Sharp Sensitivity Analysis for Inverse Propensity Weighting via Quantile Balancing

Jacob Dorn* and Kevin Guo

*Department of Economics, Princeton University, Princeton, NJ; bDepartment of Statistics, Stanford University, Stanford, CA

ABSTRACT

Inverse propensity weighting (IPW) is a popular method for estimating treatment effects from observational data. However, its correctness relies on the untestable (and frequently implausible) assumption that all confounders have been measured. This article introduces a robust sensitivity analysis for IPW that estimates the range of treatment effects compatible with a given amount of unobserved confounding. The estimated range converges to the narrowest possible interval (under the given assumptions) that must contain the true treatment effect. Our proposal is a refinement of the influential sensitivity analysis by Zhao, Small, and Bhattacharya, which we show gives bounds that are too wide even asymptotically. This analysis is based on new partial identification results for Tan’s marginal sensitivity model. Supplementary materials for this article are available online.

1. Introduction

Estimating treatment effects from observational data is difficult because “treated” and “control” samples typically differ on many characteristics besides treatment status. For example, consumers of nutritional supplements may be wealthier or more health-conscious than those not taking supplements. One popular tool for adjusting for such baseline imbalances is Inverse Propensity Weighting (IPW) (Hirano and Imbens 2002; Austin and Stuart 2015). This technique reweights treated and untreated samples to be similar along all observed characteristics and then compares outcomes in the weighted samples. The crucial assumption underlying this approach is that the weighted samples do not systematically differ along important unobserved characteristics. This “unconfoundedness” assumption is untestable, and often implausible.

This article studies how much can be learned when unconfoundedness does not hold, but one can bound the plausible degree of unobserved confounding. In particular, given a “sensitivity assumption” controlling the degree of selection, we aim to answer two questions:

1. Sensitivity analysis. Can we bound how much the IPW point estimate from our “primary analysis” might change if unobserved confounding were properly accounted for?
2. Partial identification. Can we characterize the most informative bounds that could possibly be obtained from the sensitivity assumption with even an infinite amount of observational data?

The specific sensitivity assumption used in this article is the “marginal sensitivity model” of Tan (2006), which is a variant of Rosenbaum’s famous “I” sensitivity model (Rosenbaum 1987, 2002, 2010) that is better suited for IPW analyses. This sensitivity assumption is quite popular in causal inference; see Tan (2006), Kallus and Zhou (2018), Kallus and Zhou (2020a, 2020b), Kallus, Mao, and Zhou (2019), Zhao, Small, and Bhattacharya (2019), Lee, Bargagli-Stoffi, and Dominici (2020), Rosenman et al. (2020), Rosenman and Owen (2021), and Soriano et al. (2021) for an incomplete list of references. As we will see, it lends itself to computationally-efficient sensitivity analyses which are simple enough to explain to any practitioner comfortable with IPW.

Recently, Zhao, Small, and Bhattacharya (2019) (hereafter ZSB) introduced an interpretable IPW sensitivity analysis for the marginal sensitivity model that has been largely responsible for the recent resurgence of interest in this sensitivity assumption. However, they did not answer the partial identification question, leaving open the possibility that more informative bounds could be obtained from the same data and assumptions. Indeed, there are no existing partial identification results for the marginal sensitivity model that can be used to benchmark a sensitivity analysis.

The first main contribution of this article is to provide a complete answer to the partial identification question (2). We derive closed-form expressions for the largest and smallest values of the “usual” estimands (e.g., average treatment effect) compatible with the marginal sensitivity assumption. These expressions show that the ZSB bounds are essentially always conservative because they ignore an infinite collection of constraints implied by the distribution of observed characteristics. Tan (2006) also identified these constraints, but deemed it intractable to incorporate them all in a sensitivity analysis. In contrast, our partial identification results show that this collection can actually be reduced to a single constraint which is easy to incorporate.

Our second main contribution is to introduce a new IPW sensitivity analysis, which we call the quantile balancing
method. The method is a simple refinement of the ZSB sensitivity analysis, and has several desirable features:

(i) The quantile balancing sensitivity interval is always a subset of the ZSB interval. Outside of knife-edge cases, it is a strict subset.
(ii) When the outcome's conditional quantiles can be estimated consistently, the bounds converge to the sharp partial identification region for the average treatment effect (the best possible bounds that can be obtained under the marginal sensitivity model). With some abuse of terminology, we say that quantile balancing is "sharp."
(iii) Under standard assumptions for IPW inference, the bounds can be converted into confidence intervals using the same percentile bootstrap scheme proposed by ZSB.
(iv) When the estimated quantiles are inconsistent, the sensitivity interval is too wide rather than too narrow and the confidence intervals over-cover rather than under-cover. In other words, our intervals are guaranteed to be valid, regardless of the quality of the additional input we demand.

We apply the quantile balancing method in several simulated examples and one real-data application, and find that it can substantially tighten the ZSB bounds when the covariates are good predictors of the outcome. We also extend our analysis to Augmented IPW (AIPW) estimators. That analysis shows that a slight refinement of the ZSB method is sharp under "additive-noise" data generating processes, though the refinement makes little difference in practice. One shortcoming we will mention up-front is that our statistical guarantees assume the outcome is continuously-distributed in order to enable quantile regression. Since our partial identification results also apply to discrete outcomes, we conjecture that the quantile balancing procedure could be modified to give sharp bounds in that setting too.

1.1. Setting and Background

We consider the Neyman–Rubin potential outcomes model with a binary treatment (Neyman 1923; Rubin 1974). We observe iid samples \((X_i, Y_i, Z_i)\) from a distribution \(P\), where \(X_i \in \mathcal{X} \subseteq \mathbb{R}^d\) is a vector of covariates, \(Z_i \in \{0,1\}\) is a binary treatment assignment indicator, and \(Y_i \in \mathbb{R}\) is a real-valued outcome.

We assume that each sample \((X_i, Y_i, Z_i)\) is obtained by coarsening a “full data” sample \((X_i, Y_i(0), Y_i(1), Z_i, U_i)\). Here, \(Y_i(0)\) and \(Y_i(1)\) are potential outcomes and \(U_i\) is a vector of unobserved confounders of unspecified dimension. The observed outcome is related to the potential outcomes through the consistency relation \(Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)\).

The goal is to use the observed data to draw inferences about a causal estimand \(\psi_0\). For the purposes of exposition, we initially focus on the counterfactual means \(\psi_T = \mathbb{E}[Y(1)]\) and \(\psi_C = \mathbb{E}[Y(0)]\), although the examples of most practical interest are the average treatment effect (ATE) and the average treatment effect on the treated (ATT):

\[
\psi_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)] \\
\psi_{\text{ATT}} = \mathbb{E}[Y(1) - Y(0) | Z = 1].
\]

With minor modification, our identification results can also be applied to more complex estimands, including policy values (Kallus and Zhou 2018; Athey and Wager 2021) and weighted average treatment effects. However, we do not present those extensions in this article.

Under the unconfoundedness assumption \((Y(0), Y(1)) \perp\!
\!
\perp Z | X\), all of the above quantities can be consistently estimated from the observed data using inverse propensity weighting. IPW estimators work by reweighting the observed sample by some function of the propensity score \(c(x) := P(Z = 1 | X = x)\). For example, if the estimand of interest is \(\psi_T\), the (stabilized) IPW estimator is given by (1):

\[
\hat{\psi}_T = \frac{\mathbb{E}_n[Y_i Z_i / \hat{c}(X_i)]}{\mathbb{E}_n[Z / \hat{c}(X_i)]}, \tag{1}
\]

Here, \(\hat{c}(\cdot)\) is an estimate of the propensity score \(c(\cdot)\) and \(\mathbb{E}_n[\cdot]\) is shorthand for \(\frac{1}{n} \sum_i \cdot\). An unstabilized version of \(\hat{\psi}_T\) which uses only the numerator of (1) is also common. Related estimators for the other estimands considered will be denoted by \(\hat{\psi}_C, \hat{\psi}_{\text{ATE}},\) and \(\hat{\psi}_{\text{ATT}}\). See the articles by Austin and Stuart (2015) or Hirano and Imbens (2002) for their exact formulas.

We will assume some conditions which are required for identification and estimation under unconfoundedness: overlap \((0 < c(X) < 1\) almost surely) and one outcome moment \((\mathbb{E}_P[Y] < \infty)\). However, we will not assume unconfoundedness.

2. The Marginal Sensitivity Model

The marginal sensitivity model introduced by Tan (2006) is a relaxation of unconfoundedness which has been applied in many causal inference problems. This one-parameter sensitivity assumption allows for the existence of unobserved confounders \(U\), but limits the degree of selection bias that can be attributed to these confounders.

