Bootstrapping Sensitivity Analysis

Qingyuan Zhao

Department of Statistics, The Wharton School
University of Pennsylvania

May 23, 2018 @ ACIC

Based on:


Why sensitivity analysis?

- Unless we have perfectly executed randomized experiment, causal inference is based on some unverifiable assumptions.
- In observational studies, the most commonly used assumption is ignorability or no unmeasured confounding:

\[ A \perp \perp Y(0), Y(1) \mid X. \]

We can only say this assumption is “plausible”.
- Sensitivity analysis asks: what if this assumption does not hold? Does our qualitative conclusion still hold?
- This question appears in multiple settings:
  1. Confounded observational studies.
  2. Survey sampling with missing not at random (MNAR).
  3. Longitudinal study with non-ignorable dropout.
- In general, this means that the target parameter (e.g. average treatment effect) is only partially identified.
Overview: Bootstrapping sensitivity analysis

Point-identified parameter: Efron’s bootstrap

\[ \text{Bootstrap} \]
Point estimator $\rightarrow$ Confidence interval

Partially identified parameter: An analogy

\[ \text{Optimization} \quad \text{Percentile Bootstrap} \quad \text{Minimax inequality} \]
Extrema estimator $\rightarrow$ Confidence interval

Rest of the talk

Apply this idea to IPW estimators in a marginal sensitivity model.
Some existing sensitivity models

Generally, we need to specify how unconfoundedness is violated.

1. **Y models:** Consider a specific difference between the conditional distribution \( Y(a) \mid X, A \) and \( Y(a) \mid X \).
   - Usually called “pattern mixture models”.
   - Robins (1999, 2002); Birmingham et al. (2003); Vansteelandt et al. (2006); Daniels and Hogan (2008).

2. **A models:** Consider a specific difference between the conditional distribution \( A \mid X, Y(a) \) and \( A \mid X \).
   - Usually called “selection models”.
   - Scharfstein et al. (1999); Birmingham et al. (2003); Vansteelandt et al. (2006); Scharfstein et al. (2014).

3. **Simultaneous models:** Consider a range of A models and/or Y models and report the “worst case” result.
   - Cornfield et al. (1959); Rosenbaum (2002); Ding and VanderWeele (2016); Fogarty (2017).

**Our sensitivity model:** A hybrid of 2nd and 3rd.
Rosenbaum’s sensitivity model

▶ Imagine there is an unobserved confounder $U$ that “summarizes” all confounding, so $A \perp \perp Y(0), Y(1) \mid X, U$.

▶ Let $e_0(x, u) = \mathbb{P}_0(A = 1 \mid X = x, U = u)$.

Rosenbaum’s sensitivity model

\[ \mathcal{R}(\Gamma) = \left\{ e(x, u) : \frac{1}{\Gamma} \leq \text{OR}(e(x, u_1), e(x, u_2)) \leq \Gamma, \forall x \in X, u_1, u_2 \right\}, \]

where $\text{OR}(p_1, p_2) := [p_1/(1 - p_1)]/[p_2/(1 - p_2)]$ is the odds ratio.

▶ Rosenbaum’s question: can we reject the sharp null hypothesis $Y(0) \equiv Y(1)$ for every $e_0(x, u) \in \mathcal{R}(\Gamma)$?

▶ Robins (2002): we don’t need to assume the existence of $U$. Let $U = Y(1)$ when the goal is to estimate $E[Y(1)]$. 
Our sensitivity model

Let \( e_0(x) = \mathbb{P}_0(A = 1|X = x) \) be the propensity score.

Marginal sensitivity models

\[
\mathcal{M}(\Gamma) = \left\{ e(x, y) : \frac{1}{\Gamma} \leq OR(e(x, y), e_0(x)) \leq \Gamma, \forall x \in \mathcal{X}, y \right\}.
\]

Compare this to Rosenbaum’s model:

\[
\mathcal{R}(\Gamma) = \left\{ e(x, u) : \frac{1}{\Gamma} \leq OR(e(x, u_1), e(x, u_2)) \leq \Gamma, \forall x \in \mathcal{X}, u_1, u_2 \right\}.
\]

Tan (2006) (first?) considered this model, but he did not consider statistical inference in finite sample.

