SENSITIVITY ANALYSIS FOR INVERSE PROBABILITY WEIGHTING ESTIMATORS VIA THE PERCENTILE BOOTSTRAP

QINGYUAN ZHAO, DYLAN S. SMALL AND BHASWAR B. BHATTACHARYA

Department of Statistics, The Wharton School, University of Pennsylvania

Abstract. To identify the estimand in missing data problems and observational studies, it is common to base the statistical estimation on the "missing at random" and "no unmeasured confounder" assumptions. However, these assumptions are unverifiable using empirical data and pose serious threats to the validity of the qualitative conclusions of the statistical inference. A sensitivity analysis asks how the conclusions may change if the unverifiable assumptions are violated to a certain degree. In this paper we consider a marginal sensitivity model which is a natural extension of Rosenbaum's sensitivity model for matched observational studies. We aim to construct confidence intervals based on inverse probability weighting estimators, such that the intervals have asymptotically nominal coverage of the estimand whenever the data generating distribution is in the collection of marginal sensitivity models. We use a percentile bootstrap and a generalized minimax/maximin inequality to transform this intractable problem to a linear fractional programming problem, which can be solved very efficiently. We illustrate our method using a real dataset to estimate the causal effect of fish consumption on blood mercury level.

1. Introduction

A common task in statistics is to estimate the average treatment effect in observational studies. In such problems, the estimand is not identifiable from the observed data without further assumptions. The most common identification assumption, namely the "no unmeasured confounder" (NUC) assumption, asserts that the confounding mechanism is completely determined by some observed covariates. Based on this assumption, many statistical procedures have been proposed and thoroughly studied in the past decades, including propensity score matching (Rosenbaum and Rubin, 1983), inverse probability weighting (Horvitz and Thompson, 1952), and doubly robust and machine learning estimators (Robins et al., 1994, Van Der Laan and Rubin, 2006, Chernozhukov et al., 2017).

However, the underlying NUC assumption is not verifiable using empirical data, posing serious threats to the usefulness of the subsequent statistical inferences that crucially rely on this assumption. A prominent example is the antioxidant vitamin beta carotene. Willett (1990), after reviewing observational epidemiological data, concluded that "Available data thus strongly support the hypothesis that dietary carotenoids reduce the risk of lung cancer". Quite unexpectedly, four years later a large-scale randomized controlled trial (The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group, 1994) reached the opposite conclusion and found "a higher incidence of lung cancer among the men who received

E-mail address: {qyzhao,dsmall,bhaswar}@wharton.upenn.edu.

Date: November 30, 2017.

beta carotene than among those who did not (change in incidence, 18 percent; 95 percent confidence interval, 3 to 36 percent)". The most probable reason for the disagreement between the observational studies and the randomized trial is insufficient control of confounding.

Criticism of confounding bias in observational studies dates at least back to Fisher (1958) who suggested that the association between smoking and lung cancer may be due to genetic predisposition. In response to this, Cornfield et al. (1959) conducted the first formal sensitivity analysis in an observational study. They concluded that, in order to explain the apparent odds ratio of getting lung cancer when there is no real causal effect, those with the genetic predisposition (or any hypothetical unmeasured confounder) must be 9 times more prevalent in smokers than in non-smokers. Because a genetic predisposition was seen as unlikely to have such a strong effect, this strengthened the evidence that smoking had a causal harmful effect.

Cornfield et al. (1959)'s initial sensitivity analysis only applies to binary outcomes and ignores sampling variability. These limitations were later removed in a series of pioneering work by Rosenbaum and his coauthors (Rosenbaum, 1987, Gastwirth et al., 1998, Rosenbaum, 2002b,c). Rosenbaum's sensitivity model considers all possible violations of the NUC assumption as long as the violation is less than some degree. Using a matched cohort design, Rosenbaum attempts to quantify the strength of the unmeasured confounders needed to not reject the sharp null hypothesis of no treatment effect. However, Rosenbaum's framework is only limited to matched observational studies and usually requires effect homogeneity to construct confidence intervals for the average treatment effect.

Besides matching, another widely used method in causal inference is inverse probability weighting (IPW), where the observations are weighted by the inverse of the probability of being treated (Horvitz and Thompson, 1952). IPW estimators have good efficiency properties (Hirano et al., 2003) and can be augmented with outcome regression to become "doubly robust" (Robins et al., 1994). There is much recent work aiming to improve the efficiency of IPW-type estimators by using tools developed in machine learning and high dimensional statistics (Van der Laan and Rose, 2011, Belloni et al., 2014, Athey et al., 2016, Chernozhukov et al., 2017). However, they all heavily rely on the NUC assumption. Robustness of the IPW-type estimators to unmeasured confounding bias are usually studied using pattern-mixture models (Birmingham et al., 2003) or selection models (Scharfstein et al., 1999), in which the unmeasured confounding is usually modeled parametrically. See Richardson et al. (2014) for a recent overview and Section 7.1 for more references and discussion.

In this paper we propose a new framework for sensitivity analysis of missing data and observational studies problems, which can be applied to "smooth" estimators such as the inverse probability weighting (IPW) estimator and the "doubly robust" augmented IPW estimators. We consider a marginal sensitivity model introduced in Tan (2006, Section 4.2), which is a natural modification of Rosenbaum's model. Compared to existing sensitivity analyses of IPW estimators, our sensitivity model is nonparametric in the sense that we do not require the unmeasured confounder to affect the treatment and the outcome in a parametric way. This is appealing because we can never observe the unmeasured confounder and thus cannot test any parametric assumption.

Our marginal sensitivity model measures the degree of violation of the NUC assumption by the odds ratio between the conditional probability of being treated given the measured confounders and conditional probability of being treated given the measured confounders and the outcome/potential outcome variable (see Section 3 for the precise definition). Given a user-specified magnitude for this odds-ratio, the goal is to obtain a confidence interval of

the estimand (the mean response vector in missing data problems and the average treatment effect in observational studies), with asymptotically $1-\alpha$ coverage probability, for all data generating distributions which violate the MAR assumption within this threshold (see Definition 3).

The obvious approach to compute such a confidence interval involves using the asymptotic normal distribution of the IPW estimators. However, the asymptotic sandwich variance estimators of the IPW estimates are quite complicated. Finding extrema of the point estimates and variance estimates over a collection of sensitivity models is generally computationally intractable. We circumvent this problem by using the percentile bootstrap, reducing the problem to solving many linear fractional programs. The highlights of our strategy and the summary of the results are given below:

(1) To obtain a confidence interval under the marginal sensitivity model, we simply maximize/minimize the IPW point estimates over the marginal sensitivity model over B bootstrap resamples of the data. Then we report the $\frac{\alpha}{2}$-sample quantile of the B minima and $(1-\frac{\alpha}{2})$-sample quantile of the B maxima, as the lower and upper end-points of the confidence interval, respectively (Section 4.3).
(2) The asymptotic $(1-\alpha)$ coverage of this interval follows from the limiting normal distribution of the IPW estimators and a generalized minimax/maximin inequality, which justifies the interchange of quantile and infimum/supremum (Theorem 4 and Corollary 6). In fact, our interval covers the entire partially identified region of the estimand with probability tending to $(1-\alpha)$.
(3) Maximizing/minimizing the IPW point estimate over the marginal sensitivity model is very efficient computationally. This is a linear fractional programming problem (ratio of two linear functions) which can be reduced to solving a linear program using the well-known Charnes-Cooper transformation (Section 4.4). In fact, by local perturbations, it can be shown that the solution of the resulting linear program has the same/opposite order as the outcome vectors, using which the confidence interval can be computed in time linear in sample size, for every bootstrap resample (Proposition 5).
(4) The percentile bootstrap approach and the reduction to linear programming are very general and can be extended to sensitivity analysis of other smooth estimators, such as IPW estimates of the mean of the non-respondents and the average treatment effect on the treated (Section 6.1), the augmented inverse probability weighting estimator (Section 6.2), and inference for partially identified parameters (Section 7.3).

The rest of this paper is organized as follows. We introduce the necessary notations in Section 2 and the marginal sensitivity model in Section 3. Our method for constructing confidence under the marginal sensitivity model, for the IPW estimator for the mean response with missing data, is described in Section 4. The extension to estimating the average treatment effect in observational studies is discussed in Section 5. Other extensions, including the use of augmented IPW estimators, are considered in Section 6. We review some related sensitivity analysis methods and compare our framework with Rosenbaum's sensitivity analysis in Section 7. In Section 8 we illustrate our method using a real data example and compare it with Rosenbaum's sensitivity analysis.

