.

Jureckova, J. (1981): "Tail-Behavior of Location Estimators," *Annals of Statistics*, 9, 578–585.

Kang, J., and J. Schafer (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22, 523–539.

Khan, S., and E. Tamer (2010a): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

——— (2010b): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Working Paper.

Klein, R., C. Shen, and F. Vella (2011): "Semiparametric Selection Models with Binary Outcomes," IZA DP No. 6008.

Leadbetter, M., G. Lindgren, and H. Rootzen (1983): *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag.

Lechner, M. (2008): "A Note on the Common Support Problem in Applied Evaluation Studies," *Annals of Economic and Statistics*, 91/92, 217–235.

Lee, B., J. Lessler, and E. Stuart (2011): "Weight Trimming and Propensity Score Weighting," *PLOS One*, 6.

Lew, R. (1981): "An Approximation to the Cumulative Normal Distribution with Simple Coefficients,," *Journal of the Royal Statistical Societ Series C*, 30, 299–301.

Lewbel, A. (1997): "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory*, 13, 32–51.

Lunceford, J., and M. Davidian (2004): "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects : A Comparative Study," *Statistics in Medicine*, 23, 2937–2960.

McCaffrey, D., G. Ridgeway, and A. Morral (2004): "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, 9, 403–425.

Peng, L. (2001): "Estimating the Mean of a Heavy Tailed Distribution," *Statistics and Probability Letters*, 52, 255–264.

Potter, F. (1993): "The Effect of Weight Trimming on Nonlinear Survey Estimates," in *Proceedings of the Section on Survey Research Methods & Research*. American Statistical Association.

Resnick, S. (1987): *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag: New York.

Ridgeway, G., and D. McCaffrey (2007): "Comment: Demystifying Double Robustness :A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22, 540–543.

Rosenbaum, P., and D. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Stuart, E. (2010): "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, 25, 1–21.

Traskin, M., and D. Small (2012): "Defining the Study Population for an Observational Study to Ensure Sufficient Overlap: A Tree Approach," Mimeo.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

Wooldridge, J. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141, 1281–1301.

TABLE 1 - Estimator Properties (Sample Mean-Centering, Normal *or* Laplace, $n = 100$)

	$(Y_0, Y_1, X, U) \sim$ **Normal**						$(Y_0, Y_1, X, U) \sim$ **Laplace**				
	$\beta = .25$ ($\kappa = 17$)						$\beta = .25$ ($\kappa = 5$)				
Estimator	Mean	Median	MSE	KS	Tr%		Mean	Median	MSE	KS	Tr%
Un-Trimmed	-.0039	-.0053	.0414	.9572	0.00		.0039	.0048	.0477	.5510	0.00
Tail-Trimmed-Z	-.0004	-.0003	.0493	.5756	9.00		-.0012	-.0008	.0478	.5635	9.00
Bias-Corrected-Z	-.0004	-.0002	.0512	.6254	9.00		-.0012	-.0024	.0626	.4994	9.00
Tail-Trimmed-X	-.0039	-.0054	.0398	1.043	12.0		.0044	.0039	.0436	.5042	12.0
TX-Adapt ($k_n^{(x)}$)	-.0010	-.0026	.0227	.8294	43.0		.0030	.0020	.0338	.5211	43.0
TX-Adapt (k_n)	-.0037	-.0065	.0372	1.343	9.00		-.0008	.0002	.0419	.7894	9.00

	$\beta = 1$ ($\kappa = 2$)						$\beta = 1$ ($\kappa = 2$)				
Estimator	Mean	Median	MSE	KS	Tr%		Mean	Median	MSE	KS	Tr%
Un-Trimmed	.0015	-.0013	.2017	1.989	0.00		-.0138	-.0034	.9980	2.908	0.00
Tail-Trimmed-Z	-.0005	.0005	.0330	1.142	9.00		.0006	.0004	.0356	.5238	9.00
Bias-Corrected-Z	-.0005	.0012	.0399	1.006	9.00		.0002	.0005	.0664	.9462	9.00
Tail-Trimmed-X	-.0019	-.0021	.0836	.7941	12.0		-.0043	-.0050	.0796	1.025	12.0
TX-Adapt ($k_n^{(x)}$)	-.0014	-.0009	.0265	.6937	43.0		-.0029	-.0052	.0395	1.142	43.0
TX-Adapt (k_n)	.0026	.0023	.0638	.8194	9.00		-.0027	.0006	.0912	.8528	9.00

