Introduction

Motivation

Propensity-based sample trimming is a popular way to reduce variance by dropping units with extreme propensities. In this work, we propose trimming not only units with extreme propensities, but also units with extreme conditional variances. This both reduces variance more effectively than trimming based on propensities alone, and acts as a form of outlier removal, and which can be useful in identifying sub-populations on whom the treatment has a significantly different effect than on the full-population. We also show how to perform valid inference after sample trimming, even when nuisances are estimated non-parametrically, and give methods for constructing simultaneous CIs around multiple sub-populations.

Notation

- \( n \) units \((Y_i(1), Y_i(0), X_i, Z_i)\)
- propensity score \( e(x) = \mathbb{P}(Z_i = 1 \mid X_i = x) \)
- conditional means \( \mu_{Y_i(x)}(x) = \mathbb{E}[Y_i(w) \mid X_i = x] \)
- conditional variances \( \sigma_{Y_i(x)}^2(x) = \text{var}(Y_i(w) \mid X_i = x) \)

Heteroscedasticity-aware trimming

To trim the sample, we have to choose a one-dimensional summary \( k(x) \) of the covariates and a cut-off \( \hat{\gamma} \). Standard propensity based trimming uses \( \hat{k} = 1/(\bar{e}(x)(1 - \bar{e}(x))) \); we propose instead using

\[
\hat{k}(x) = \hat{\delta}(x) \frac{\hat{\sigma}(x)}{\bar{e}(x)} + \hat{\delta}(x) \frac{\bar{\sigma}(x)}{1 - \bar{e}(x)}
\]

There are several different possible choices for \( \hat{\gamma} \):

- pre-commit to a specific cut-off, for example, the rule of thumb to drop units with propensities outside of \((0.1, 0.9)\)
- to obtain the smallest asymptotic variance, we should take

\[
\hat{\gamma} = \min_{k(x) \leq \max \hat{k}(x)} \frac{1}{2} \sum \mathbb{1}(\hat{k}(X_i) \leq \gamma)
\]

- to only trim a \( \delta \)-fraction of sample, we can take

\[
\hat{\gamma} = 1 - \delta \text{ quantile of } \hat{k}(X_1), \ldots, \hat{k}(X_n)
\]

Inference after trimming

Prior work on inference after trimming either ignores variance introduced by the sample trimming or makes strong parametric assumptions on modeling. In contrast, we show how to obtain valid inference after sample trimming with only fourth-root rate assumptions on nuisance components.

We perform inference on a smoothed estimand with an AIPW-style estimator:

\[
\tau_{S,A}^A = \frac{1}{n} \sum_{i=1}^n \tau(X_i) \Phi(\hat{\gamma} - \hat{k}(X_i)), \quad \hat{\tau}_{S,A}^A = \frac{1}{n} \sum_{i=1}^n \hat{\tau}(X_i) \Phi(\hat{\gamma} - \hat{k}(X_i))
\]

Main theorem

If \( \hat{k} \) and \( \hat{\gamma} \) converge to \( \bar{k} \) and \( \gamma \) and the usual AIPW assumptions hold, then \( \tau_{S,A}^A - \tau_{S,A}^A \) is first-order equivalent to the i.i.d. sum

\[
\frac{1}{n} \sum_{i=1}^n (Y_i - \mu_{Y_i(x)} Z_i + Y_i(0) - \mu_{Y_i(x)} (1 - Z_i)) \Phi(\gamma - \bar{k}(X_i))
\]

where \( \gamma = \mathbb{E} [\Phi(\tau - \bar{k}(X_i))] \). This allows us to perform inference on \( \tau_{S,A}^A \) using asymptotic normal approximation or the bootstrap.

Simultaneous trimming

If we want to explore multiple trimmings and choose among them, we need simultaneously valid confidence intervals, which we can construct using a bootstrap-based algorithm:

1. Estimate \( \hat{\delta}, \hat{\sigma} \) and construct \( \hat{\bar{k}} \)
2. Set \( \hat{\gamma}_m \) as in (2) and let \( \hat{A}_m = \{x : \hat{k}(x) \leq \hat{\gamma}_m\} \)
3. Compute \( \hat{\tau}_{S,A_m} \) as well as standard errors \( \hat{\sigma}_m \)
4. for \( b = 1, \ldots , B \) do
5. Sample bootstrapped \((X_1, \tilde{Z}_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Z}_n, \tilde{Y}_n)\)
6. Re-estimate \( \tilde{k}^B \) on the bootstrapped data and compute estimates \( \hat{\tau}_{S,A_m}^B \) and standard errors \( \hat{\sigma}_m^B \)
7. Compute

\[
T_b = \max_{1 \leq m \leq n} \left| \frac{\tau_{S,A_m} - \hat{\tau}_{S,A_m}}{\hat{\sigma}_m} \right|
\]

8. end for
9. Let \( q \) be the \( 1 - \alpha \) quantile of \( T_1, \ldots, T_B \)
10. return \( \hat{\tau}_{S,A_m}^B \pm q \hat{\sigma}_m^B \)

Example: NHANES data

Figure 1. Effect of smoking on blood lead levels based the 2007-2008 U.S. National Health and Nutrition Examination Survey data, for smoothed estimand with smoothing function \( \Phi(t) = 1/(1 + \exp(-t/\epsilon)) \), comparing several different trimming methods and values of \( \epsilon \).

Example: ACIC 2022 contest data

Figure 2. Estimating effect of Medicare interventions on patient costs using ACIC 2022 contest data; simultaneous confidence intervals for sub-populations of different sizes found using (2).