2. Background

In this paper, we consider sensitivity analysis in two closely related problems: estimation of the mean response with missing data and estimation of the average treatment effect in observational studies:

(1) In the missing data problem, we assume $(A_1, \boldsymbol{X}_1, Y_1)$, $(A_2, \boldsymbol{X}_2, Y_2), \ldots$ $(A_n, \boldsymbol{X}_n, Y_n)$ are i.i.d. from a joint distribution F_0, where for each subject $i \in [n] := \{1, 2, \ldots, n\}$, A_i is the indicator of non-missing response, $\boldsymbol{X}_i \in \mathscr{X} \subset \mathbb{R}^d$ is a vector of covariates, and $Y_i \in \mathbb{R}$ is the response (observed only if $A_i = 1$). In other words, we only observe $(A_i, \boldsymbol{X}_i, A_i Y_i)$ for $i \in [n]$. Based on the observed data, our goal is to estimate the *mean response* $\mu := \mathbb{E}_0[Y]$ in the complete data, where $(A, \boldsymbol{X}, Y) \sim F_0$ and $\mathbb{E}_0$ indicates that the expectation is taken over the true data generating distribution F_0.

(2) In observational studies, we observe i.i.d. $(A_1, \boldsymbol{X}_1, Y_1), (A_2, \boldsymbol{X}_2, Y_2), \ldots, (A_n, \boldsymbol{X}_n, Y_n)$, where for each subject $i \in [n]$, A_i is a binary treatment indicator (which is 1 if treated and 0 in control), $\boldsymbol{X}_i \in \mathscr{X} \subset \mathbb{R}^d$ is a vector of measured confounders, and

$$Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

is the outcome. Furthermore we assume $(A_i, \boldsymbol{X}_i, Y_i(0), Y_i(1))$ are i.i.d. from a joint distribution F_0. Here, we are using Rubin (1974)'s potential outcome notation and have assumed the stable unit treatment value assumption (Rubin, 1980). The goal is to estimate the *average treatment effect* (ATE), $\Delta := \mathbb{E}_0[Y(1)] - \mathbb{E}_0[Y(0)]$. As before, $\mathbb{E}_0$ means the expectation is taken over the data generating distribution F_0. Notice that $Y_i(0)$ and $Y_i(1)$ are never observed together, which is often referred to as the "fundamental problem of causal inference" (Holland, 1986). With the potential outcome notation, the observational studies problem is essentially two missing data problems: use $(A_i, \boldsymbol{X}_i, A_i Y_i) = (A_i, \boldsymbol{X}_i, A_i Y_i(1))$ to estimate $\mu(1) := \mathbb{E}_0[Y(1)]$, and use $(1 - A_i, \boldsymbol{X}_i, (1 - A_i) Y_i) = (1 - A_i, \boldsymbol{X}_i, (1 - A_i) Y_i(0))$ to estimate $\mu(0) := \mathbb{E}_0[Y(0)]$.

Since some responses or potential outcomes are not observed, the estimands μ and Δ defined above are not identifiable without further assumptions. For the missing data problem, Rubin (1976) used the term "missing at random" (MAR) to describe data that are missing for reasons related to completely observed variables in the data set:

Assumption 1 (Missing at random (MAR))**.** $A \perp\!\!\!\perp Y | \boldsymbol{X}$ *under* F_0.

The corresponding assumption for observational studies is that the potential outcomes $(Y(0), Y(1))$ are independent of the treatment A given the covariates $\boldsymbol{X}$ (Rosenbaum and Rubin, 1983):

Assumption 2 (Strong ignorability or no unmeasured confounder (NUC))**.** $A \perp\!\!\!\perp (Y(0), Y(1)) | \boldsymbol{X}$ *under* F_0.

With the additional assumption that no subject is missing or receives treatment/control with probability 1, the parameters μ and Δ are identifiable from the data.

Assumption 3 (Overlap)**.** *For* $\boldsymbol{x} \in \mathscr{X}$*,*

(1) *In the missing data problem,* $e_0(\boldsymbol{x}) := \mathbb{P}_0(A = 1 | \boldsymbol{X} = \boldsymbol{x}) \in (0, 1]$.
(2) *In observational studies,* $e_0(\boldsymbol{x}) \in (0, 1)$.

Since the seminal works of Rubin (1976) and Rosenbaum and Rubin (1983), many statistical methods have been developed for the missing data and observational studies problem based on first estimating the conditional probability $e_0(\boldsymbol{x})$ (often called the *propensity score* in

observational studies). One distinguished example is the *inverse probability weighting* (IPW) estimator which dates back to Horvitz and Thompson (1952),

$$\hat{\mu}_{\text{IPW}} = \frac{1}{n}\sum_{s=1}^{n}\frac{A_i Y_i}{\hat{e}(\boldsymbol{X}_i)} \quad \text{and} \quad \hat{\Delta}_{\text{IPW}} = \frac{1}{n}\sum_{s=1}^{n}\frac{A_i Y_i}{\hat{e}(\boldsymbol{X}_i)} - \frac{(1-A_i)Y_i}{1-\hat{e}(\boldsymbol{X}_i)},$$

where $\hat{e}(\boldsymbol{X})$ is a sample-estimate of $e_0(\boldsymbol{X})$. Observe that under Assumptions 1 and 3,

$$\mu = \mathbb{E}_0\left[\frac{AY}{\mathbb{P}_0(A=1|\boldsymbol{X},Y)}\right] = \mathbb{E}_0\left[\frac{AY}{\mathbb{P}_0(A=1|\boldsymbol{X})}\right] = \mathbb{E}_0\left[\frac{AY}{e_0(\boldsymbol{X})}\right], \tag{1}$$

which implies that $\hat{\mu}_{\text{IPW}}$ can consistently estimate the mean response μ if $\hat{e}$ converges to e_0. Notice that the first equality in (1) is due to the tower property of conditional expectation and is always true if the denominator $\mathbb{P}_0(A=1|\boldsymbol{X},Y)$ is positive with probability 1. The second equality in (1) uses $\mathbb{P}_0(A=1|\boldsymbol{X},Y) = \mathbb{P}_0(A=1|\boldsymbol{X}) \neq 0$ by Assumptions 1 and 3. Similarly, $\hat{\Delta}_{\text{IPW}}$ can consistently estimate Δ under Assumptions 2 and 3.

3. Sensitivity models

A critical issue of the above approach is that the MAR and NUC assumptions are not verifiable using the data because some responses/potential outcomes are not observed. Therefore, if the MAR or NUC assumption is violated, the IPW estimator $\hat{\mu}_{\text{IPW}}$ or $\hat{\Delta}_{\text{IPW}}$ is biased and the confidence interval based on it (and any other existing estimator that assumes MAR or NUC) does not cover μ or Δ at the nominal rate. In practice, statisticians often hope the data analyst can use her subject knowledge to justify the substantive reasonableness of MAR/NUC assumptions (Little and Rubin, 2014). However, speaking realistically, the MAR and NUC assumptions are perhaps never strictly satisfied, so the statistical inference is always subject to criticism of insufficient control of confounding.

To rebut such criticism and make the statistical analysis more credible, a natural question is how sensitive the results are to the violation of the MAR/NUC assumptions. This is clearly an important question, and, in fact, sensitivity analysis is often suggested or required for empirical publications.[1] Many sensitivity analysis methods have been subsequently developed in the missing data and observational studies problems. In this paper we will consider a sensitivity model that is closely related to Rosenbaum's sensitivity model (Rosenbaum, 1987, 2002c). We refer the reader to Robins (1999), Scharfstein et al. (1999), Imbens (2003), Altonji et al. (2005), Hudgens and Halloran (2006), Vansteelandt et al. (2006), McCandless et al. (2007), VanderWeele and Arah (2011), Richardson et al. (2014), Ding and VanderWeele (2016) for alternative sensitivity analysis methods and Section 7 for a more detailed discussion.

We begin with a description of the marginal sensitivity model used in this paper, which is also considered by Tan (2006). To this end, with a slight abuse of notation, let $e_0(\boldsymbol{x}, y) = \mathbb{P}_0(A=1|\boldsymbol{X}=\boldsymbol{x}, Y=y)$. Notice that, when showing the IPW estimator is consistent in (1), the MAR assumption is useful because it implies that

$$e_0(\boldsymbol{x}, y) = e_0(\boldsymbol{x}). \tag{2}$$

Similarly, the NUC assumption in the observational studies problem implies that $e_a(\boldsymbol{x}, y) = e_a(\boldsymbol{x})$, where $e_a(\boldsymbol{x}, y) := \mathbb{P}_0(A=1|\boldsymbol{X}=\boldsymbol{x}, Y(a)=y)$, for $a \in \{0,1\}$.

[1]For example, sensitivity analysis is required in the Patient-Centered Outcome Research Institute (PCORI) methodology standards for handling missing data, see standard MD-4 in `https://www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-standards`. See also Little et al. (2012).