**Assumption \(\Lambda\) (Marginal sensitivity model).** There exists a vector of unmeasured confounders \(U\) that, if measured, would lead to unconfoundedness: \((Y(0), Y(1)) \perp\!
\!
\perp Z | (X, U)\). However, within each stratum of the observed covariates, measuring \(U\) can only change the odds of treatment by at most a factor of \(\Lambda\), that is, if we set \(e_0(x, u) := P(Z = 1 | X = x, U = u)\), then (2) holds with probability one:

\[
\Lambda^{-1} \leq \frac{e_0(X, U) / [1 - e_0(X, U)]}{c(X) / [1 - c(X)]} \leq \Lambda. \tag{2}
\]

The statement of the marginal sensitivity model presented in Tan (2006) and Zhao, Small, and Bhattacharya (2019) uses the potential outcomes \((Y(0), Y(1))\) in place of the unobserved variable \(U\). However, as pointed out by a referee, these assumptions are equivalent.

To avoid confusion between \(e_0\) and \(e\), we will follow Kallus and Zhou (2020b) and refer to \(e_0\) as the "true propensity score" and \(e\) as the "nominal propensity score."

Like Rosenbaum’s famous “\(\Gamma\) sensitivity model,” **Assumption \(\Lambda\)** controls the degree of unobserved confounding with a single parameter. When \(\Lambda = 1\), measuring additional confounders cannot change the odds of treatment at all, that is, treatment assignment is unconfounded. As \(\Lambda\) increases, stronger forms of confounding are allowed. For advice on how
to choose this parameter, see Hsu and Small (2013). For more on the relationship between this and Rosenbaum's model, see Zhao, Small, and Bhattacharya (2019, sec. 7.1). The marginal sensitivity assumption is "nonparametric" in the sense that no assumptions are needed about how \( e_0 \) depends on \( u \). Even the dimension of the vector \( U \) does not need to be specified.

To see how Assumption \( \Lambda \) can be used for sensitivity analysis, begin by considering how an oracle statistician who observed the confounders \( U_i \) might estimate \( \psi_T \). One strategy would be to use the IPW estimator (3), which is consistent under weak assumptions

\[
\psi_T^* = \frac{\sum_{i=1}^{n} Y_i Z_i / e_0(X_i, U_i)}{\sum_{i=1}^{n} Z_i / e_0(X_i, U_i)}.
\]

In reality, \( \{U_i\}_{i \in n} \) are not observed, but under Assumption \( \Lambda \), it is possible to bound the true propensity score \( e_0(X_i, U_i) \). In particular, the vector \( (e_0(X_1, U_1), \ldots, e_0(X_n, U_n)) \) must belong to the ZSB constraint set \( \mathcal{E}_n(\Lambda) \) defined in (4)

\[
\mathcal{E}_n(\Lambda) = \left\{ \bar{e} \in \mathbb{R}^n : \Lambda^{-1} \leq \bar{e}/(1 - \bar{e}) e(X_i)/[1 - e(X_i)] \leq \Lambda \right\}.
\]

ZSB proposed bounding the oracle statistician’s IPW estimator (3) with the largest and smallest IPW estimates that can be obtained using putative propensities in \( \mathcal{E}_n(\Lambda) \)

\[
[\hat{\psi}_{T,ZSB}^-, \hat{\psi}_{T,ZSB}^+] = \left[ \min_{e \in \mathcal{E}_n(\Lambda)} \sum_{i=1}^{n} Y_i Z_i / \bar{e}_i, \max_{e \in \mathcal{E}_n(\Lambda)} \sum_{i=1}^{n} Y_i Z_i / \bar{e}_i \right].
\]

Since the interval (5) contains the consistent estimator \( \psi_T^* \), the distance between the true estimand \( \psi_T \) and the nearest point in the sensitivity interval tends to zero. ZSB show that this conclusion holds even if the nominal propensity score \( e \) is replaced by a suitably consistent estimate \( \bar{e} \) in the definition of \( \mathcal{E}_n(\Lambda) \), which is important for practical applications as \( e \) is typically not known in observational studies.

This simple idea is intuitive enough to explain to any practitioner who is comfortable with IPW and has been extended to the confounders \( U_i \) might estimate \( \psi_T \). One strategy would be to use the IPW estimator (3), which is consistent under weak assumptions

\[
\psi_T^* = \frac{\sum_{i=1}^{n} Y_i Z_i / e_0(X_i, U_i)}{\sum_{i=1}^{n} Z_i / e_0(X_i, U_i)}.
\]

In reality, \( \{U_i\}_{i \in n} \) are not observed, but under Assumption \( \Lambda \), it is possible to bound the true propensity score \( e_0(X_i, U_i) \). In particular, the vector \( (e_0(X_1, U_1), \ldots, e_0(X_n, U_n)) \) must belong to the ZSB constraint set \( \mathcal{E}_n(\Lambda) \) defined in (4)

\[
\mathcal{E}_n(\Lambda) = \left\{ \bar{e} \in \mathbb{R}^n : \Lambda^{-1} \leq \bar{e}/(1 - \bar{e}) e(X_i)/[1 - e(X_i)] \leq \Lambda \right\}.
\]

ZSB proposed bounding the oracle statistician’s IPW estimator (3) with the largest and smallest IPW estimates that can be obtained using putative propensities in \( \mathcal{E}_n(\Lambda) \)

\[
[\hat{\psi}_{T,ZSB}^-, \hat{\psi}_{T,ZSB}^+] = \left[ \min_{e \in \mathcal{E}_n(\Lambda)} \sum_{i=1}^{n} Y_i Z_i / \bar{e}_i, \max_{e \in \mathcal{E}_n(\Lambda)} \sum_{i=1}^{n} Y_i Z_i / \bar{e}_i \right].
\]

Since the interval (5) contains the consistent estimator \( \psi_T^* \), the distance between the true estimand \( \psi_T \) and the nearest point in the sensitivity interval tends to zero. ZSB show that this conclusion holds even if the nominal propensity score \( e \) is replaced by a suitably consistent estimate \( \bar{e} \) in the definition of \( \mathcal{E}_n(\Lambda) \), which is important for practical applications as \( e \) is typically not known in observational studies.

This simple idea is intuitive enough to explain to any practitioner who is comfortable with IPW and has been extended to the confounders other than \( \psi_T \). ZSB also consider \( \psi_{ATE} \) and \( \psi_{ATT} \). Related work by Kallus and Zhou (2018), Kallus and Zhou (2020a), Kallus, Mao, and Zhou (2019), and Lee, Bargagli-Stoffi, and Dominici (2020) takes the idea substantially further. Tan (2006) applied a similar idea to a different propensity-score-based estimator and Aronow and Lee (2013), Miratrix, Wager, and Zubizarreta (2018), and Tudball et al. (2019) used similar approaches in survey sampling problems.

### 2.1. Sharpness and Data-Compatibility

The aforementioned works do not address the asymptotic optimality of the interval \([\hat{\psi}_{T,ZSB}^-, \hat{\psi}_{T,ZSB}^+]\). Does it converge to a limiting set containing all values of \( \psi_T \) compatible with Assumption \( \Lambda \) and no others? Sensitivity analyses with this asymptotic optimality property are called "sharp" in the partial identification literature.

Sharpness is important for interpreting the results of a sensitivity analysis. If the primary analysis finds a positive treatment effect but the bounds associated with a very small value of \( \Lambda \) include zero, one might be tempted to conclude that the primary analysis is sensitive to unobserved confounding. However, unless the bounds are known to be sharp, this inference is not warranted even in large samples. Perhaps the bounds were just too conservative.

Despite its attractive features, the ZSB sensitivity analysis is not sharp. It can be arbitrarily conservative. To illustrate this, consider a simple joint distribution of observables:

\[
X \sim \mathcal{N}(0, \sigma^2)
\]

\[
Z \mid X \sim \text{Bernoulli}(\frac{1}{2})
\]

\[
Y \mid X, Z \sim \mathcal{N}(X, 1).
\]

Suppose that a data analyst receives iid samples \((X_i, Y_i, Z_i)\) from this distribution and is willing to posit that Assumption \( \Lambda \) is satisfied with \( \Lambda = 2 \). Let \( \phi(\cdot) \) and \( z_\tau \) denote the density and \( \tau \) th quantile of the standard normal distribution, respectively. The following result, which follows from Theorem 2 in Section 3.1, writes the set of values of \( \psi_T \) compatible with Assumption \( \Lambda \) explicitly in terms of these quantities and shows that this "partially identified" set is smaller than the limiting ZSB interval.

**Corollary 1 (ZSB is asymptotically conservative).** Let \((X_i, Y_i, Z_i)\) be iid samples from the joint distribution (6).

1. The set of values of \( \psi_T \) compatible with the bound \( \Lambda = 2 \) and the distribution (6) is the interval \([\pm \phi(z_\tau/3)] \approx [\pm 0.27] \).

2. However, with probability one, \([\pm 0.27 \sqrt{\sigma^2 + 1}] \subseteq [\hat{\psi}_{T,ZSB}^-, \hat{\psi}_{T,ZSB}^+] \) for all large \( n \).