	$\beta = 2$ ($\kappa = 1.25$)						$\beta = 2$ ($\kappa = 1.5$)				
Estimator	Mean	Median	MSE	KS	Tr%		Mean	Median	MSE	KS	Tr%
Un-Trimmed	.0077	-.0022	4.053	8.222	0.00		.0447	-.0030	21.70	7.562	0.00
Tail-Trimmed-Z	.0003	.0005	.0224	.8336	9.00		-.0014	-.0014	.0264	.3650	9.00
Bias-Corrected-Z	.0004	.0004	.0299	.6913	9.00		-.0017	-.0022	.0585	.6997	9.00
Tail-Trimmed-X	.0079	-.0033	4.053	8.220	12.0		-.0034	-.0026	.2788	3.292	12.0
TX-Adapt ($k_n^{(x)}$)	.0003	-.0024	.0428	1.054	43.0		.0010	.0006	.0549	1.392	43.0
TX-Adapt (k_n)	.0223	.0019	.8403	7.317	9.00		.0056	-.0019	.3690	3.514	9.00

"Untrimmed" is the untrimmed estimator $\tilde{\theta}_n$; "Tail-Trimmed-Z" is the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and "Bias-Corrected-Z" is the bias-corrected tail-trimmed $\hat{\theta}_n^{(tz:o)}$: both use *sample mean-centering* for trimming. "Tail-Trimmed-X" is $\theta_n^{(tx)}$; and "TX-Adapt (k)" is the adaptive version $\hat{\theta}_n^{(tx)}$ of $\theta_n^{(tx)}$. KS is the Kolmogorov-Smirnov test statistic divided by its 5% critical value: values above 1 indicate rejection of standard normality at the 5% level. Tr% is the percent of observations Z_i trimmed. κ is the tail index of $Z = h(X)Y$. Other than KS, all values are averages over the randomly drawn 10,000 samples.

TABLE 2 - Estimator Properties (Sample Mean-Centering, , Normal *and* Laplace, $n = 100$)

	$(Y_0, Y_1, X) \sim$ **Normal,** $U \sim$ **Laplace**						$(Y_0, Y_1, X) \sim$ **Laplace,** $U \sim$ **Normal**				
	$\beta = .25$						$\beta = .25$				
Estimator	Mean	Median	MSE	KS	Tr%		Mean	Median	MSE	KS	Tr%
Un-Trimmed	.0022	.0021	.0419	.7364	0.00		-.0006	.0020	.0490	.5030	0.00
Tail-Trimmed-Z	-.0005	-.0003	.0486	.3556	9.00		-.0014	-.0013	.0473	.4847	9.00
Bias-Corrected-Z	-.0006	-.0005	.0506	.4074	9.00		-.0011	-.0017	.0615	.4717	9.00
Tail-Trimmed-X	.0017	.0009	.0400	.7734	12.0		.0062	.0023	.0424	.4512	12.0
TX-Adapt ($k_n^{(x)}$)	.0018	.0015	.0227	.6324	43.0		.0004	-.0014	.0325	.6202	43.0
TX-Adapt (k_n)	.0004	-.0004	.0367	.4649	9.00		.0012	.0022	.0421	.6672	9.00
	$\beta = 1$						$\beta = 1$				
Estimator	Mean	Median	MSE	KS	Tr%		Mean	Median	MSE	KS	Tr%
Un-Trimmed	.0005	-.0008	.0701	.5615	0.00		-.0276	-.0007	5.809	5.327	0.00
Tail-Trimmed-Z	-.0001	.0003	.0349	.8038	9.00		-.0000	.0012	.0331	.8253	9.00
Bias-Corrected-Z	-.0002	.0000	.0416	.5961	9.00		.0004	.0012	.0587	.7352	9.00
Tail-Trimmed-X	-.0000	-.0009	.0621	.3751	12.0		-.0004	-.0013	.1220	1.796	12.0
TX-Adapt ($k_n^{(x)}$)	.0004	-.0010	.0261	.4987	43.0		.0029	.0020	.0388	.9124	43.0
TX-Adapt (k_n)	.0056	.0048	.0535	.5131	9.00		-.0041	.0026	.1775	2.886	9.00
	$\beta = 2$						$\beta = 2$				
Estimator	Mean	Median	MSE	KS	Tr%		Mean	Median	MSE	KS	Tr%
Un-Trimmed	.0007	-.0005	.2402	4.074	0.00		.0218	.0041	1.279	8.992	0.00
Tail-Trimmed-Z	-.0005	-.0004	.0251	.8352	9.00		-.0008	.0002	.0239	.8842	9.00
Bias-Corrected-Z	.0002	-.0005	.0414	.8932	9.00		-.0004	-.0007	.0447	1.042	9.00
Tail-Trimmed-X	.0023	-.0034	.1689	2.953	12.0		.0219	.0044	1.281	9.078	12.0
TX-Adapt ($k_n^{(x)}$)	.0011	.0013	.0349	.6414	43.0		.0013	.0038	.0742	1.219	43.0
TX-Adapt (k_n)	.0063	-.0002	.1200	2.028	9.00		.0182	.0021	3.411	9.732	9.00