When the MAR assumption is violated, equation (2) is no longer valid. The following well-known proposition (proof included in Appendix A.1 for completeness) shows that $e_0(\boldsymbol{x}, y)$ is generally not identifiable from the data without the MAR assumption. For this reason we shall refer to a user-specified function $e_0(\boldsymbol{x}, y)$ as a *sensitivity model*.

Proposition 1. *In the missing data problem and assuming Assumption 3 holds,* $e_0(\boldsymbol{x}, y) = \mathbb{P}_0(A = 1|\boldsymbol{X} = \boldsymbol{x}, Y = y)$ *is not identifiable from the data.*

In this paper we consider the following collection of sensitivity models in which the degree of violation of the MAR assumption is quantified by the odd ratio of the "complete data" selection probability $e_0(\boldsymbol{x}, y)$ and the "observed data" selection probability $e_0(\boldsymbol{x})$.

Definition 1 (Marginal Sensitivity Model)**.** Fix a parameter $\Lambda \geq 1$.

(1) For the missing data problem, we assume $e(\boldsymbol{x}, y) \in \mathcal{E}(\Lambda)$, where

$$\mathcal{E}(\Lambda) = \Big\{ e(\boldsymbol{x}, y) : \frac{1}{\Lambda} \leq \mathrm{OR}(e(\boldsymbol{x}, y), e_0(\boldsymbol{x})) \leq \Lambda, \text{ for all } \boldsymbol{x} \in \mathscr{X}, y \in \mathbb{R} \Big\}. \tag{3}$$

(2) For the observational studies problem, we assume $e_a(\boldsymbol{x}, y) \in \mathcal{E}(\Lambda)$ for $a = 0, 1$.

The relationship between this sensitivity model and Rosenbaum's sensitivity model is examined in Section 7.2.

Remark 1. To understand Definition 1, it can be conceptually easier to imagine that there is an unobserved variable U that "summarizes" all unmeasured confounding. Definition 4 means that the odds ratio between $\mathbb{P}_0(A|\boldsymbol{X}, U = u_1)$ and $\mathbb{P}_0(A|\boldsymbol{X}, U = u_2)$ is always bounded by Λ^{-1} and Λ, but the relation between U and the potential outcomes are not constrained in any way. This leads to an alternative "added variable" representation of sensitivity model (Rosenbaum, 2002c). Robins (2002) pointed out that it suffices to consider the conditional probabilities when U is either one of the potential outcomes. In the remainder of the paper we will follow Robin's suggestion to simplify the notation.

Remark 2. It is often convenient to use the logistic representation of the marginal sensitivity model (3). For simplicity, let us consider the missing data problem. Denote

$$g_0(\boldsymbol{x}) = \mathrm{logit}(e_0(\boldsymbol{x})) = \log \frac{e_0(\boldsymbol{x})}{1 - e_0(\boldsymbol{x})} \quad \text{and} \quad g_0(\boldsymbol{x}, y) = \mathrm{logit}(e_0(\boldsymbol{x}, y)), \tag{4}$$

and $h_0(\boldsymbol{x}, y) = g_0(\boldsymbol{x}) - g_0(\boldsymbol{x}, y)$ be the logit-scale difference of the observed data selection probability and the complete data selection probability. If we write further introduce the notation $e^{(h)}(\boldsymbol{x}, y) = [1 + \exp(h(\boldsymbol{x}, y) - g_0(\boldsymbol{x}))]^{-1}$, then

$$\mathcal{E}(\Lambda) = \{e^{(h)}(\boldsymbol{x}, y) : h \in \mathcal{H}(\lambda)\}, \text{ where } \mathcal{H}(\lambda) = \{h : \mathscr{X} \times \mathbb{R} \to \mathbb{R} \text{ and } ||h||_\infty \leq \lambda\}, \tag{5}$$

that is the marginal sensitivity model puts bound on the L_∞-norm of h.

4. Confidence Interval for the Mean Response

The primary goal of this paper is to construct confidence intervals for μ and Δ under a specified collection of sensitivity models. For simplicity, we focus on the missing data problem in this section. The simple extension to observational studies is described in Section 5.

4.1. **Confidence interval in sensitivity analysis.** We begin by briefly extending the definition of sensitivity models. In Definition 1, a postulated sensitivity model $e_0(\boldsymbol{x}, y)$ is compared with the data-identifiable $e_0(\boldsymbol{x}) = \mathbb{P}_0(A = 1|\boldsymbol{X} = \boldsymbol{x})$. This model is most appropriate when $e_0(\boldsymbol{X})$ is estimated non-parametrically using the data, but this is only possible for low-dimensional problems due to the curse of dimensionality. More often, for example, in propensity score matching or IPW, $e_0(\boldsymbol{x})$ is estimated by a parametric model. In this case it is more appropriate to compare the postulated sensitivity model $e_0(\boldsymbol{x}, y)$ with the best parametric approximation to $e_0(\boldsymbol{x})$:

$$\begin{aligned} e_{\boldsymbol{\beta_0}}(\boldsymbol{x}) &= \arg\min_{\boldsymbol{\beta}\in\Theta} \mathrm{KL}\big(\mathbb{P}_0(A = 1|\boldsymbol{X} = \boldsymbol{x}) \,\big|\big|\, \mathbb{P}_{\boldsymbol{\beta}}(A = 1|\boldsymbol{X} = \boldsymbol{x})\big) \\ &= \arg\max_{\boldsymbol{\beta}\in\Theta} \mathbb{E}_0 \left[A \cdot \log e_{\boldsymbol{\beta}}(\boldsymbol{X}) + (1 - A) \cdot \log(1 - e_{\boldsymbol{\beta}}(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right], \end{aligned} \tag{6}$$

where KL stands for the Kullback-Leibler divergence and $\{e_{\boldsymbol{\beta}}(\boldsymbol{x}) := \mathbb{P}_{\boldsymbol{\beta}}(A = 1|\boldsymbol{X} = \boldsymbol{x}),\ \boldsymbol{\beta} \in \Theta \subset \mathbb{R}^d\}$ is a family of parametric models for the missingness probability. We recommend to use the logistic regression model

$$e_{\boldsymbol{\beta}}(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}}},\ \text{or equivalently}\ g_{\boldsymbol{\beta}}(\boldsymbol{x}) = \mathrm{logit}(e_{\boldsymbol{\beta}}(\boldsymbol{x})) = \boldsymbol{\beta}'\boldsymbol{x}, \tag{7}$$

which works seamlessly with our sensitivity model because the degree of violation of MAR is quantified by odds ratios. However our framework can be easily applied to other parametric models $e_{\boldsymbol{\beta}}(\boldsymbol{x})$ (e.g. different links).

Definition 2 (Parametric Marginal Sensitivity Model)**.** Fix a parameter a $\Lambda \geq 1$, this collection of sensitivity models assumes the true missingness probability $e_0(\boldsymbol{x}, y) = \mathrm{P}_0(A = 1|\boldsymbol{X} = \boldsymbol{x}, Y = y)$ satisfies

$$e_0(\boldsymbol{x}, y) \in \mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda) := \Big\{e(\boldsymbol{x}, y) :\ \frac{1}{\Lambda} \leq \mathrm{OR}(e(\boldsymbol{x}, y), e_{\boldsymbol{\beta}_0}(\boldsymbol{x})) \leq \Lambda \text{ for all } \boldsymbol{x} \in \mathscr{X}, y \in \mathbb{R}\Big\}. \tag{8}$$

As in (5), the constraint (8) is equivalent to $h_{\boldsymbol{\beta}_0}(\boldsymbol{x}, y) \in \mathcal{H}(\lambda)$ for $h_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = g_{\boldsymbol{\beta}}(\boldsymbol{x}) - g_0(\boldsymbol{x}, y)$ and $\lambda = \log\Lambda$.

Next we define the estimand (mean response) under a specific sensitity model h:

$$\begin{aligned} \mu^{(h)} &= \left(\mathbb{E}_0\left[\frac{A}{e^{(h)}(\boldsymbol{X}, Y)}\right]\right)^{-1} \mathbb{E}_0\left[AY\left(1 + e^{h(\boldsymbol{X},Y) - g_{\boldsymbol{\beta}_0}(\boldsymbol{X})}\right)\right] \\ &= \left(\mathbb{E}_0\left[\frac{A}{e^{(h)}(\boldsymbol{X}, Y)}\right]\right)^{-1} \mathbb{E}_0\left[\frac{AY}{e^{(h)}(\boldsymbol{X}, Y)}\right] \end{aligned} \tag{9}$$

where $e^{(h)}(\boldsymbol{x}, y) = (1 + e^{h(\boldsymbol{x},y) - g_{\boldsymbol{\beta}_0}(\boldsymbol{x})})^{-1}$. By the definition of $h_{\boldsymbol{\beta}}$ and the first equality in (1), it is easy to see that $\mathbb{E}_0\left[\frac{A}{e^{(h_{\boldsymbol{\beta}_0})}(\boldsymbol{X},Y)}\right] = 1$ and $\mu^{(h_{\boldsymbol{\beta}_0})}$ are equal to the true mean response μ. The set $\{\mu^{(h)} :\ e^{(h)} \in \mathcal{E}_{\boldsymbol{\beta}_0}(\lambda)\}$ will be referred to as the partially identifiable region under $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$; see Section 7.3 for more discussion.