The precise meaning of (i) is the following: for any \( \psi_T \in [\pm \phi(z_\tau/3)] \), it is possible to construct a distribution \( Q \) for the full data \((X, Y(0), Y(1), Z, U)\) which marginalizes to (6), satisfies Assumption \( \Lambda \) with \( \Lambda = 2 \), and has \( \mathbb{E}_Q[Y(1)] = \psi_T \). On the other hand, for any \( \psi_T \) not in this interval, it is impossible to construct such a distribution.

Corollary 1 implies that the ZSB interval typically includes many values of \( \psi \) which cannot possibly be reconciled with the data. The explanation for this conservatism is that the odds-ratio bound (2) does not capture all of the restrictions on the true propensity score \( e_0 \). Additional information can be found in the marginal distribution of the observed characteristics. For example, in the context of Corollary 1, consider the putative propensity score (7):

\[
\bar{e}(x, u) = \begin{cases} 1/3 & \text{if } x < 0 \\ 2/3 & \text{if } x \geq 0 \end{cases}
\]

This certainly satisfies the odds-ratio bound (2)—and is therefore a possible value of \( \bar{e} \) in the ZSB optimization problem (5)—but it could not possibly be the true propensity score \( e_0 \). If it were, we would observe \( P(Z = 1 | X \geq 0) = \frac{2}{3} \), while the observed data distribution \( P \) demands that \( P(Z = 1 | X \geq 0) = \frac{3}{4} \). Another way of saying this is that \( \bar{e} \) does not marginalize to the nominal propensity score:

\[
1/2 = P(Z = 1 | X = x)
\]

\[
= \int P(Z = 1 | X = x, U = u) \, dP(u | X = x)
\]
\[ \neq \int \bar{e}(x, u) \, dP(u | X = x) \]
\[ = \begin{cases} 
1/3 & \text{if } x < 0 \\
2/3 & \text{if } x \geq 0 
\end{cases}. \]

In short, this choice of \( \bar{e} \) is allowed in the domain of the ZSB optimization problem but is incompatible with the distribution of observed data.

This example suggests that it should be possible to improve upon the ZSB bounds by only optimizing over the subset of \( \mathcal{E}_n(\Lambda) \) which is "data compatible." However, this is easier said than done, because the observed data distribution actually imposes an infinite number of constraints on putative propensity scores \( \bar{e} \). For example, the true \( e_0 \) "balances" all integrable functions \( h : \mathcal{X} \to \mathbb{R} \):

\[ \mathbb{E}[h(X)Z/e_0(X, U)] = \mathbb{E}[h(X)\mathbb{E}[Z|X, U]/e_0(X, U)] \]
\[ = \mathbb{E}[h(X)e_0(X, U)/e_0(X, U)] \]
\[ = \mathbb{E}[h(X)]. \]

Every such \( h \) gives rise to a testable "balancing constraint" \( \psi \) which can be used to rule out incompatible values of \( \bar{e} \):

\[ \frac{\mathbb{E}_n[h(X)Z/\bar{e}]}{\mathbb{E}_n[Z/\bar{e}]} \approx \mathbb{E}[h(X)]. \]

In other words, any sharp sensitivity analysis must contend with an infinite number of constraints, which is typically computationally intractable (Beresteanu, Molchanov, and Molinari 2011; Davezies and D'Haultfoeuille 2016). Previous works have considered relaxing these constraints by balancing only a finite set of functions (Tan 2006; Tudball et al. 2019), but the resulting bounds are generally not sharp.

While this article proceeds under the "superpopulation" model of causal inference, the idea that observable quantities can constrain unobserved variables can also be applied in the "finite population" model. See Tudball et al. (2019) for an application of this idea to partial identification in survey sampling problems.

### 3. Partial Identification Results

In this section, we show that at the population level, it is possible to characterize the sharp bounds for \( \psi_0 \in \{ \psi_T, \psi_C, \psi_{ATT}, \psi_{ATE} \} \) without ignoring or relaxing any of the infinitely many balancing constraints in the true propensity score. We apply these partial identification results to finite-sample sensitivity analysis in Section 4.

To state these results formally, we need a few pieces of additional notation. Recall that Assumption \( \Lambda \) requires the true propensity score \( e_0(X, U) \) to satisfy the following odds-ratio bound:

\[ \Lambda^{-1} \leq \frac{e_0(X, U)/[1 - e_0(X, U)]}{e(X)/[1 - e(X)]} \leq \Lambda. \]

Therefore, it is natural to define \( \mathcal{E}_\infty(\Lambda) \) to be the set of all random variables \( \bar{E} \) which satisfy the same condition:

\[ \mathcal{E}_\infty(\Lambda) := \left\{ \bar{E} : \Lambda^{-1} \leq \frac{\bar{E}(1 - \bar{E})}{e(X)(1 - e(X))} \leq \Lambda \text{ with probability one} \right\}. \]

This can be viewed as the "population" version of the ZSB constraint set \( \mathcal{E}_n(\Lambda) \).

Additionally, we define the conditional distribution function \( F(y|x, z) \) and quantile function \( Q_t(x, z) \) by:

\[ F(y|x, z) = P(Y \leq y \mid X = x, Z = z) \]
\[ Q_t(x, z) = \inf(q \in \mathbb{R} : F(q|x, z) \geq t). \]

Since these functions only refer to observed quantities, they are identified from the observed-data distribution.

#### 3.1. Partial Identification via Quantile Balancing

Our first partial identification result shows that to compute optimal bounds for \( \psi_T \), the infinitely-many balancing constraints described in Section 2.1 can actually be reduced to a single constraint. In particular, it suffices to minimize/maximize the function \( \bar{E} \rightarrow \mathbb{E}[YZ/\bar{E}] \) over the set of putative propensity scores \( \bar{E} \in \mathcal{E}_\infty(\Lambda) \) that "balance" a particular conditional quantile of \( Y \).

**Theorem 1 (Optimal bounds for \( \psi_T \)).** For any \( \Lambda \geq 1 \), the set of values of \( \psi_T \) compatible with the observed data distribution and Assumption \( \Lambda \) is a closed interval \[ \psi_T^- , \psi_T^+ \]. Moreover, if we define \( \tau = \frac{\Lambda}{\Lambda + 1} \), then the interval endpoints solve (11) and (12).

\[ \psi_T^- = \min_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \mathbb{E}[YZ/\bar{E}] \text{ subject to } \mathbb{E}[Q_{\tau - \tau}(X, 1)Z/\bar{E}] = \mathbb{E}[Q_{\tau - \tau}(X, 1)] \]
\[ \psi_T^+ = \max_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \mathbb{E}[YZ/\bar{E}] \text{ subject to } \mathbb{E}[Q_\tau(X, 1)Z/\bar{E}] = \mathbb{E}[Q_\tau(X, 1)]. \]

We will highlight a few important takeaways from this theorem. First, if one adds additional balancing constraints of the form \( \mathbb{E}[h(X)Z/\bar{E}] = \mathbb{E}[h(X)] \) in (11) and (12), the value of these problems will not change. Thus, for the purposes of computing population-level bounds, the quantile balancing constraints in Theorem 1 capture all the information in the observed data. Second, the fact that only a single conditional quantile appears in each of the sharp bounds for \( \psi_T \) reflects a special advantage of the marginal sensitivity model. For alternative sensitivity assumptions, sharp bounds often involve distinct quantiles \( Q_t(x) \) for each covariate level (Lee 2009; Masten and Poirier 2018), complicating estimation by potentially requiring estimates of the entire conditional quantile process (Masten, Poirier, and Zhang 2020; Semenova 2020). Third, this result shows that the ZSB sensitivity analysis for IPW can only be sharp when the conditional quantiles of \( Y \) do not depend on \( X \) at all, and can therefore be refined outside pathological cases. AIPW-based variants of the ZSB sensitivity analysis will generally refine the IPW bounds since some of the variability in the quantiles of \( Y \) will be absorbed by the regression function. We discuss AIPW sensitivity analysis in Section 4.2.

We can extend the theorem to other estimands. To bound \( \psi_C \), exchange the labels "treated" and "control" and apply Theorem 1. Sharp bounds on \( \psi_{ATT} \) can be translated into sharp bounds on \( \psi_{ATT} \) using the relation \( \psi_{ATT} = \frac{\mathbb{E}[Y|\psi_C - \psi_C]}{P(Z = 1)} \).
Corollary 2 (Optimal bounds for $\psi_C$ and $\psi_{ATT}$). In the setting of Theorem 1, the partially identified set for $\psi_C$ is the interval $[\psi_C^-, \psi_C^+]$, where the interval endpoints solve (13) and (14):

$$
\psi_C^- = \min_{E \in \mathcal{E}_C(\Lambda)} \mathbb{E}[Y - Z] \quad \text{subject to} \quad \mathbb{E}[Q_t - \tau(X, 0)] - Z = \mathbb{E}[Q_t - \tau(X, 0)]
$$

(13)

$$
\psi_C^+ = \max_{E \in \mathcal{E}_C(\Lambda)} \mathbb{E}[Y - Z] \quad \text{subject to} \quad \mathbb{E}[Q_t - \tau(X, 0)] - Z = \mathbb{E}[Q_t - \tau(X, 0)]
$$

(14)

The partially identified set for $\psi_{ATT}$ is the interval $[\psi_{ATT}^-, \psi_{ATT}^+]$, where $\psi_{ATT} = \mathbb{E}[Y - \psi_C^+ / \psi_C^-]$, subject to $\mathbb{E}[\bar{Y} - \psi_C^+] < \mathbb{E}[\bar{Y} - \psi_C^-]$.