"Untrimmed" is the untrimmed estimator $\tilde{\theta}_n$; "Tail-Trimmed-Z" is the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and "Bias-Corrected-Z" is the bias-corrected tail-trimmed $\hat{\theta}_n^{(tz:o)}$: both use *sample mean-centering* for trimming. "Tail-Trimmed-X" is $\theta_n^{(tx)}$; and "TX-Adapt (k)" is the adaptive version $\hat{\theta}_n^{(tx)}$ of $\theta_n^{(tx)}$. KS is the Kolmogorov-Smirnov test statistic divided by its 5% critical value: values above 1 indicate rejection of standard normality at the 5% level. Tr% is the percent of observations Z_i trimmed. κ is the tail index of $Z = h(X)Y$. Other than KS, all values are averages over the randomly drawn 10,000 samples.

TABLE 3 - Rejection Frequencies (Sample Mean-Centering, $n = 100$)

$(Y_0, Y_1, X, U) \sim$ **Normal**

β	Untrimmed	Tail-Trimmed-Z	Bias-Corrected-Z	Tail-Trimmed-X	TX-Adapt ($k_n^{(x)}$)	TX-Adapt (k_n)
.25	.013, .051, .094	.011, .046, .099	.011, .046, .100	.010, .051, .093	.013, .047, .093	.006, .048, .099
1	.015, .029, .040	.008, .052, .091	.010, .050, .101	.018, .045, .087	.008, .047, .105	.016, .061, .102
2	.013, .023, .029	.008, .052, .094	.014, .052, .096	.013, .022, .029	.014, .051, .105	.012, .017, .021

$(Y_0, Y_1, X, U) \sim$ **Laplace**

β	Untrimmed	Tail-Trimmed-Z	Bias-Corrected-Z	Tail-Trimmed-X	TX-Adapt ($k_n^{(x)}$)	TX-Adapt (k_n)
.25	.015, .059, .104	.010, .051, .108	.009, .051, .102	.013, .051, .097	.011, .051, .104	.010, .040, .092
1	.019, .032, .047	.010, .045, .084	.015, .047, .091	.012, .049, .099	.008, .048, .099	.014, .045, .089
2	.008, .015, .018	.009, .046, .101	.013, .051, .094	.026, .047, .067	.012, .054, .104	.027, .041, .061

$(Y_0, Y_1, X) \sim$ **Normal**, $U \sim$ **Laplace**

β	Untrimmed	Tail-Trimmed-Z	Bias-Corrected-Z	Tail-Trimmed-X	TX-Adapt ($k_n^{(x)}$)	TX-Adapt (k_n)
.25	.014, .047, .101	.007, .053, .103	.009, .056, .101	.012, .051, .100	.009, .043, .102	.013, .047, .095
1	.013, .043, .085	.006, .046, .101	.011, .049, .104	.011, .051.100	.012.,053, .109	.015, .049, .097
2	.025, .051, .076	.011, .047, .094	.014, .047, .093	.025, .058, .089	.012, .053, .102	.023, .056, .083

$(Y_0, Y_1, X) \sim$ **Laplace**, $U \sim$ **Normal**

β	Untrimmed	Tail-Trimmed-Z	Bias-Corrected-Z	Tail-Trimmed-X	TX-Adapt ($k_n^{(x)}$)	TX-Adapt (k_n)
.25	.016, .043, .089	.010, .054, .103	.007, .055, .106	.011, .043, .099	.009, .051, .101	.013, .049, .089
1	.015, .026, .035	.011, .049, .097	.013, .047, .093	.023, .051, .084	.013, .051, .095	.020, .046, .077
2	.002, .002, .002	.013, .041, .091	.015, .050, .092	.002, .002, .002	.019, .053, .091	.012, .017, .022

Values are rejection frequencies of the null hypothesis ATE = 0, at the 1%, 5%, 10% levels. "Untrimmed" is the untrimmed estimator $\tilde{\theta}_n$. "Tail-Trimmed" is the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and "Bias-Corrected" is the bias-corrected tail-trimmed $\hat{\theta}_n^{(tz:o)}$: both use *sample mean-centering* for trimming "Tail-Trimmed-X" is $\theta_n^{(tx)}$; "TX-Adapt (k)" is the adaptive version $\hat{\theta}_n^{(tx)}$ of $\theta_n^{(tx)}$ with based on using $\gamma_n = X_{(k)}^{(a)}$. In this study $k_n^{(x)} > k_n$.