Remark 3. In defining $\mu^{(h)}$ and $e^{(h)}$, we slightly abused the notation and did not indicate that they also depend on the "baseline" choice of parametric or nonparametric model of $e_0(\boldsymbol{X})$. This is because the choice of the sensitivity model is a natural consequence of the choice of the working model for $e_0(\boldsymbol{X})$. If $e_0(\boldsymbol{X})$ is modeled nonparametrically (or parametrically), then we should use the collection of nonparametric (or parametric) sensitivity models $\mathcal{E}(\Lambda)$ (or $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$). For notational simplicity, hereafter we will often refer to $h(\boldsymbol{x}, y) \in \mathcal{H}(\lambda)$ instead of $e^{(h)}(\boldsymbol{x}, y)$ as the sensitivity model, since the former does not depends on the choice of

the working model for $e_0(\boldsymbol{X})$. Consequently we will also call $\mathcal{H}(\lambda)$ a collection of sensitivity models without specifying which parametric/nonparametric model was used for $e_0(\boldsymbol{X})$.

Remark 4. Compared to Definition 4, the only difference in (8) is that the postulated sensitivity model $e_0(\boldsymbol{x}, y)$ is compared to the parametric model $e_{\boldsymbol{\beta}_0}(\boldsymbol{x})$ instead of the non-parametric probability $e_0(\boldsymbol{x})$. In other words, the parametric sensitivity model considers both

(1) Model misspecification, that is, $e_{\boldsymbol{\beta}_0}(\boldsymbol{x}) \neq e_0(\boldsymbol{x})$; and
(2) Missing not at random, that is, $e_0(\boldsymbol{x}) \neq e_0(\boldsymbol{x}, y)$.

This is arguably a desirable feature, because the term "sensitivity analysis" is also widely used as the analysis of an empirical study's robustness to parametric modeling assumptions.

With these notations, we can now define what is meant by a confidence interval under a collection of sensitivity models:

Definition 3. A data-dependent interval $[L, U]$ is called a $(1-\alpha)$-confidence interval for the mean response μ under the collection of sensitivity models $\mathcal{H}(\lambda)$ (may corresponds to $\mathcal{E}(\Lambda)$ or $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$, see Remark 3), if

$$\mathbb{P}_0(\mu \in [L, U]) \geq 1 - \alpha \tag{10}$$

is true for any data generating distribution F_0 such that $h_0 \in \mathcal{H}(\lambda)$ or $h_{\boldsymbol{\beta}_0} \in \mathcal{H}(\lambda)$, depending on whether $\mathcal{E}(\Lambda)$ or $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$ is used. The interval is said to be an asymptotic $(1-\alpha)$-confidence interval, if $\liminf_{n\to\infty} \mathbb{P}_0(\mu \in [L, U]) \geq 1 - \alpha$.

4.2. **The IPW Point Estimates.** Intuitively, a confidence interval $[L, U]$ as in Definition 3 must at least include a point estimate of $\mu^{(h)}$ for every $h \in \mathcal{H}(\lambda)$. To this end, let h be a postulated sensitivity model. The corresponding missingness probability $e^{(h)}(\boldsymbol{x}, y)$ can then be estimated by

$$\hat{e}^{(h)}(\boldsymbol{x}, y) = \frac{1}{1 + e^{h(\boldsymbol{x},y) - \hat{g}(\boldsymbol{x},y)}}, \tag{11}$$

where

$$\hat{g}(\boldsymbol{x}) = \begin{cases} \operatorname{logit}(\hat{\mathbb{P}}(A = 1 | \boldsymbol{X} = x)) & \text{if} \quad e^{(h)}(\boldsymbol{x}, y) \in \mathcal{E}(\Lambda) \\ \operatorname{logit}(\mathbb{P}_{\hat{\boldsymbol{\beta}}}(A = 1 | \boldsymbol{X} = x)) = \hat{\boldsymbol{\beta}}' \boldsymbol{x} & \text{if} \quad e^{(h)}(\boldsymbol{x}, y) \in \mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda). \end{cases} \tag{12}$$

We consider two point estimates of $\tilde{\mu}^{(h)}$ or $\mu^{(h)}$, namely the IPW estimate and a stabilized version of it:

(1) The IPW estimator of $\tilde{\mu}^{(h)}$ is

$$\hat{\mu}^{(h)}_{\mathrm{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\hat{e}^{(h)}(\boldsymbol{X}_i, Y_i)},$$

where $\hat{e}^{(h)}$ is as in (11).

(2) It is well known that, even under the MAR assumption, the IPW estimator can be unstable when the missingness probability $e_0(\boldsymbol{x})$ (or the parametric approximation $e_{\boldsymbol{\beta}_0}(\boldsymbol{x})$) is close to 0 for some $\boldsymbol{x} \in \mathscr{X}$ (Kang and Schafer, 2007). To alleviate this issue, the *stabilized IPW* (SIPW) estimator, obtained by normalizing the weights, is often used in practice:

$$\hat{\mu}^{(h)} = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{\hat{e}^{(h)}(\boldsymbol{X}_i, Y_i)} \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\hat{e}^{(h)}(\boldsymbol{X}_i, Y_i)} \right]. \tag{13}$$

It is easy to see that $\hat{\mu}^{(h)}$ estimates $\mu^{(h)}$.

Compared to IPW, the SIPW estimator is sample bounded (Robins et al., 2007), that is,

$$\hat{\mu}^{(h)} \in \left[\min_{i:A_i=1} Y_i, \max_{i:A_i=1} Y_i\right].$$

This property is even more desirable in sensitivity analysis because $e^{(h)}$ is almost always not the true missingness probability, so the total unnormalized weights $\frac{1}{n}\sum_{i=1}^n A_i/\hat{e}^{(h)}(\boldsymbol{X}_i, Y_i)$ can be very different from 1. For this reason, we will use the SIPW estimator in the remainder of this paper.

Heuristically, the confidence interval $[L, U]$ should at least contain the range of SIPW point estimates, $\left[\inf_{h\in\mathcal{H}(\lambda)} \hat{\mu}^{(h)}, \sup_{h\in\mathcal{H}(\lambda)} \hat{\mu}^{(h)}\right]$. We defer the numerical computation of the extrema of SIPW point estimates till Section 4.4. For now we will focus on constructing the confidence interval $[L, U]$ assuming the range of point estimates is given.

4.3. **Constructing the Confidence Interval.** To construct a confidence interval for μ, we need to consider the sampling variability of the SIPW estimator described above. In the MAR setting, the most common way to estimate the variance is the asymptotic sandwich formula or the bootstrap (Efron and Tibshirani, 1994, Austin, 2016). In sensitivity analysis, we also need to consider all possible violations of the MAR assumption in $\mathcal{H}(\lambda)$. In this case, optimizing the estimated asymptotic variance over $\mathcal{H}(\lambda)$ is generally computationally intractable, as explained below.

4.3.1. *The Union Method.* We begin by showing how individual confidence intervals of $\mu^{(h)}$ for $h \in H(\lambda)$ can be combined into a confidence interval in sensitivity analysis.

Proposition 2. *Suppose there exists data-dependent intervals $[L^{(h)}, U^{(h)}]$ such that*

$$\liminf_{n\to\infty} \mathbb{P}_0(\mu^{(h)} \in [L^{(h)}, U^{(h)}]) \geq 1-\alpha$$

holds for every $h \in \mathcal{H}(\lambda)$.