Sharp bounds for $\psi_{ATE}$ can be obtained by subtracting sharp bounds for $\psi_T$ and $\psi_C$. Equivalently, these bounds can be obtained by solving optimization problems with two quantile balancing constraints. Although this result is superficially similar to Theorem 1 and Corollary 2, its proof requires a novel construction, which we discuss in Section 3.3.

Theorem 2 (Optimal bounds for $\psi_{ATE}$). For any $\Lambda \geq 1$, the set of values of $\psi_{ATE}$ compatible with the observed data distribution and Assumption $\Lambda$ is a closed interval $[\psi_{ATE}^-, \psi_{ATE}^+]$ where

$$
\psi_{ATE}^+ = \psi_T^+ - \psi_C^- \quad \text{and} \quad \psi_{ATE}^- = \psi_T^- - \psi_C^+.
$$

In certain special cases, the partially identified set for $\psi_{ATE}$ can be computed more explicitly. These explicit bounds are useful for gaining intuition about the main factors that make a causal estimate more or less robust to unobserved confounding. Corollary 3, which is a corollary of our later work, gives such bounds in the Gaussian outcome model (15)

$$
\begin{align*}
X &\sim P_X \\
| X &\sim \text{Bernoulli}(e(X)) \\
Y &\sim N(\mu(X, Z), \sigma^2(X)).
\end{align*}
$$

3.2. Data-Compatible Propensity Scores

Although the qualitative implications of Corollary 3 are plausible, we nevertheless find the quantile balancing formulas of Section 3.1 to be counterintuitive. After all, it is certainly not true that every random variable $\bar{E} \in \mathcal{E}_C(\Lambda)$ satisfying $\mathbb{E}[Q_t - \tau(X, 1)/\bar{E}] = \mathbb{E}[Q_t - \tau(X, 1)]$ could plausibly be the true propensity score $e_0(X, U)$. Indeed, the constraints of the quantile-balancing optimization problems do not even enforce that $\mathbb{E}[Z/\bar{E}] = 1$. In our intuition for why the ZSB procedure is conservative suggests the quantile balancing formulas should be conservative as well.

To explain how these results are possible, we begin by characterizing which random variables $\bar{E}$ could plausibly be the true propensity score $e_0(X, U)$. The calculation (8) indicates that $\bar{E}$ should at least satisfy $\mathbb{E}[h(X)Z/\bar{E}] = \mathbb{E}[h(X)]$ for all integrable $h$, or equivalently, $\mathbb{E}[Z/\bar{E}] = 1$. Proposition 1 shows that for the purposes of bounding $\psi_T$, this is actually the only constraint on $\bar{E}$ implied by the distribution of observables. Similar results appear in Birmingham, Rotnitzky, and Fitzmaurice (2003), Robins, Rotnitzky, and Scharfstein (2000), Tan (2006), Graham (2011), Hristache and Patilea (2017), Franks, Amour, and Feller (2020), and Zhao, Small, and Bhattacharya (2019).

Proposition 1 (Characterizing data-compatible propensity scores). For any random variable $\bar{E} \in \mathcal{E}_C(\Lambda)$ satisfying $\mathbb{E}[Z/\bar{E}] = 1$, there is a distribution $Q$ for $(X, Y(0), Y(1), Z, U)$ with the following properties:

(i) The distribution of the observables $(X, Y, Z)$ is the same under $P$ and $Q$.

(ii) $Q$ satisfies Assumption $\Lambda$.

(iii) $\mathbb{E}_Q[Y(1)] = \mathbb{E}_P[YZ/\bar{E}]$.

In short, this result says that $\mathbb{E}[YZ/\bar{E}]$ is a plausible value of $\psi_T$ as long as $\mathbb{E}[Z/\bar{E}] = 1$. It is not hard to show that the converse also holds: if $\psi$ is a plausible value of $\psi_T$, then $\psi = \mathbb{E}[YZ/\bar{E}]$ for some random variable $\bar{E}$ satisfying $\mathbb{E}[Z/\bar{E}] = 1$. As a result, the optimal bounds for $\psi_T$ can be obtained by solving the variational problems in Corollary 4.

Corollary 4. The partially identified set for $\psi_T$ is an interval whose endpoints solve:

$$
\psi_T^- = \min_{\bar{E} \in \mathcal{E}_C(\Lambda)} \mathbb{E}[YZ/\bar{E}] \quad \text{subject to} \quad \mathbb{E}[Z/\bar{E}] = 1
$$

(17)

$$
\psi_T^+ = \max_{\bar{E} \in \mathcal{E}_C(\Lambda)} \mathbb{E}[YZ/\bar{E}] \quad \text{subject to} \quad \mathbb{E}[Z/\bar{E}] = 1.
$$

(18)

Even though the variational problems (17) and (18) can be infinite-dimensional optimization problems with infinitely-many constraints, they have several nice features that enable them to be solved explicitly. Some straightforward algebraic manipulation shows that the problem (18) can be written as

$$
\begin{align*}
&\maximize \mathbb{E}[E[YZ/\bar{E}] \\
&\text{subject to} \quad \mathbb{E}[Z/\bar{E}] = 1
\end{align*}
$$

(19)

and $1 + \frac{1 - e(X)}{e(X)} \Lambda^{-1} \leq 1/\bar{E} \leq 1 + \frac{1 - e(X)}{e(X)} \Lambda$.

Not only is this problem linear in the decision “variable” $1/\bar{E}$, it also separates across levels of $X$. Therefore, it suffices to
separately solve (20) for each \( x \in \mathcal{X} \)

\[
\text{maximize } \mathbb{E}[YZ/\tilde{E}|X = x] \\
\text{subject to } \mathbb{E}[Z/\tilde{E}|X = x] = 1 \quad (20)
\]

and \( 1 + \frac{1-e(x)}{e(x)} \Lambda^{-1} \leq 1/\tilde{E} \leq 1 + \frac{1-e(x)}{e(x)} \Lambda \).

The problem (20) requires us to maximize one expectation subject to an equality constraint on another expectation. This resembles the problem solved by the Neyman–Pearson lemma, and in fact is a special case of the generalization due to Dantzig and Wald (1951). The optimization problems posed in Theorem 1 also fall in this class. It turns out that both of these problems have a common solution, given in Proposition 2.

**Proposition 2 (Formulas for the worst-case propensity scores).** There exist \( \tilde{E}_-, \tilde{E}_+ \in \mathcal{E}_\infty(\Lambda) \) satisfying \( \mathbb{E}[Z/\tilde{E}_-|X] = \mathbb{E}[Z/\tilde{E}_+|X] = 1 \) and also (21) and (22):

\[
\frac{1}{\tilde{E}_-} = \begin{cases} 
1 + \frac{1-e(x)}{e(x)} \Lambda^{-1} & \text{if } Y < Q_{1-\tau}(X, 1) \\
1 + \frac{1-e(x)}{e(x)} \Lambda^{-1} & \text{if } Y > Q_{1-\tau}(X, 1) 
\end{cases} \quad (21)
\]

\[
\frac{1}{\tilde{E}_+} = \begin{cases} 
1 + \frac{1-e(x)}{e(x)} \Lambda^{-1} & \text{if } Y > Q_{\tau}(X, 1) \\
1 + \frac{1-e(x)}{e(x)} \Lambda^{-1} & \text{if } Y < Q_{\tau}(X, 1) 
\end{cases} \quad (22)
\]

Further, \( \tilde{E}_- \) solves both (11) and (17), and \( \tilde{E}_+ \) solves both (12) and (18).

The form of the propensity score \( \tilde{E}_+ \) gives us insight into the confounding structure which maximizes \( \psi_T \): in the worst case, all observations with “high” values of \( Y \) are unlikely to be treated and thus receive large propensity weight, while all observations with “low” values of \( Y \) are likely to be treated and thus receive small propensity weight. The cutoff between high and low is chosen to satisfy the data-compatibility condition \( \mathbb{E}[Z/\tilde{E}_-|X] = 1 \).

This argument presented in this section extends immediately to \( \psi_C \) by swapping treatment and control labels, extends to \( \psi_{ATT} \) by the argument given in Section 3.1, and can extend to other sensitivity models of the form \( e_{\min}(X) \leq e_0(X, U) \leq e_{\max}(X) \) by modifying the constraints of (20).