Relationship between the two models: \( \mathcal{M}(\sqrt{\Gamma}) \subseteq \mathcal{R}(\Gamma) \subseteq \mathcal{M}(\Gamma) \).

For observational studies, we assume both

\[
\mathbb{P}_0(A = 1|X, Y(1)), \mathbb{P}_0(A = 1|X, Y(0)) \in \mathcal{M}(\Gamma).
\]

\[\text{1 The second part needs “compatibility”: } e(x, y) \text{ marginalizes to } e_0(x).\]
Parametric extension

In practice, the propensity score \( e_0(X) = P_0(A = 1|X) \) is often estimated by a parametric model.

Definition (Parametric marginal sensitivity models)

\[
\mathcal{M}_{\beta_0}(\Gamma) = \left\{ e(x, y) : \frac{1}{\Gamma} \leq \text{OR}(e(x, y), e_{\beta_0}(x)) \leq \Gamma, \forall x \in \mathcal{X}, y \right\},
\]

where \( e_{\beta_0}(x) \) is the best parametric approximation of \( e_0(x) \).

This model covers both

1. Model misspecification, that is, \( e_{\beta_0}(x) \neq e_0(x) \); and
2. Missing not at random, that is, \( e_0(x) \neq e_0(x, y) \).
Logistic representations

1. **Rosenbaum’s sensitivity model:**

   \[
   \text{logit}(e(x, u)) = g(x) + \gamma u,
   \]

   where \(0 \leq U \leq 1\) and \(\gamma = \log \Gamma\).

2. **Marginal sensitivity model:**

   \[
   \text{logit}(e^{(h)}(x, y)) = \text{logit}(e_0(x)) + h(x, y),
   \]

   where \(\|h\|_\infty = \sup |h(x, y)| \leq \gamma\).

   ▶ Due to this reason we also call it marginal \(L_\infty\)-sensitivity model.

3. **Parametric marginal sensitivity model:**

   \[
   \text{logit}(e^{(h)}(x, y)) = \text{logit}(e_{\beta_0}(x)) + h(x, y),
   \]

   where \(\|h\|_\infty = \sup |h(x, y)| \leq \gamma\).
Confidence interval I

- For simplicity, consider the “missing data” problem where $Y = Y(1)$ is only observed if $A = 1$.
- Observe i.i.d. samples $(A_i, X_i, A_i Y_i), i = 1, \ldots, n$.
- The estimand is $\mu_0 = E_0[Y]$, however it is only partially identified under a simultaneous sensitivity model.

Goal 1 (Coverage of true parameter)

Construct a data-dependent interval $[L, U]$ such that

$$\mathbb{P}_0(\mu_0 \in [L, U]) \geq 1 - \alpha$$

whenever $e_0(X, Y) = \mathbb{P}_0(A = 1|X, Y) \in \mathcal{M}(\Gamma)$. 
Confidence interval II

- The inverse probability weighting (IPW) identity:

\[ \mathbb{E}_0[Y] = \mathbb{E}\left[ \frac{AY}{e_0(X, Y)} \right] \overset{MAR}{=} \mathbb{E}\left[ \frac{AY}{e_0(X)} \right]. \]

- Define

\[ \mu^{(h)} = \mathbb{E}_0\left[ \frac{AY}{e^{(h)}(X, Y)} \right] \]

- Partially identified region: \( \{ \mu^{(h)} : e^{(h)} \in \mathcal{M}(\Gamma) \} \).