1. *Let $L = \inf_{h\in\mathcal{H}(\lambda)} L^{(h)}$, $U = \sup_{h\in\mathcal{H}(\lambda)} U^{(h)}$. Then $[L, U]$ is an asymptotic $(1-\alpha)$-confidence interval of μ under the collection of sensitivity models $\mathcal{H}(\lambda)$.*
2. *Moreover, if there exists $\alpha' \in [0, \alpha]$ (not depending on h) such that*

$$\limsup_{n\to\infty} \mathbb{P}_0\big(\mu^{(h)} < L^{(h)}\big) \leq \alpha' \quad \textit{and} \quad \limsup_{n\to\infty} \mathbb{P}_0\big(\mu^{(h)} > U^{(h)}\big) \leq \alpha - \alpha', \tag{14}$$

for all $h \in \mathcal{H}(\lambda)$, then the union interval $[L, U]$ covers the partially identified region with probability at least $1-\alpha$,

$$\liminf_{n\to\infty} \mathbb{P}_0\left(\{\mu^{(h)} : h \in \mathcal{H}(\lambda)\} \subseteq [L, U]\right) \geq 1-\alpha.$$

Proposition 2 suggests the following way to construct a confidence interval for μ in sensitivity analysis using the asymptotic distribution of $\hat{\mu}_1^{(h)}$. Using the general theory of Z-estimation, it is not difficult to establish that

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

(see, for example, Lunceford and Davidian (2004) or Corollary 9 in the Appendix). Then using the sandwich variance estimator $(\hat{\sigma}^{(h)})^2$, an asymptotically $(1-\alpha)$-confidence interval of $\mu^{(h)}$ is

$$[L_{\text{sand}}^{(h)}, U_{\text{sand}}^{(h)}] = \left[\hat{\mu}^{(h)} - z_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^{(h)}}{\sqrt{n}}, \hat{\mu}^{(h)} + z_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^{(h)}}{\sqrt{n}}\right],$$

where $z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$-quantile. Then, by Proposition 2, an asymptotically $(1-\alpha)$-confidence interval under the collection of sensitivity models is $[L_{\text{sand}}, U_{\text{sand}}]$,

$$L_{\text{sand}} = \inf_{h \in \mathcal{H}(\lambda)} L^{(h)}_{\text{sand}}, \quad U_{\text{sand}} = \sup_{h \in \mathcal{H}(\lambda)} U^{(h)}_{\text{sand}}.$$

However, the standard error $\hat{\sigma}^{(h)}$ is a very complicated function of h (see Corollary 9) and numerical optimization over $h \in \mathcal{H}(\lambda)$ is practically infeasible.

4.3.2. *The Percentile Bootstrap.* The centerpiece of our proposal is to use the percentile bootstrap to construct $[L^{(h)}, U^{(h)}]$. Next we introduce the necessary notations to describe the bootstrap procedure. Let $\mathbb{P}_n$ be the empirical measure of the sample $\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_n$, where $\boldsymbol{T}_i = (A_i, \boldsymbol{X}_i', A_i Y_i)$, and $\hat{\boldsymbol{T}}_1, \hat{\boldsymbol{T}}_2, \ldots, \hat{\boldsymbol{T}}_n$ be i.i.d. resamples from the empirical measure. Let $\hat{\hat{\mu}}^{(h)}$ the SIPW estimate (13) computing using the bootstrap resamples $\{\hat{\boldsymbol{T}}_i\}_{i \in [n]}$. For $h \in \mathcal{H}(\lambda)$, the percentile bootstrap confidence interval of $\mu^{(h)}$ is given by

$$[L^{(h)}, U^{(h)}] = \left[Q_{\frac{\alpha}{2}}(\hat{\hat{\mu}}^{(h)}), Q_{1-\frac{\alpha}{2}}(\hat{\hat{\mu}}^{(h)}) \right], \tag{15}$$

where $Q_\alpha(\hat{\hat{\mu}})$ is the α-percentile of $\hat{\hat{\mu}}$ in the bootstrap distribution, that is,

$$Q_\alpha(\hat{\hat{\mu}}) := \inf\{t : \hat{\mathbb{P}}_n(\hat{\hat{\mu}} \le t) \ge \alpha\},$$

where $\hat{\mathbb{P}}_n$ is the bootstrap resampling distribution.

We begin by showing that the percentile bootstrap interval $[L^{(h)}, U^{(h)}]$ is an asymptotically valid confidence interval of $\mu^{(h)}$ for the parametric sensitivity model $e^{(h)} \in \mathcal{E}_{\beta_0}(\Lambda)$. Notice that bootstrap is generally not valid if the missingness probability is modeled nonparametrically (Abadie and Imbens, 2008).

Theorem 3 (Validity of the Percentile Bootstrap)**.** *In the logistic model* (7) *and under Assumption 4 in Appendix B, for every* $e^{(h)} \in \mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$ *we have*

$$\limsup_{n \to \infty} \mathbb{P}_0 \left(\mu^{(h)} < L^{(h)} \right) \le \frac{\alpha}{2} \quad \textit{and} \quad \limsup_{n \to \infty} \mathbb{P}_0 \left(\mu^{(h)} > U^{(h)} \right) \le \frac{\alpha}{2}.$$

The proof of this result, which invokes the general theory of bootstrap for Z-estimators (Wellner and Zhan, 1996, Kosorok, 2006, Chapter 10), is explained in Appendix B.

Our percentile bootstrap confidence interval under the collection of sensitivity models $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$ is given by $[L, U]$ where

$$L = Q_{\frac{\alpha}{2}}\Big(\inf_{h \in \mathcal{H}(\lambda)} \hat{\hat{\mu}}^{(h)} \Big) \quad \text{and} \quad U = Q_{1-\frac{\alpha}{2}}\Big(\sup_{h \in \mathcal{H}(\lambda)} \hat{\hat{\mu}}_b^{(h)} \Big). \tag{16}$$

The important thing to observe here is that the infimum/supremum is inside the quantile function in (16), which makes the computation especially efficient using linear programming (see Section 4.4). The interchange of quantile and infimum/supremum is justified in the following generalized (von Neumann's) minimax/maximin inequalities (see Cohen (2013) for a similar result for finite sets).

Lemma 1. *Let* L, U *be as defined in* (17)*. Then*

$$L \le \inf_{h \in \mathcal{H}(\lambda)} L^{(h)} \quad \textit{and} \quad U \ge \sup_{h \in \mathcal{H}(\lambda)} U^{(h)}.$$

Asymptotic validity of the confidence interval $[L, U]$ in Equation (16) then immediately follows from the validity of the union method (Proposition 2), the validity of the percentile bootstrap (Theorem 3), and Lemma 1.

Theorem 4. *Under the same assumptions as in Theorem 3, $[L, U]$ is an asymptotic $(1-\alpha)$-confidence interval of the mean response μ, under the collection of sensitivity models $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$. Furthermore, $[L, U]$ covers the partially identified region $\{\mu^{(h)} : e^{(h)} \in \mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)\}$ with probability at least $1-\alpha$.*

4.4. **Range of SIPW Point Estimates: Linear Fractional Programming.** Theorem 4 transformed the sensitivity analysis problem to computing the extrema of the SIPW point estimates, $\inf_{h\in\mathcal{H}(\lambda)} \hat{\hat{\mu}}_b^{(h)}$ and $\sup_{h\in\mathcal{H}(\lambda)} \hat{\hat{\mu}}_b^{(h)}$. In practice, we only repeat this over B ($\ll N$) random resamples and compute the interval by

$$L_B = Q_{\frac{\alpha}{2}}\left(\left(\inf_{h\in\mathcal{H}(\lambda)} \hat{\hat{\mu}}_b^{(h)}\right)_{b\in[B]}\right), \quad U_B = Q_{1-\frac{\alpha}{2}}\left(\left(\sup_{h\in\mathcal{H}(\lambda)} \hat{\hat{\mu}}_b^{(h)}\right)_{b\in[B]}\right). \tag{17}$$

For notational simplicity, below we consider how to compute the extrema $\inf_{h\in\mathcal{H}(\lambda)} \hat{\mu}^{(h)}$ and $\sup_{h\in\mathcal{H}(\lambda)} \hat{\mu}^{(h)}$ using the full observed data instead of the resampled data. Recalling (12) and (13), this is equivalent to solving

$$\text{max or min } \frac{\sum_{i=1}^n A_i Y_i \left(1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)}\right)}{\sum_{i=1}^n A_i \left(1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)}\right)} \text{ subject to } z_i \in \left[\Lambda^{-1}, \Lambda\right], \text{ for all } i \in [n], \tag{18}$$

where the optimization variables are $z_i = e^{h(\boldsymbol{X}_i, Y_i)}$ for $i \in [n]$. All the other variables are observed or can be estimated from the data. Notice that $\hat{g}(\boldsymbol{x})$ needs to be re-estimated in every bootstrap resample. Without loss of generality, assume that the first $1 \le m < n$ responses are observed, that is $A_1 = A_2 = \cdots A_m = 1$ and $A_{m+1} = \cdots = A_n = 0$, and suppose that the observed responses are in decreasing order, $Y_1 \ge Y_2 \ge \cdots \ge Y_m$. Then (18) simplifies to

$$\text{max or min } \frac{\sum_{i=1}^m Y_i \left(1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)}\right)}{\sum_{i=1}^m \left(1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)}\right)}, \text{ subject to } z_i \in \left[\Lambda^{-1}, \Lambda\right], \ 1 \le i \le m. \tag{19}$$

This optimization problem is the ratio of two linear functions of the decision variables $\boldsymbol{z}$, hence called a *linear fractional programming*. It can be transformed to linear programming by the Charnes-Cooper transformation (Charnes and Cooper, 1962). Denote

$$\bar{z}_i = \frac{z_i}{\sum_{i=1}^m (1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)})}, \text{ for } 1 \le s \le m, \text{ and } t = \frac{1}{\sum_{i=1}^m (1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)})}.$$

This translates (19) into the following linear programming:

$$\begin{aligned} \text{max or min} \quad & \sum_{i=1}^m Y_i e^{-\hat{g}(\boldsymbol{X}_i)} \bar{z}_i + t\left(\sum_{i=1}^m Y_i\right) \\ \text{subject to} \quad & t \ge 0, \ \Lambda^{-1} t \le \bar{z}_i \le \Lambda t, \text{ for } 1 \le i \le m, \\ & \sum_{i=1}^m e^{-\hat{g}(\boldsymbol{X}_i)} \bar{z}_i + t\left(\sum_{i=1}^m A_i\right) = 1. \end{aligned} \tag{20}$$

Therefore, the range of the SIPW point estimate can be computed efficiently by solving the above linear program. Furthermore, the following result shows that the solution of (20) must have the same or opposite order as the outcomes $\boldsymbol{Y}$, which enables even faster computation of (20).