### 3.3. Data Compatibility for the ATE

To extend the argument from Section 3.2 to the ATE requires additional care. Although \( \psi_{ATE}^+ = \psi_T^+ - \psi_C^- \) is certainly a valid upper bound for the partially identified set for \( \psi_{ATE} \), it is not obviously a sharp one. **Proposition 1** only implies that there exists a distribution \( Q \) matching the observed-data distribution which has \( \mathbb{E}_Q[Y(1)] = \psi_T^+ \) and another distribution \( Q' \) which has \( \mathbb{E}_Q[Y(0)] = \psi_C^- \), but these distributions need not be the same. In other words, the two bounds may not be simultaneously achievable.

**Theorem 2** indicates that the worst-case bounds on the counterfactual means are simultaneously achievable in the marginal sensitivity model. This is a surprising result, given that simultaneous achievability is not expected to hold in the closely-related Rosenbaum sensitivity model. In that model, Yadlowsky et al. (2022) derived sharp bounds on \( \psi_T \) and \( \psi_C \) but required an extra symmetry assumption on the distribution of potential outcomes to establish sharpness of the resulting ATE bounds.

The key to our bounds on \( \psi_{ATE} \) is the following claim, which strengthens **Proposition 1**.

**Proposition 3 (Simultaneous achievability).** For any random variable \( \tilde{E} \in \mathcal{E}_\infty(\Lambda) \) satisfying \( \mathbb{E}[Z/\tilde{E}|X] = \mathbb{E}[(1-Z)/(1-\tilde{E})|X] = 1 \), there is a distribution \( Q \) for the full data \((X, Y(0), Y(1), Z, U)\) with the following properties:

(i) The distribution of the observables \((X, Y, Z)\) is the same under \( P \) and \( Q \).

(ii) \( Q \) satisfies **Assumption A**.

(iii) \( \mathbb{E}_Q[Y(1)] = \mathbb{E}_P[YZ/\tilde{E}] \) and \( \mathbb{E}_Q[Y(0)] = \mathbb{E}_P[Y(1-Z)/(1-\tilde{E})] \).

Unlike **Proposition 1**, this result does not follow from the existing data-compatibility characterizations of Birmingham, Rotnitzky, and Fitzmaurice (2003), Robins, Rotnitzky, and Scharfstein (2000), Tan (2006), and Zhao, Small, and Bhattacharya (2019) and instead requires an original construction. Given this result, one can derive **Theorem 2** as a consequence of **Theorem 1** and **Corollary 2**.

### 4. Sensitivity Analysis

In this section, we give our proposals for translating the population-level partial identification results of Section 3 into practical sensitivity analyses. Our main proposal, which we call the *quantile balancing* method, conducts a sensitivity analysis for IPW estimators by modifying the ZSB proposal to incorporate the sufficient constraints derived in Section 3.1. We also discuss extensions of our sensitivity analysis to the AIPW estimator of Robins, Rotnitzky, and Zhao (1994) which are simpler to implement but only sharp under homoscedasticity.

Throughout this section, we take \( \Lambda \geq 1 \) to be fixed and set \( \tau = \Lambda/(\Lambda + 1) \).

#### 4.1. Sensitivity Analysis via Quantile Balancing

We begin by describing our IPW sensitivity analysis for the average treated potential outcome. **Theorem 1** implies that the largest value of \( \psi_T \) compatible with **Assumption A** solves the optimization problem (23):

\[
\psi_T^+ = \max_{\tilde{E} \in \mathcal{E}_\infty(\Lambda)} \frac{\mathbb{E}[YZ/\tilde{E}]}{\mathbb{E}[Z/\tilde{E}]} \quad \text{subject to} \\
\left( \frac{\mathbb{E}[Q_T(X, 1)Z/\tilde{E}]}{\mathbb{E}[Z/\tilde{E}]} \right) = \left( \frac{\mathbb{E}[Q_T(X, 1)Z/e(X)]}{\mathbb{E}[Z/e(X)]} \right). \quad (23)
\]

In the above display, we have included an additional constraint \( \mathbb{E}[Z/\tilde{E}] = \mathbb{E}[Z/e(X)] \) which motivates our finite-sample procedure without affecting the optimization problem value.

Our proposal is to estimate \( \psi_T^+ \) by replacing all of the unknown quantities in (23) with empirical counterparts. We estimate \( \psi_T^+ \) by following the same principle. To translate these estimates into confidence intervals, we employ the same simple percentile bootstrap scheme as ZSB.

We will be concrete about what optimization problem we are proposing to solve. Let \( Q_T(x, z) \) be an estimate of the conditional
quantile function of \( Y \) obtained by some kind of quantile regression (e.g., Stone 1977; Koenker and Bassett 1978; Meinshausen 2006; Athey, Tibshirani, and Wager 2019). Let \( \hat{c} \) be the data analyst’s estimate of the nominal propensity score \( c \) from their primary analysis. We define \( \hat{\psi}_T^+ \) as the solution to the empirical maximization problem (24):

\[
\hat{\psi}_T^+ = \max_{\hat{c} \in \hat{c}(\Lambda)} \frac{\mathbb{E}_n[Z/e(\hat{c})]}{\mathbb{E}_n[Z/e(\hat{c})]} \quad \text{subject to} \quad \left( \frac{\mathbb{E}_n[\hat{Q}_T(X, 1)/e(X)]}{\mathbb{E}_n[Z/e(\hat{c})]} \right) = \left( \frac{\mathbb{E}_n[\hat{Q}_T(X, 1)/e(\hat{c})]}{\mathbb{E}_n[Z/e(\hat{c})]} \right). 
\]

The lower bound \( \hat{\psi}_T^- \) is defined similarly, but with maximization replaced by minimization and \( \hat{Q}_T(x, z) \) replaced by another quantile estimate \( \hat{Q}_t(x, z) \). We call \( \hat{\psi}_T^+ \) and \( \hat{\psi}_T^- \) the quantile balancing bounds for \( \psi_T \).

Two features of this proposal require some explanation. The first feature to explain is the inclusion of the constraint \( \mathbb{E}_n[Z/e(\hat{c})] = \mathbb{E}_n[\hat{Q}_T(X, 1)/e(\hat{c})] \) in (24). At the population level, Theorem 1 shows that only the constraint \( \mathbb{E}[Q_T(X, 1)/\bar{E}] = \mathbb{E}[Q_T(X, 1)/\bar{e}(X)] \) is relevant. However, in finite samples, this additional constraint improves robustness when \( \hat{Q}_T \) is an inaccurate estimate of \( Q_T \) and also simplifies the associated computation. The second feature to explain is why the right-hand side of the constraints in (24) have an “IPW” form (i.e., \( \mathbb{E}_n[\hat{Q}_T(X, 1)/\bar{e}(X)] \)) rather than a “sample average” form (i.e., \( \mathbb{E}_n[\hat{Q}_T(X, 1)] \)). If \( \mathbb{E}_n[\hat{Q}_T(X, 1)/\bar{e}(X)] \neq \mathbb{E}_n[\hat{Q}_T(X, 1)] \), then a sample average version of (24) may have no feasible propensities. With the IPW form, \( \hat{c}_i = \hat{c}(X_i) \) is always feasible.

Now that we have explained our proposed sensitivity analysis, we will collect several immediate properties of the quantile balancing bounds:

(i) When \( \Lambda = 1 \) (i.e., no confounding is allowed), the quantile balancing bounds collapse to the usual IPW estimate of \( \psi_T \) under unconfoundedness.

(ii) The quantile balancing bounds are sample bounded, that is, \( \min_i Y_i \leq \hat{\psi}_T^- \leq \hat{\psi}_T^+ \leq \max_i Y_i \).

(iii) The quantile balancing bounds are always a subset of the ZSB bounds and, outside of knife-edge cases, are a strict subset.

(iv) The optimization problem (24) is convex and can be solved efficiently. In fact, it reduces to a standard quantile regression problem. See Appendix A, supplementary materials for implementation details.

The Property (i) leads us to call quantile balancing a “sensitivity analysis for IPW.” One can also apply quantile balancing to unstabilized IPW estimators at the cost of Properties (ii) and (iii). See Appendix B, supplementary materials for computational details, including for Augmented IPW estimators.

The quantile balancing idea extends to other causal estimands. To compute bounds for \( \psi_C \), one only needs to exchange the definitions of “treated” and “control” and solve the same optimization problem. Subtracting the bounds for \( \psi_T \) and \( \psi_C \) gives bounds for \( \psi_{ATE} \), and bounds for \( \psi_{ATT} \) follow from a similar principle (see Appendix A, supplementary materials for the exact formula).

To form confidence intervals based on quantile balancing, we follow Zhao, Small, and Bhattacharya (2019) and propose using the percentile bootstrap. If \( \{\hat{\psi}_T^+, \hat{\psi}_T^-\} \) are quantile balancing bounds estimated in the \( b \)th of \( B \) bootstrap samples, we report the quantile balancing \( 1 - \alpha \) confidence interval as

\[
\text{CI}(\alpha) = \bigl[\hat{Q}_{\alpha/2}(\{\hat{\psi}_T^-\})_{b \in [B]}, \hat{Q}_{1-\alpha/2}(\hat{\psi}_T^+)_{b \in [B]}\bigr].
\]

As is standard for bootstrap-based IPW inference, we require re-estimating the nominal propensity score separately in each bootstrap replication. That requirement does not extend to the conditional quantiles. While the conditional quantiles can be re-estimated within bootstraps, our inference results will also apply if they are taken from the main dataset. This helps keep inference computationally tractable.