Goal 2 (Coverage of partially identified region)

Construct a data-dependent interval \([L, U]\) such that

\[ \mathbb{P}_0\left( \{ \mu^{(h)} : e^{(h)} \in \mathcal{M}(\Gamma) \} \subseteq [L, U] \right) \geq 1 - \alpha. \]

- Imbens and Manski (2004) have discussed the difference between these two Goals.
An intuitive idea: “The Union Method”

- Suppose for any \( h \), we have a confidence interval \([L^{(h)}, U^{(h)}]\) such that
  \[
  \lim \inf_{n \to \infty} P_0(\mu^{(h)} \in [L^{(h)}, U^{(h)}]) \geq 1 - \alpha
  \]

- Let \( L = \inf_{\|h\|} L^{(h)} \) and \( U = \sup_{\|h\|} U^{(h)} \), so \([L, U]\) is the union interval.

**Theorem**

1. \([L, U]\) satisfies Goal 1 asymptotically.
2. Furthermore if the intervals are “congruent”: \( \exists \alpha' < \alpha \) such that

\[
\limsup_{n \to \infty} P_0(\mu^{(h)} < L^{(h)}) \leq \alpha', \quad \limsup_{n \to \infty} P_0(\mu^{(h)} > U^{(h)}) \leq \alpha - \alpha'.
\]

Then \([L, U]\) satisfies Goal 2 asymptotically.
Practical challenge: How to take the union?

- Suppose \( \hat{g}(x) \) is an estimate of \( \logit(e_0(x)) \).
- For a specific difference \( h \), we can estimate \( e^{(h)}(x, y) \) by

\[
\hat{e}^{(h)}(x, y) = \frac{1}{1 + e^{h(x,y)} - \hat{g}(x,y)}.
\]

- This leads to an (stabilized) IPW estimate of \( \mu^{(h)} \):

\[
\hat{\mu}^{(h)} = \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{\hat{e}^{(h)}(X_i, Y_i)} \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\hat{e}^{(h)}(X_i, Y_i)} \right].
\]

- Using the general Z-estimation theory,

\[
\sqrt{n} \left( \hat{\mu}^{(h)} - \mu^{(h)} \right) \xrightarrow{d} N(0, (\sigma^{(h)})^2)
\]

- Therefore we can use \( [L^{(h)}, U^{(h)}] = \hat{\mu}^{(h)} \mp z_{\alpha/2} \cdot \hat{\sigma}^{(h)} / \sqrt{n} \).

- However, computing the union interval requires solving a complicated optimization problem.
Bootstrapping sensitivity analysis

Point-identified parameter: Efron’s bootstrap

\[ \text{Bootstrap} \]

Point estimator \rightarrow \text{Confidence interval}

Partially identified parameter: An analogy

\[ \text{Optimization} \quad \text{Percentile Bootstrap} \quad \text{Minimax inequality} \]

Extrema estimator \rightarrow \text{Confidence interval}

A simple procedure for simultaneous sensitivity analysis

1. Generate \( B \) random resamples of the data. For each resample, compute the extrema of IPW estimates under \( \mathcal{M}_{\beta_0}(\Gamma) \).

2. Construct the confidence interval using \( L = Q_{\alpha/2} \) of the \( B \) minima and \( U = Q_{1-\alpha/2} \) of the \( B \) maxima.

Theorem

\([L, U]\) achieves Goal 2 for \( \mathcal{M}_{\beta_0}(\Gamma) \) asymptotically.
Proof of the Theorem

Partially identified parameter: Three ideas

**Optimization**

1. **Percentile Bootstrap**
2. **Minimax inequality**

Extrema estimator $\Rightarrow$ Confidence interval

1. The sampling variability of $\hat{\mu}^{(h)}$ can be captured by bootstrap. The percentile bootstrap CI is given by

$$
\left[ Q_{\frac{\alpha}{2}} \left( \hat{\mu}^{(h)}_b \right), \ Q_{1-\frac{\alpha}{2}} \left( \hat{\mu}^{(h)}_b \right) \right].
$$

2. Generalized minimax inequality:

$$
Q_{\frac{\alpha}{2}} \left( \inf_{h} \hat{\mu}^{(h)}_b \right) \leq \inf_{h} Q_{\frac{\alpha}{2}} \left( \hat{\mu}^{(h)}_b \right) \leq \sup_{h} Q_{1-\frac{\alpha}{2}} \left( \hat{\mu}^{(h)}_b \right) \leq Q_{1-\frac{\alpha}{2}} \left( \sup_{h} \hat{\mu}^{(h)}_b \right).
$$