Proposition 5. *Suppose $(z_i)_{i=1}^m$ solves the maximization problem in (20). Then $(z_i)_{i=1}^m$ has the same order as $(Y_i)_{i=1}^m$, that is, if $Y_{s_1} > Y_{s_2}$ then $z_{s_1} > z_{s_2}$, for $1 \le s_1 \ne s_2 \le m$. Furthermore, there exists a solution $(z_i)_{i=1}^m$ and a threshold M such that $z_i = \Lambda$, if $Y_i \ge M$, and $z_i = \frac{1}{\Lambda}$, if $Y_i < M$. The same conclusion holds for the minimizer in (20) with $(Y_i)_{i=1}^m$ replaced by $(-Y_i)_{s=1}^m$.*

Proposition 5 implies that we only need to compute the objective of (19) for at most m choices of $(z_i)_{i\in[m]}$ by enumerating the index where it changes from Λ to Λ^{-1}:

$$\left\{(z_i)_{i\in[m]} : \exists\ a \in [m] \text{ with } z_i = \Lambda,\ \text{for } 1 \le i \le a,\ \text{and } z_i = \frac{1}{\Lambda},\ \text{for } a+1 \le i \le m\right\}. \quad (21)$$

It is easy to see that we only need $O(m)$ time to compute $\sum_{i=1}^m Y_i(1+z_i e^{-\hat{g}(\boldsymbol{X}_i)})$ and $\sum_{i=1}^m (1+z_i e^{-\hat{g}(\boldsymbol{X}_i)})$, for all the m choices in (21). Hence, the computational complexity to solve the linear fractional programming (19) is $O(m)$. This is the best possible rate since it takes $O(m)$ time to just compute the objective once.

The above discussion suggests that the interval $[L_B, U_B]$, which entails finding $\inf_{h\in\mathcal{H}(\lambda)} \hat{\hat{\mu}}_b^{(h)}$ and $\sup_{h\in\mathcal{H}(\lambda)} \hat{\hat{\mu}}_b^{(h)}$ for every $1 \le b \le B$, can be computed extremely efficiently. The computational complexity is $O(nB + n\log n)$ ignoring the time spent to fit the logistic propensity score models, where the extra $n\log n$ is needed for sorting the data. Indeed, even under the MAR assumption, we need $O(nB)$ time to compute the Bootstrap confidence interval of μ , which is recommended in practice by Austin (2016). In conclusion, our proposal requires almost no extra cost to conduct a sensitivity analysis for the IPW estimator than to obtain its bootstrap confidence interval under MAR.

5. Confidence Interval for the ATE in the Sensitivity Model

As discussed in Section 2, an observational study is essentially two missing data problems, so the framework developed above can be easily extended. Recall that $e_a(\boldsymbol{x}, y) = \mathbb{P}_0(A = 1|\boldsymbol{X} = \boldsymbol{x}, Y(a) = y)$. Suppose we use a parametric model $e_{\boldsymbol{\beta}_0}(\boldsymbol{x})$ to model the propensity score $\mathbb{P}_0(A = 1|\boldsymbol{X} = \boldsymbol{x})$ using the observed covariates. The parametric sensitivity model assumes

$$\frac{1}{\Lambda} \le \mathrm{OR}(e_a(\boldsymbol{x}, y), e_{\boldsymbol{\beta}_0}(\boldsymbol{x})) \le \Lambda,\ \text{for all } \boldsymbol{x} \in \mathscr{X}, y \in \mathbb{R}, a \in \{0, 1\}. \quad (22)$$

Let $g_a(\boldsymbol{x}, y) = \mathrm{logit}(e_a(\boldsymbol{x}, y))$, $g_{\boldsymbol{\beta}}(\boldsymbol{x}) = \mathrm{logit}(e_{\boldsymbol{\beta}}(\boldsymbol{x}))$ and $h_a^{(\boldsymbol{\beta})}(\boldsymbol{x}, y) = g_{\boldsymbol{\beta}}(\boldsymbol{x}) - g_a(\boldsymbol{x}, y)$, for $a \in \{0, 1\}$. Then (22) is equivalent to assuming $h_{a,\boldsymbol{\beta}_0} \in \mathcal{H}(\lambda)$ for $a \in \{0, 1\}$.

As in (9) define, for $h \in \mathcal{H}(\lambda)$,

$$\mu^{(h)}(a) = \left(\mathbb{E}\left[\frac{A^a(1-A)^{1-a}}{e_a^{(h)}(\boldsymbol{X}, Y)}\right]\right)^{-1} \mathbb{E}_0\left[\frac{A^a(1-A)^{1-a}Y}{e_a^{(h)}(\boldsymbol{X}, Y)}\right], \quad (23)$$

where $e_a^{(h)}(\boldsymbol{x}, y) = \left[1 + e^{(-1)^{a+1}h(\boldsymbol{x},y) - \mathrm{logit}(\mathbb{P}_{\boldsymbol{\beta}_0}(A=a|\boldsymbol{X}=\boldsymbol{x}))}\right]^{-1}$. Denote $\Delta^{(h_0,h_1)} := \mu^{(h_1)}(1) - \mu^{(h_0)}(0)$. It is straightforward to show that $\Delta^{(h_0^{(\boldsymbol{\beta}_0)}, h_1^{(\boldsymbol{\beta}_0)})} = \Delta$. For this reason, as before, let

$$\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda) := \{(e_0^{(h_1)}(\boldsymbol{x}, y), e_1^{(h_2)}(\boldsymbol{x}, y)) : h_1, h_2 \in \mathcal{H}(\lambda)\} \quad (24)$$

be the collection of parametric sensitivity models in the observational studies problem. These can be estimated as in (11) by

$$\hat{e}_a^{(h)}(\boldsymbol{x}, y) = \frac{1}{1 + e^{(-1)^{a+1}[h(\boldsymbol{x},y) - \hat{g}(\boldsymbol{x})]}}, \quad (25)$$

where $\hat{g}(\boldsymbol{x}) = g_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x})$ is the estimated propensity score. Using this we can define the SIPW estimate of the average treatment effect as follows:

$$\hat{\Delta}^{(h_0,h_1)} := \hat{\mu}^{(h_1)}(1) - \hat{\mu}^{(h_0)}(0), \tag{26}$$

where

$$\hat{\mu}^{(h_1)}(1) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\hat{e}_1^{(h_1)}(\boldsymbol{X}_i, Y_i)}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{e}_1^{(h_1)}(\boldsymbol{X}_i, Y_i)}\right],$$

and $\hat{\mu}^{(h_1)}(0)$ defined analogously.

Now, as in Section 4.3.2, using the percentile bootstrap we can obtain a asymptotically valid interval for $\Delta^{(h_0,h_1)}$:

$$\left[Q_{\frac{\alpha}{2}}(\hat{\hat{\Delta}}^{(h_0,h_1)}), Q_{1-\frac{\alpha}{2}}(\hat{\hat{\Delta}}^{(h_0,h_1)})\right],$$

where $Q_{\frac{\alpha}{2}}(\hat{\hat{\Delta}}^{(h_0,h_1)})$ is the α-th bootstrap quantile of the SIPW estimates (26), as defined in Section 4.3.2. Then, interchanging the maximum/minimum and the quantile as in Theorem 6, we get a confidence interval for the average-treatment effect for the collection of sensitivity models (24).