### 4.2. Implications for AIPW Sensitivity Analysis

The quantile balancing sensitivity analysis described above requires the data analyst to perform several quantile regressions. Our partial identification results imply that, in certain “additive-noise” data generating processes, a data analyst whose primary analysis was conducted using the AIPW estimator can perform sharp sensitivity analysis without performing any quantile regressions.

To explain how, we begin by describing the modeling assumption. Suppose the observed outcome \( Y \) has the following signal-plus-noise representation:

\[
Y = \mu(X, Z) + \epsilon \quad \text{with} \quad \mathbb{E} [\epsilon] = 0, \epsilon \independent (X, Z).
\]

Such models frequently arise in the regression applications (see, e.g., Hastie, Tibshirani, and Friedman 2001, chap. 3) and fit quite well in the real-data example we present in Section 5.2.

The additive-noise assumption (26) implies that the conditional quantiles of the residuals \( \epsilon \) are constant. In particular, the assumption implies \( Q_t(x, z) = \mu(x, z) + Q_t(\epsilon) \), where \( Q_t(\epsilon) \) is the \( r \)th quantile of the noise. Therefore, Theorem 1 and some algebra imply that the sharp upper bound for \( \psi_T \) has the following formula:

\[
\psi_T^+ = \max_{\hat{c} \in \hat{c}(\Lambda)} \left\{ \mathbb{E}[\mu(X, 1) + \mathbb{E}[Y - \mu(X, 1)/\bar{E}]/\mathbb{E}[Z/\bar{E}]] \right\}
\]

subject to \( \mathbb{E}[Z/\bar{E}] = \mathbb{E}[Z/e(X)] \).

Similar formulas can be derived for \( \psi_T^-, \psi_C^+, \psi_C^- \). This formula is convenient after an AIPW primary analysis, which requires estimates of all the nuisance parameters in this equation.

A natural estimate of \( \psi_T^+ \) is the finite-sample analogue of (27):

\[
\hat{\psi}_{T\text{AIPW}}^+ = \max_{\hat{c} \in \hat{c}(\Lambda)} \left\{ \mathbb{E}_n[\hat{\mu}(X, 1) + \mathbb{E}_n[(Y - \hat{\mu}(X, 1)/\bar{E})]/\mathbb{E}_n[Z/\bar{E}]] \right\}
\]

subject to \( \mathbb{E}_n[Z/\bar{E}] = \mathbb{E}_n[Z/e(X)] \).

The estimated bound \( \hat{\psi}_{T\text{AIPW}}^+ \) grows with \( \Lambda \) and recovers the original (stabilized) AIPW estimator when \( \Lambda = 1 \). One can also not divide by \( \mathbb{E}_n[Z/\bar{E}] \) in (28) to recover the unstabilized AIPW estimator at \( \Lambda = 1 \).

The estimator (28) slightly modifies the proposal in Section 6.2 of Zhao, Small, and Bhattacharya (2019) to include the balancing constraint \( \mathbb{E}_n[Z/\bar{E}] = \mathbb{E}_n[Z/e(X)] \). In theory, this constraint is necessary to achieve sharpness in the additive-noise...
model (26). However, the simulations presented in Section 5 find that when the additive-noise model holds, this constraint scarcely refines the stabilized point estimates while somewhat degrading the coverage of bootstrap confidence intervals.

4.3. Theoretical Properties

We now state some theoretical properties of the quantile balancing bounds [\(\hat{\psi}^-, \hat{\psi}^+\)] which apply when the outcome \(Y\) has a continuous distribution. In short, the bounds are sharp when quantiles are estimated consistently and are valid even when quantiles are estimated inconsistently. Moreover, the percentile bootstrap yields valid confidence intervals if standard IPW inference conditions are satisfied and quantiles are estimated parametrically.

To obtain these results, we need a few conditions. The first condition collects some standard IPW consistency requirements which we expect the data analyst to have already assumed in their primary analysis.

**Condition 1 (IPW assumptions).** The nominal propensity score \(e\) satisfies \(e \leq e(X) \leq 1 - e\) with probability one for some \(e > 0\). The estimated nominal propensity score \(\hat{e}(\cdot) \equiv \hat{e}(\cdot, [X, Z]_{1 \leq n})\) is uniformly consistent, and the variance of \(Y\) is finite.

The second condition requires that the outcome \(Y\) has a bounded conditional density which is positive near the relevant conditional quantiles. This is a common identification condition for quantile regression (Athey, Tibshirani, and Wager 2019; Belloni et al. 2019). However, it means our theoretical guarantees do not apply when \(Y\) is discrete.

**Condition 2 (Density).** The conditional distribution of \(Y \mid X, Z\) has a uniformly bounded density \(f(y|x, z)\). For each \((x, z) \in \mathcal{X} \times [0, 1]\), the map \(y \mapsto f(y|x, z)\) is continuous and positive near \(Q_{1-\tau}(x, z)\) and \(Q_{\tau}(x, z)\).

Finally, we make some assumptions about how the quantiles are estimated. For the standard linear quantile regression method of Koenker and Bassett (1978), one only needs to check that the regressors in the quantile regression have finite variance. We cover generic (possibly nonlinear) methods by requiring sample splitting to avoid overfitting. The specific form of sample splitting analyzed in our proofs is “cross-fitting” (Schick 1986; Newey and Robins 2017; Chernozhukov et al. 2018), but leave-one-out or out-of-bag quantile estimates perform similarly in simulations. Based on our practical experience, we recommend using some kind of sample splitting even when the quantile model is linear.

**Condition 3 (Quantile estimates).** For each \(t \in [1 - \tau, \tau]\), one of the following holds for the estimated quantile function \(\hat{Q}_t\):

(i) \(\hat{Q}_t(x, z) = \hat{\beta}_t(z)\top h(x)\) for some fixed “features” \(h_j(x)\) with finite variance.

(ii) \(\hat{Q}_t(x, z)\) is estimated using cross-fitting and satisfies Condition N in the supplementary materials.

**Condition 3 is essentially “algorithmic,” and neither (i) nor (ii) impose any accuracy requirements on the estimated conditional quantiles. The restrictions in (ii) are technical to state but very mild. Under Conditions 1 and 2, they are satisfied by quantile estimates based on nearest-neighbors (Stone 1977), kernels (Bhattacharya and Gangopadhyay 1990), and random forests (Meinshausen 2006; Athey, Tibshirani, and Wager 2019).

Under these conditions, we have the following result on the asymptotic sharpness of the quantile balancing bounds.

**Theorem 3 (Sharpness and robustness).** For any \(\psi_0 \in [\psi, \psi_C, \psi_{ATT}, \psi_{ATE}]\), let \([\hat{\psi}^-, \hat{\psi}^+\)\] be its partially identified interval under Assumption A and let \([\hat{\psi}^-, \hat{\psi}^+]\) be the corresponding quantile balancing interval. Assume Conditions 1, 2, and 3.

(i) If the quantile regression estimates are consistent, then \(\hat{\psi}^- \overset{p}{\rightarrow} \psi^-\) and \(\hat{\psi}^+ \overset{p}{\rightarrow} \psi^+\).

(ii) Even if the quantile models are misspecified, we still have \(\hat{\psi}^- \leq \psi^- + a_n\) and \(\psi^- + b_n \leq \hat{\psi}^+\), where \(a_n = o_P(1)\) and \(b_n = o_P(1)\).

The same conclusions hold for the AIPW-based bounds introduced in Section 4.2 when the outcome regression is estimated by linear regression, that is, sharpness under an additive-noise model and validity in general. However, while AIPW is doubly-robust under unconfoundedness, the validity of the corresponding AIPW quantile balancing bounds relies on correct specification of the nominal propensity score.

The result (ii) shows that even when quantiles are not estimated consistently, the quantile balancing bounds are still valid; we will offer some intuition on why this novel robustness property holds. At the population level, the worst-case propensity score \(\hat{E}_{\psi}^-\) defined in Proposition 2 “balances” all integrable functions of \(X\), so intuitively, we should expect that it “nearly” balances the estimated quantile function \(\hat{Q}_t(\cdot, 1)\) in finite samples even if \(\hat{Q}_t(\cdot, 1)\) is not particularly close to \(Q_t(\cdot, 1)\). That suggests a vector of propensities very close to the true worst-case propensity vector will belong to the feasible set \(E_{\psi}(\Lambda)\). Since the quantile balancing upper bound \(\hat{\psi}_T^+\) is defined as a maximum over the feasible set, it will be at least as large as a quantity close to \(\psi_T^+\). This roughly explains why validity holds even under misspecification.