Union CI
Computation

Partially identified parameter: Three ideas

3. **Optimization**
   Extrema estimator

3. **Percentile Bootstrap**

3. **Minimax inequality**

\[ \Rightarrow \] Confidence interval

3. Computing extrema of \( \hat{\mu}^{(h)} \) is a **linear fractional programming**:
   Let \( z_i = e^{h(X_i, Y_i)} \), we just need to solve
   \[
   \max \quad \min \quad \frac{\sum_{i=1}^{n} A_i Y_i \left(1 + z_i e^{-\hat{g}(X_i)}\right)}{\sum_{i=1}^{n} A_i \left(1 + z_i e^{-\hat{g}(X_i)}\right)},
   \]
   subject to \( z_i \in [\Gamma^{-1}, \Gamma], \quad i = 1, \ldots, n. \)

- This can be converted to a linear programming.
- Moreover, the solution \( z \) must have the same/opposite order as \( Y \), so the time complexity can be reduced to \( O(n) \) (optimal).

The role of Bootstrap

Comapred to the union method, the workflow is greatly simplified:

1. No need to **derive** \( \sigma^{(h)} \) analytically (though we could).
2. No need to **optimize** \( \sigma^{(h)} \) (which is very challenging).
Comparison with Rosenbaum’s sensitivity analysis

<table>
<thead>
<tr>
<th></th>
<th>Rosenbaum’s paradigm</th>
<th>New bootstrap approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td>Sample</td>
<td>Super-population</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>Matching</td>
<td>Weighting</td>
</tr>
<tr>
<td><strong>Sensitivity model</strong></td>
<td>$\mathcal{M}(\sqrt{\Gamma}) \subseteq \mathcal{R}(\Gamma) \subseteq \mathcal{M}(\Gamma)$</td>
<td></td>
</tr>
<tr>
<td><strong>Inference</strong></td>
<td>Bounding $p$-value</td>
<td>CI for ATE/ATT</td>
</tr>
<tr>
<td><strong>Effect modification</strong></td>
<td>Constant effect</td>
<td>Allow for heterogeneity</td>
</tr>
<tr>
<td><strong>Extension</strong></td>
<td>Carefully developed for observational studies</td>
<td>Can be applied to missing data problems</td>
</tr>
</tbody>
</table>
Example

Fish consumption and blood mercury

- 873 controls: \( \leq 1 \) serving of fish per month.
- 234 treated: \( \geq 12 \) servings of fish per month.
- Covariates: gender, age, income (very imbalanced), race, education, ever smoked, \# cigarettes.

Implementation details

- Rosenbaum’s method: 1-1 matching, CI constructed by Hodges-Lehmann (assuming causal effect is constant).
- Our method (percentile Bootstrap): stabilized IPW for ATT w/wo augmentation by outcome linear regression.
Results

- Recall that $\mathcal{M}(\sqrt{\Gamma}) \subseteq \mathcal{R}(\Gamma) \subseteq \mathcal{M}(\Gamma)$.

**Figure:** The solid error bars are the range of point estimates and the dashed error bars (together with the solid bars) are the confidence intervals. The circles/triangles/squares are the mid-points of the solid bars.
References I


Possible open problems for discussion

- We didn’t require $e^{(h)}(x, y)$ to be “compatible” (marginalizes to $e_0(x)$). Can we use this property to reduce the length of CI?
  - Not a problem for Rosenbaum’s analysis because he conditions on one treated in every pair.
  - Compatibility between A model and Y model?

- Can we consider different sensitivity models?
  - Example: require $h(x, y)$ to be Lipschitz continuous. Considered in the paper, still a LFP.
  - Can we do “average case” sensitivity analysis, e.g. $L_1$ instead of $L_\infty$?

- How can we extend this framework to other semiparametric estimation problems (e.g. Longitudinal data with non-ignorable dropout)?