Corollary 6. *Under the same assumptions in Theorem 3,*

$$\left[Q_{\frac{\alpha}{2}}\left(\inf_{h_0,h_1\in\mathcal{H}(\lambda)}\hat{\hat{\Delta}}^{(h_0,h_1)}\right), Q_{1-\frac{\alpha}{2}}\left(\inf_{h_0,h_1\in\mathcal{H}(\lambda)}\hat{\hat{\Delta}}^{(h_0,h_1)}\right)\right] \tag{27}$$

is an asymptotic $(1-\alpha)$*-confidence interval of the average treatment effect* Δ *under the collection of parametric sensitivity models* (24)*.*

The interval in (27) can be computed efficiently using linear fractional programming as in Section 4.4. To simplify notation, assume, without loss of generality, the first $m \leq n$ units are treated ($A = 1$) and the rest are the control ($A = 0$), and that the outcomes are ordered decreasingly among the first m units and the other $n-m$ units. Then, as in (19), computing the interval (27) is equivalent to solving the following optimization problem:

$$\begin{aligned} \text{maximize or minimize} \quad & \frac{\sum_{i=1}^{m} Y_i(1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)})}{\sum_{i=1}^{m}(1 + z_i e^{-\hat{g}(\boldsymbol{X}_i)})} - \frac{\sum_{i=m+1}^{n} Y_i(1 + z_i e^{\hat{g}(\boldsymbol{X}_i)})}{\sum_{i=m+1}^{n}(1 + z_i e^{\hat{g}(\boldsymbol{X}_i)})} \\ \text{subject to} \quad & \frac{1}{\Lambda} \leq z_i \leq \Lambda, \text{ for } 1 \leq i \leq n, \end{aligned} \tag{28}$$

where $z_i = e^{h_1(\boldsymbol{X}_i, Y_i)}$, for $1 \leq i \leq m$ and $z_i = e^{-h_0(\boldsymbol{X}_i, Y_i)}$, for $m+1 \leq i \leq n$. Note that the variables $(z_i)_{i=1}^m$ and $(z_i)_{i=m+1}^n$ are separable in (28), so we can solve the maximization/minimization problem in (28) by solving one maximization/minimization problem for $(z_i)_{i=1}^m$ and one minimization/maximization problem for $(z_i)_{i=m+1}^n$. Therefore, similar to the missing data problem, the time complexity to obtain the range of the SIPW estimates $\hat{\Delta}$, over a range of B bootstrap resamples, is only $O(nB + n\log n)$.

6. Extensions

In this section we discuss three extensions of the general framework described in Section 4.

6.1. **Mean of the Non-Respondents and Average Treatment Effect on the Treated.** In many applications, it is also interesting to estimate the *mean of the non-respondents* $\mu_0 = \mathbb{E}_0[Y|A=0]$, and the *average treatment effect on the treated* (ATT) $\Delta_1 = \mathbb{E}_0[Y(1)-Y(0)|A=1]$. The method described in Section 4 can be easily applied to these estimands, as described below.

To begin with note that

$$\mu_0 = \mathbb{E}_0[Y|A=0] = \frac{\mathbb{E}_0[Y\cdot 1\{A=0\}]}{\mathbb{P}_0(A=0)} = \frac{\mathbb{E}_0\left[(1-e_0(\boldsymbol{X},Y))Y\right]}{\mathbb{P}_0(A=0)} = \frac{\mathbb{E}_0\left[\frac{1-e_0(\boldsymbol{X},Y)}{e_0(\boldsymbol{X},Y)}AY\right]}{\mathbb{P}_0(A=0)}.$$

Then as in Theorem 4, under the collection of parametric sensitivity models $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$,

$$\left[Q_{\frac{\alpha}{2}}\left(\left(\inf_{h\in\mathcal{H}(\lambda)}\hat{\hat{\mu}}_{0b}^{(h)}\right)_{b\in[B]}\right),\ Q_{1-\frac{\alpha}{2}}\left(\left(\inf_{h\in\mathcal{H}(\lambda)}\hat{\hat{\mu}}_{0b}^{(h)}\right)_{b\in[B]}\right)\right], \tag{29}$$

is an asymptotic $(1-\alpha)$-confidence interval of the non-respondent mean μ_0, where the SIPW estimate is

$$\hat{\mu}_0^{(h)} = \frac{\sum_{i=1}^n e^{h(\boldsymbol{X}_i,Y_i)-\hat{g}(\boldsymbol{X}_i)}A_iY_i}{\sum_{i=1}^n e^{h(\boldsymbol{X}_i,Y_i)-\hat{g}(\boldsymbol{X}_i)}A_i},$$

where $\hat{g}$ is as in (12) and $\hat{\hat{\mu}}_{01}^{(h)}, \hat{\hat{\mu}}_{02}^{(h)}, \ldots, \hat{\hat{\mu}}_{0B}^{(h)}$ are the B bootstrap resamples of $\hat{\mu}_0^{(h)}$. As before, the interval (29) can be computed efficiently using linear fractional programming. In particular, it is easy to verify that Proposition 5 still holds and thus we only need to consider the $O(n)$ candidate solutions in Equation (21).

For the ATT, notice that $\Delta_1 = \mathbb{E}[Y(1)|A=1] - \mathbb{E}[Y(0)|A=1]$. The first term is identifiable using the data and the second term is the non-respondent mean by treating $Y(0)$ as the response. We can use the same procedure in Section 5 to obtain confidence interval of Δ_1 under $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$, using the analogously defined SIPW estimates $\hat{\Delta}_1^{(h_0)}$.

6.2. **Augmented Inverse Probability Weighting.** Besides the basic IPW and SIPW estimators, another commonly used estimator in missing data and observational studies is the *augmented inverse probability weighting* (AIPW) estimator which has a double robustness property explained below (Robins et al., 1994). As usual, we consider the missing data problem first for simplicity. Apart from the missingness probability, the AIPW estimator also utilizes another nuisance parameter, $f_0(\boldsymbol{x}) = \mathbb{E}_0[Y|A=1,\boldsymbol{X}=\boldsymbol{x}]$. Suppose this is estimated by $\hat{f}(\boldsymbol{x})$ from the sample, for example, by linear regression, and $e_0(\boldsymbol{x})$ is estimated by $\hat{e}(\boldsymbol{x})$. Then the AIPW estimator (with weight stabilization) is given by

$$\hat{\mu}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^n \frac{A_iY_i}{\hat{e}(\boldsymbol{X}_i)} - \frac{A_i-\hat{e}(\boldsymbol{X}_i)}{\hat{e}(\boldsymbol{X}_i)}\hat{f}(\boldsymbol{X}_i).$$

Let $\bar{e}(\boldsymbol{x})$ and $\bar{f}(\boldsymbol{x})$ be the large-sample limits of $\hat{e}(\boldsymbol{x})$ and $\hat{f}(\boldsymbol{x})$. When $e_0(\boldsymbol{x})$ and $f_0(\boldsymbol{x})$ are estimated non-parametrically, then $\bar{e}(\boldsymbol{x}) = e_0(\boldsymbol{x})$ and $\bar{f}(\boldsymbol{x}) = f_0(\boldsymbol{x})$. When $e_0(\boldsymbol{x})$ and $f_0(\boldsymbol{x})$ are estimated parametrically, $\bar{e}(\boldsymbol{x}) = e_{\boldsymbol{\beta}_0}(\boldsymbol{x})$ and $\bar{f}(\boldsymbol{x}) = f_{\boldsymbol{\theta}_0}(\boldsymbol{X})$, the best parametric approximations (see (6)). Then it is easy to show that $\hat{\mu}_{\text{AIPW}}$ always estimates (regardless of the correctness of MAR)

$$\hat{\mu}_{\text{AIPW}} \xrightarrow{P} \mathbb{E}_0[Y] + \mathbb{E}_0\left[\frac{A-\bar{e}(\boldsymbol{X})}{\bar{e}(\boldsymbol{X})}(Y-\bar{f}(\boldsymbol{X}))\right]. \tag{30}$$

Under MAR (Assumption 1), by first taking expectation conditioning on $\boldsymbol{X}$ it is straightforward to show that the second term is 0 if $\bar{e}(\boldsymbol{x}) = e_0(\boldsymbol{X})$ or $\bar{f}(\boldsymbol{x}) = f_0(\boldsymbol{x})$. In other words,

$\hat{\mu}_{\text{AIPW}}$ is consistent for μ if MAR holds and at least one of $\hat{e}(\boldsymbol{x})$ and $\hat{f}(\boldsymbol{x})$ is consistent, a property called *double robustness*.