The validity of the confidence interval (25) follows under stronger parametric assumptions. We prove an inference result assuming the nominal propensity score is estimated by a correctly-specified parametric model and the conditional quantiles are estimated by a (potentially misspecified) parametric model.

**Theorem 4 (Inference).** Let \([\hat{\psi}^-, \hat{\psi}^+]\) be as in Theorem 3, and let CI(\(\alpha\)) as be as in (25). Suppose Conditions 1, 2, and 3(i) are satisfied, and also that the nominal propensity score is estimated by a regular parametric model (e.g., logistic regression). Then we have

\[
\liminf_{n \to \infty} \Pr([\hat{\psi}^-, \hat{\psi}^+] \subseteq \text{CI}(\alpha)) \geq 1 - \alpha
\]  

(29)

for any \(\alpha \in (0, 1)\).
are correctly specified. In our simulations, the use of cross-fit conditional quantile estimates largely resolves the issue with minimal effect on point estimates, so we advocate for the use of such estimators in practice. Although we do not have theoretical support for the confidence interval CI(α) when quantiles are estimated by a nonlinear model, we find that approach performs reasonably well in the simulations of Section 5 as long as cross-fit quantiles are used.

5. Numerical Examples

In this section, we illustrate the finite-sample performance of our proposed sensitivity analyses on several simulated datasets and one real-data example.

5.1. Simulated Data

We consider two data-generating processes (DGPs) in our simulated examples. The two DGPs differ in the conditional distribution of $Y$ given $(X, Z)$, but otherwise can be described as follows:

$$X \sim \text{Uniform}([-1, 1]^5)$$

$$Z | X \sim \text{Bernoulli} \left( \frac{1}{1 + \exp(-\sum_{j=1}^{5} X_j / \sqrt{5})} \right)$$

$$Y | X, Z \sim N(\mu(X), \sigma^2(X))$$

(30)

In the first DGP, we use $\mu(x) = x_1 + \cdots + x_5$ and $\sigma(x) = 1$. In the second DGP, we use $\mu(x) = \frac{3}{2} \text{sign}(x_1) + \text{sign}(x_2)$ and $\sigma(x) = 2 + \text{sign}(x_3) + \text{sign}(x_4)$. The estimand of interest is the ATE and we fix $\Lambda = 2$, that is, unobserved confounders can double or halve the odds of treatment.

We compare five methods for obtaining bounds on the partially identified set:

1. **QB-Linear** applies the quantile balancing method of Section 4 with quantiles estimated using linear quantile regression on $X_1, \ldots, X_5$.
2. **QB-Forest** applies quantile balancing with quantiles estimated using the random forest method from Athey, Tibshirani, and Wager (2019).
3. **ZSB** applies the main IPW method from Zhao, Small, and Bhattacharya (2019), described in Section 2.1.
4. **ZSB-AIPW** applies the AIPW-based method from Section 6.2 of Zhao, Small, and Bhattacharya (2019), described in Section 4.2. This requires an estimate of the outcome model $\mu(X, Z) = \mathbb{E}[Y|X, Z]$. We use a situationally-appropriate outcome model, linear regression in DGP1 and random forest regression in DGP2.
5. **AIPW+1** applies the AIPW-based method introduced in Section 4.2. We call this AIPW+1 because it refines ZSB-AIPW to incorporate an additional “one-balancing” constraint $\mathbb{E}_n[Z/\bar{e}] = \mathbb{E}_n[Z/\hat{e}(X)]$.

All methods estimate the nominal propensity score by logistic regression. We use 5-fold cross-fitting in all of our quantile regressions. We do not re-estimate quantiles or random forest models within bootstraps.

Figure 1 shows the distribution of upper and lower bound point estimates from each of these five methods, estimated using 2000 simulations with $n = 1000$ observations each. Simulations at other sample sizes are presented in Appendix B, supplementary materials. Dashed lines indicate the true partially identified region. The results conform to the asymptotic predictions of

![Figure 1. Boxplots of the ATE upper and lower bound point estimates for both DGPs and all considered methods. The dashed line indicates the boundary of the true partially identified set. In DGP1, all methods but ZSB are correctly specified and give reasonably accurate bounds. In DGP2, the Forest method is well-suited to the piecewise-constant conditional quantiles and gives the most accurate bounds.](image-url)
Section 4: (i) when the quantile models are “correctly specified,” the quantile balancing point estimates are nearly unbiased; (ii) under misspecification, the range of QB point estimates is too wide rather than too narrow; (iii) the ZSB range of point estimates is too wide in both cases; and (iv) AIPW-based methods give nearly-sharp bounds in the additive-noise DGP1 but conservative bounds in the heteroscedastic DGP2. We also find that the $+1$ constraint in AIPW+1, which is necessary for sharpness in theory, has minimal practical impact in either DGP.

Figure 2 shows the coverage for 95% bootstrap confidence intervals based on each of the five methods. In DGP1, both quantile balancing methods have nearly nominal coverage, but AIPW-based methods undercover and the $+1$ constraint exacerbates the undercoverage. In DGP2 the QB-Forest method achieves nearly nominal coverage, while all other methods overcover. The ZSB method overcovers for both DGPs.

5.2. Real Data

In this section, we apply our proposed sensitivity analysis to a subsample of data from the 1966–1981 National Longitudinal Survey (NLS) of Older and Young Men. We wish to estimate the impact of union membership on wages. Specifically, we consider the ATE of union membership on log wages. For illustrative reasons, we focus on the 1978 cross-section of Young Men and restrict our attention to craftsmen and laborers not enrolled in school. Our estimates are thus based on a sample of 668 respondents with measurements of wages, union membership, and eight covariates.

For our primary analysis, we use IPW to adjust for baseline imbalances in covariates between union and nonunion samples. Table 1 reports the covariate balance between union and nonunion samples before and after weighting by the (estimated) inverse propensity score. On several important characteristics, inverse propensity weighting dramatically improves balance across the two samples.

The IPW point estimate of the ATE is 0.23 with an associated 90% confidence interval of [0.18, 0.27]. Thus, our primary analysis concludes that union membership has a positive effect on wages, at least on average among craftsmen and laborers. Both the point estimate and the confidence interval are in agreement with prior literature studying the same problem using cross-sectional data. See Jakubson (1991) and Johnson (1975) for overviews. An AIPW-based primary analysis gives the same point estimate and confidence interval, up to rounding. Freeman (1984), Mellow (1981), and many other economists have argued that cross-sectional estimates of the union premium overestimate the true causal effect because higher-skill workers are simultaneously more likely to be selected for union jobs and earn higher wages. Here, “skill” refers to an unobserved confounder which is only partially captured by the measured covariates. Is it plausible that the positive effect we find in the IPW analysis could be entirely due to selection on skill? A sensitivity analysis may help address this question.

Figure 3 reports point estimate ranges and 90% bootstrap confidence intervals from quantile balancing, the ZSB-IPW method, and the ZSB-AIPW method for several values of the sensitivity parameter $\Lambda$. For quantile balancing, we estimate conditional quantiles using linear quantile regression with 5-fold cross fitting. For AIPW, we use linear regression for the outcome model.

All three sensitivity analyses show that the positive effect found in the primary analysis is fairly robust to unobserved confounding, but quantile balancing and ZSB-AIPW refine the
baseline ZSB-IPW interval. Even if the odds of union membership for "skilled" workers were nearly double ($\Lambda = 1.9$) the odds for "typical" workers with the same observed covariates, the quantile balancing and AIPW sensitivity analyses analysis would still find a statistically significant positive treatment effect. Meanwhile, when $\Lambda = 1.8$, the ZSB confidence intervals already include the null. In this application, quantile balancing only slightly refines the ZSB range. Moreover, quantile balancing and ZSB-AIPW yield very similar ranges and confidence intervals. This is to be expected from the discussion in Section 4.2, as an "additive noise" model appears quite plausible in this application.

To put these sensitivities in context, we follow Kallus and Zhou (2020b) and compute the degree to which the (estimated) odds of union membership could change if measured confounders were omitted from the dataset. Caveats to this approach and more sophisticated empirical calibration strategies are discussed in Hsu and Small (2013), Zhang and Small (2020), and Cinelli and Hazlett (2020). No measured confounders except Laborer and South were able to nearly double or halve the odds of union membership for any respondent. We interpret these results as showing that the qualitative conclusions of the primary analysis are fairly robust to unobserved confounding by skill.

Incidentally, longitudinal estimates of union wage effects—which control for individual-specific effects like "skill"—come to similar conclusions as the one suggested by our sensitivity analysis. Although treatment effect estimates from longitudinal studies are generally smaller than those from cross-sectional studies, they still find evidence in favor of the "union premium" (Chamberlain 1982; Freeman 1984; Jakubson 1991).

6. Conclusion

We have shown that quantile balancing—a simple modification of the popular ZSB sensitivity analysis— is feasible, robust, and sharp. This new sensitivity analysis for IPW is based on novel partial identification results for Tan (2006)'s marginal sensitivity model.