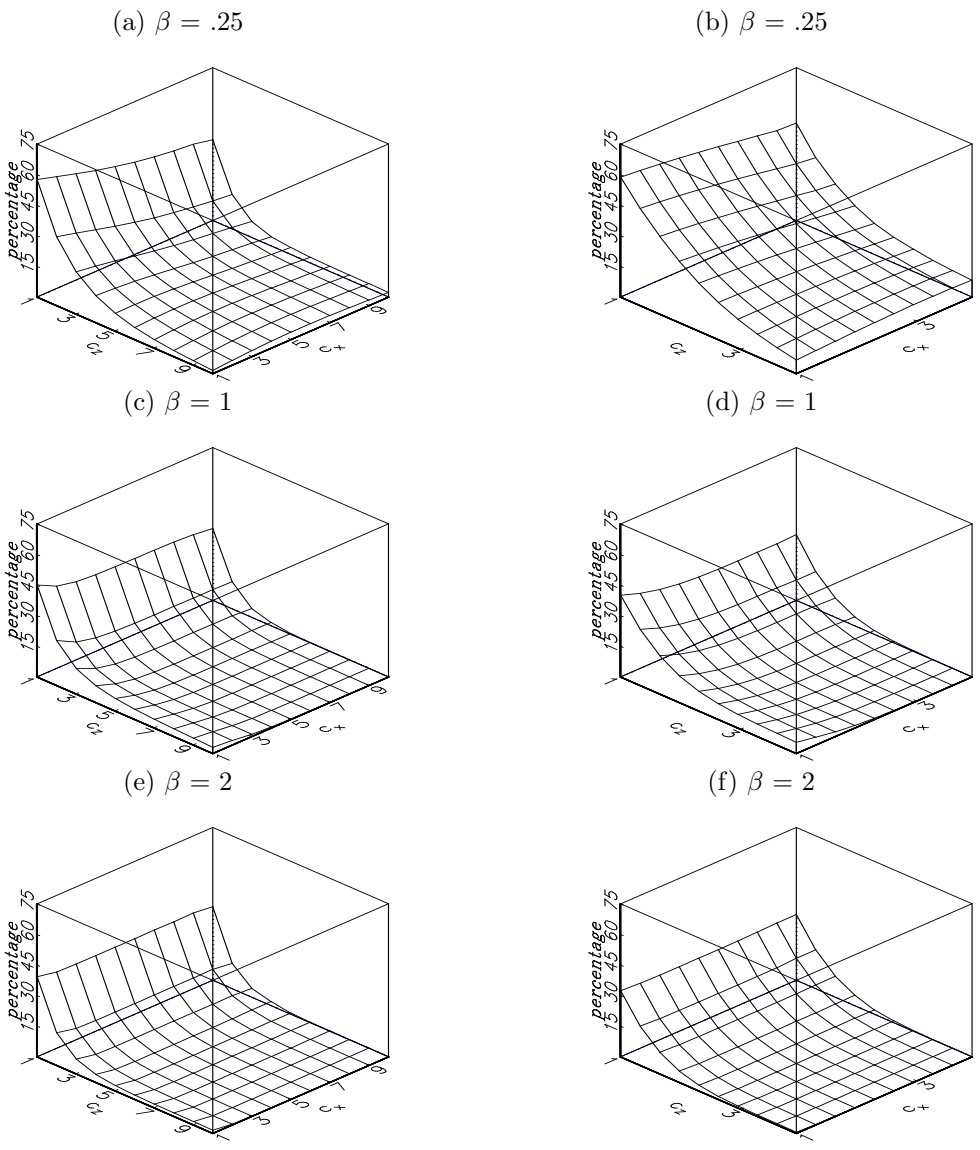

Figure 1: $P(|Z| > c_z \mid |X| > c_x)$: (Y_1, Y_2, U, X) are iid Laplace (left panels) or Normal (right panels).