We will point to several interesting directions for future work. While our partial identification results focus on counterfactual means and a few treatment effects, it should be possible to extend our partial identification results to more complex estimands of the type considered in Kallus, Mao, and Zhou (2019), Kallus and Zhou (2020a), Kallus and Zhou (2018), Kallus and Zhou (2020b), and Lee, Bargagli-Stoffi, and Dominici (2020). Perhaps a similarly compact sensitivity analysis could even apply to dynamic treatment regimes. Future work could also investigate data-compatibility in the finite population model. In addition, while our IPW identification arguments generalize to any sensitivity assumption that only restricts the propensity score in a pointwise fashion (i.e., $e_{\min}(x) \leq e(x, u) \leq e_{\max}(x)$), the practicality of our sensitivity analysis and its theoretical properties rely on the marginal sensitivity model quite heavily. It would be interesting to see if a practical and sharp sensitivity analysis could be developed for other sensitivity assumptions in this class.

Supplementary Materials

Online supplement: implementation, simulations, and proofs Details on implementation, additional simulation results, and proofs of the propositions, corollaries, and theorems in this manuscript. (pdf file)

Acknowledgments

We are grateful for comments from Guillaume Basse, Bo Honoré, Nathan Kallus, Michal Kolesár, David Lee, Ulrich Müller, Karl Schulze, Dylan Small, Angela Zhou, Qingyuan Zhao, Mikkel Plagborg-Møller, and seminar participants at Berkeley, Cambridge, and Princeton.

Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2039656. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
References

Aaronov, P. M., and Lee, D. D. K. K. (2013), “Interval Estimation of Population Means Under Unknown but Bounded Probabilities of Sample Selection,” *Biometrics*, 100, 235–240. [2647]

Athey, S., Tibshirani, J., and Wager, S. (2019), “Generalized Random Forests,” *Annals of Statistics*, 47, 1148–1178. [2651,2652,2653]

Athey, S., and Wager, S. (2021), “Policy Learning with Observational Data,” *Econometrica*, 89, 133–161. [2646]

Austin, P. C., and Stuart, E. A. (2015), “Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies,” *Statistics in Medicine*, 34, 3661–3679. [2645,2646]

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernández-Val, I. (2019), “Conditional Quantile Processes Based on Series or Many Regressors,” *Journal of Econometrics*, 213, 4–29. Annals: In Honor of Roger Koenker. [2652]

Beresteanu, A., Molchanov, I., and Molinari, F. (2011), “Sharp Identification Regions in Models with Convex Moment Predictions,” *Econometrica*, 79, 1785–1821. [2648]

Bhattacharya, P. K., and Gangopadhyay, A. K. (1990), “Kernel and Nearest-Neighbor Estimation of a Conditional Quantile,” *Annals of Statistics*, 18, 1400–1415. [2652]

Birmingham, J., Rotnitzky, A., and Fitzmaurice, G. M. (2003), “Pattern-Mixture and Selection Models for Analysing Longitudinal Data with Monotone Missing Patterns,” *Journal of the Royal Statistical Society, Series B*, 65, 275–297. [2649,2650]

Chamberlain, G. (1982), “Multivariate Regression Models for Panel Data,” *Journal of Econometrics*, 18, 5–46. [2655]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), “Double/Debiased Machine Learning for Treatment and Structural Parameters,” *Econometric Journal*, 21, C1–C68. [2652]

Cinelli, C., and Hazlett, C. (2020), “Making Sense of Sensitivity: Extending Omitted Variables Bias,” *Journal of the Royal Statistical Society, Series B*, 82, 39–67. [2649,2655]

Dantzig, G. B., and Wald, A. (1951), “On the Fundamental Lemma of Neyman and Pearson,” *Annals of Mathematical Statistics*, 22, 87–93. [2650]

Davies, L., and D’Haultfoeuille, X. (2016), “A New Characterization of Identified Sets in Partially Identified Models.” Working paper. [2648]

Franks, A. M., Amour, A. D., and Feller, A. (2020), “Flexible Sensitivity Analysis for Observational Studies Without Observable Implications,” *Journal of the American Statistical Association*, 115, 1730–1746. [2649]

Freeman, R. B. (1982), “Longitudinal Analyses of the Effects of Trade Dantzig, G. B., and Wald, A. (1951), “On the Fundamental Lemma of Neyman and Pearson,” *Annals of Mathematical Statistics*, 22, 87–93. [2650]

Hristache, M., and Patilea, V. (2017), “Conditional Moment Models with Convex Moment Predictions,” *Econometrica*, 84, 33–50. [2651,2652,2655]

Kallus, N., and Zhou, A. (2018), “Confounding-Robust Policy Improvement” in *Advances in Neural Information Processing Systems* 31, eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. pp. 9269–9279. Curran Associates, Inc. [2645,2646,2647,2655]

Lee, D. (2009), “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” *Review of Economic Studies*, 76, 1071–1102. [2648]

Lee, K., Bargagli-Stoffi, F. J., and Dominici, F. (2020), “Causal Rule Ensemble: Interpretable Inference of Heterogeneous Treatment Effects.” [2645,2647,2655]

Masten, A. S., and Poirier, A. (2018), “Identification of Treatment Effects Under Conditional Partial Independence,” *Econometrica*, 86, 317–351. [2648]

Masten, A. S., Poirier, A., and Zhang, L. (2020), “Assessing Sensitivity to Unconfoundedness: Estimation and Inference.” [2648]

Meinshausen, N. (2006), “Quantile Regression Forests,” *Journal of Machine Learning Research*, 7, 983–999. [2651,2652]

Mellow, W. (1981), “Unionism and Wages: A Longitudinal Analysis,” *Review of Economics and Statistics*, 63, 43–52. [2654]

Murray, L. W., Wager, S., and Zubiagaureta, J. R. (2018), “Shape-Constrained Partial Identification of a Population Mean Under Unknown Probabilities of Sample Selection,” *Biometrika*, 105, 103–114. [2647]

Newey, W. K., and Robins J. M. (2017), “Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation,” CeMMAP working papers CWP41/17, Centre for Microdata Methods and Practice, Institute for Fiscal Studies. [2652]

Neyman, J. (1923), “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” *Statistical Science*, 5, 465–472. [2646]

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000), “Sensitivity Analysis for Selection Bias Unmeasured Confounding in Missing Data and Causal Inference Models,” in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, eds. M. E. Halloran and D. Berry, pp. 1–94. New York: Springer. [2649,2650]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), “Estimation of Regression-Coefficients When Some Regressors are not Always Observed,” *Journal of the American Statistical Association*, 89, 846–866. [2650]

Rosenbaum, P. R. (1987), “Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies,” *Biometrika*, 74, 13–26. [2645]

——— (2002), “Covariance Adjustment in Randomized Experiments and Observational Studies,” *Statistical Science*, 17, 286–327. [2645]

——— (2005), “Heterogeneity and Causality,” *The American Statistician*, 59, 147–152. [2649]

——— (2010), *Design of Observational Studies*, New York: Springer. [2645]

Rosenbaum, E., Basse, G., Owen, A., and Balocchi, M. (2020), “Combining Observational and Experimental Datasets Using Shrinkage Estimators.” [2645]

Rosenbaum, E. T. R., and Owen, A. B. (2021), “Designing Experiments Informed by Observational Studies,” *Journal of Causal Inference*, 9, 147–171. [2645]

Rubin, D. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701. [2646]

Schick, A. (1986), “On Asymptotically Efficient Estimation in Semiparametric Models,” *Annals of Statistics*, 14, 1139–1151. [2652]
Semenova, V. (2020), “Better Lee Bounds.” [2648]
Soriano, D., Ben-Michael, E., Bickel, P. J., Feller, A., and Pimentel, S. D. (2021), “Interpretable Sensitivity Analysis for Balancing Weights. [2645]
Stone, C. J. (1977), “Consistent Nonparametric Regression,” Annals of Statistics, 5, 595–620. [2651,2652]
Tan, Z. (2006), “A Distributional Approach for Causal Inference Using Propensity Scores,” Journal of the American Statistical Association, 101, 1619–1637. [2645,2646,2647,2648,2649,2650,2655]
Tudball, M., Zhao, Q., Hughes, R., Tilling, K., and Bowden, J. (2019), “An Interval Estimation Approach to Sample Selection Bias.” [2647,2648]
Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J. C., and Tian, L. (2022), “Bounds on the Conditional and Average Treatment Effect with Unobserved Confounding Factors,” The Annals of Statistics [2650]
Zhang, B., and Small, D. S. (2020), “A Calibrated Sensitivity Analysis for Matched Observational Studies with Application to the Effect of Second-Hand Smoke Exposure on Blood Lead Levels in Children,” Journal of the Royal Statistical Society, Series C, 69, 1285–1305. [2655]
Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019), “Sensitivity Analysis for Inverse Probability Weighting Estimators via the Percentile Bootstrap.” Journal of the Royal Statistical Society, Series B, 81, 735–761. [2645,2646,2647,2649,2650,2651,2653]