When MAR does not hold, by taking expectation conditioning on $\boldsymbol{X}$ and Y, (30) implies that

$$\hat{\mu}_{\text{AIPW}} - \mu \xrightarrow{P} \mathbb{E}_0\left[\frac{e_0(\boldsymbol{X}, Y) - \bar{e}(\boldsymbol{X})}{\bar{e}(\boldsymbol{X})}(Y - \bar{f}(\boldsymbol{X}))\right].$$

Now, as in Section 4, we consider the collection of parametric sensitivity models $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$ so $\bar{e}(\boldsymbol{x}) = e_{\boldsymbol{\beta}_0}(\boldsymbol{x})$. For $h \in \mathcal{H}(\lambda)$, define

$$\tilde{\mu}_{\bar{f}}^{(h)} = \mu + \mathbb{E}_0\left[\frac{e_0(\boldsymbol{X}, Y) - e^{(h)}(\boldsymbol{X}, Y)}{e^{(h)}(\boldsymbol{X}, Y)}(Y - \bar{f}(\boldsymbol{X}))\right],$$

where $e^{(h)}(\boldsymbol{x}, y) = \left[1 + e^{h(\boldsymbol{x},y) - \text{logit}(\mathbb{P}_0(A=1|\boldsymbol{X}=x))}\right]^{-1}$. As before, it is obvious that $\tilde{\mu}_{\bar{f}}^{(h_{\boldsymbol{\beta}_0})} = \mu$, the true mean response, where $h_{\boldsymbol{\beta}_0}(\boldsymbol{x}, y) = \text{logit}(e_{\boldsymbol{\beta}_0}(\boldsymbol{x})) - \text{logit}(e_0(\boldsymbol{x}, y))$.

The AIPW estimator of $\tilde{\mu}_{\bar{f}}^{(h)}$ is

$$\begin{aligned}\hat{\mu}_{\text{AIPW}}^{(h)} &= \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{e}^{(h)}(\boldsymbol{X}_i)} - \frac{A_i - \hat{e}^{(h)}(\boldsymbol{X}_i)}{\hat{e}^{(h)}(\boldsymbol{X}_i)}\hat{f}(\boldsymbol{X}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{A_i(Y_i - \hat{f}(\boldsymbol{X}_i))}{\hat{e}^{(h)}(\boldsymbol{X}_i)},\end{aligned} \tag{31}$$

where $\hat{e}^{(h)}(\boldsymbol{x}, y) = \left[1 + e^{h(\boldsymbol{x},y) - \hat{g}(\boldsymbol{x},y)}\right]^{-1}$ and where $\hat{g}$ is as in (12). The second term on the right hand side of (31) is not sample bounded, so it is often preferable to use the stabilized weights. This results in the following stabilized AIPW (SAIPW) estimator:

$$\begin{aligned}\hat{\mu}_{\text{SAIPW}}^{(h)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\boldsymbol{X}_i) + \frac{1}{\frac{1}{n}\sum_{i=1}^{n}A_i\hat{e}^{(h)}(\boldsymbol{X}_i)}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{A_i(Y_i - \hat{f}(\boldsymbol{X}_i))}{\hat{e}^{(h)}(\boldsymbol{X}_i)}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\boldsymbol{X}_i) + \frac{\sum_{i=1}^{n}A_i(Y_i - \hat{f}(\boldsymbol{X}_i))(1 + e^{h(\boldsymbol{X}_i,Y_i) - \hat{g}(\boldsymbol{X}_i)})}{\sum_{i=1}^{n}A_i(1 + e^{h(\boldsymbol{X}_i,Y_i) - \hat{g}(\boldsymbol{X}_i)})}.\end{aligned} \tag{32}$$

As before, this estimates μ when $h = h_{\boldsymbol{\beta}_0}$.

Compared to the SIPW estimator (13), in (32) we replace the response Y_i by $Y_i - \hat{f}(\boldsymbol{X}_i)$ and add an offset term $\frac{1}{n}\sum_{i=1}^{n}\hat{f}(\boldsymbol{X}_i)$. Therefore, computing the extrema of (32) can still be formulated as linear fractional programming and the numerical computation is efficient. To construct asymptotically valid confidence intervals, the outcome regression model $\hat{f}(\boldsymbol{X}_i)$ must be parametric (for example, linear regression). The Z-estimation framework in Appendix B can then be extended to show that Theorem 3 (validity of percentile bootstrap) still holds for SAIPW.

Similar to Sections 5 and 6, the above procedures can be extended to observational studies for estimating the ATE Δ, the non-respondent mean μ_0, or the ATT Δ_1, using analogously defined SAIPW estimates $\hat{\Delta}_{\text{SAIPW}}$, $\hat{\mu}_{0,\text{SAIPW}}$, and $\hat{\Delta}_{1,\text{SAIPW}}$, respectively.

6.3. **Lipschitz Constraints in the Sensitivity Model.** So far we have focused on the marginal sensitivity models,

$$\mathcal{E}(\Lambda) \text{ or } \mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda) = \{e^{(h)}(\boldsymbol{x}, y) : h \in \mathcal{H}(\lambda)\}.$$

Although this model is very easy to interpret, some deviations $h \in \mathcal{H}(\lambda)$ may be deemed unlikely because the function h is not smooth. Here, we consider an extension of our sensitivity model which assumes h is also Lipschitz-continuous. Formally, define

$$\mathcal{E}(\Lambda) \text{ or } \mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda) := \{e^{(h)}(\boldsymbol{x}, y) : h \in \mathcal{H}_{\lambda,L}\},$$

where

$$\mathcal{H}_{\lambda,L} = \mathcal{H}(\lambda) \bigcap \left\{ h : \frac{|h(\boldsymbol{x}_1, y_1) - h(\boldsymbol{x}_2, y_2)|}{d((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2))} \leq L, \text{ for all } \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathscr{X}, \text{ and } y_1, y_2 \in \mathbb{R} \right\},$$

and d is a distance metric defined on the space $\mathscr{X} \times \mathbb{R}$ and $L > 0$ is the Lipschitz constant.

As $\mathcal{H}_{\lambda,L} \subseteq \mathcal{H}(\lambda)$, the validity of the percentile bootstrap obviously holds for functions in $\mathcal{H}_{\lambda,L}$. To obtain a range of point estimates and confidence intervals under $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda, L)$, we just need to add the following $n(n-1)$ constraints in the optimization problem (18):

$$z_i \leq e^{L \cdot d((\boldsymbol{X}_i, Y_i), (\boldsymbol{X}_j, Y_j))} z_j, \text{ for all } 1 \leq i \neq j \leq n,$$

These constraints are linear in $\{z_i\}_{i=1}^m$ and the resulting optimization problem is still a linear fractional program, which can be efficiently computed. However, Proposition 5 no longer holds, so we cannot use the algorithm described after Proposition 5 to solve the optimization problem corresponding to $\mathcal{H}_{\lambda,L}$.

7. Discussion

7.1. **Related Frameworks of Sensitivity Analysis.** If there is no unmeasured confounder, the potential outcome $Y(a)$ is independent of the treatment A given $\boldsymbol{X}$, $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}$, for $a \in \{0, 1\}$. Existing sensitivity analysis methods have considered at least three types of relaxations of this assumption:

(1) Pattern-mixture models consider a *specific* difference between the conditional distribution $Y(a)|\boldsymbol{X}, A$ and $Y(a)|\boldsymbol{X}$ (e.g. Robins, 1999, 2002, Birmingham et al., 2003, Vansteelandt et al., 2006, Daniels and Hogan, 2008).
(2) Selection models consider a *specific* difference between the conditional distribution $A|\boldsymbol{X}, Y(a)$ and $A|\boldsymbol{X}$ (e.g. Scharfstein et al., 1999, Gilbert et al., 2003, 2013).
(3) Rosenbaum's sensitivity models consider a *range* of possible selection models so that within a matched set the probabilies of getting treated are no more different than a number Λ in odds ratio. A worst case p-value of Fisher's sharp null hypothesis is then reported (e.g. Rosenbaum, 2002c, Chapter 4).

The first two approaches have the desirable property that, under a specified deviation, one can often use existing theory to derive an asymptotically normal (and sometimes efficient) estimator of the causal effect. However, they are arguably more difficult to interpret than Rosenbaum's sensitivity model because it is impossible to exhaust all possible deviations in this way. Often, one considers just a few functional forms of the deviation and hopes the results of the sensitivity analysis can be extended to "similar" functional forms (see e.g. Brumback et al., 2004).

Our marginal sensitivity model can be regarded as a hybrid of the selection model and Rosenbaum's approach, in that we consider a range of possible differences between $A|\boldsymbol{X}, Y(a)$ and $A|\boldsymbol{X}$. This model was considered first introduced by Tan (2006), who noticed that the range of IPW point estimates can be computed by linear programming. However, Tan (2006) did not consider sampling variation of the bounds, thus his method has limited applicability in practice.

7.2. **Comparison with Rosenbaum's sensitivity analysis.** Rosenbaum (1987) proposed to quantify the degree of violation of the MAR/NUC assumption based on the largest odds ratio of $e_0(\boldsymbol{x}, y_1)$ and $e_0(\boldsymbol{x}, y_2)$:

Definition 4 (Rosenbaum's Sensitivity Models)**.** Fix a parameter a $\Gamma \geq 1$ which will quantify the degree of violation from the MAR assumption.

(1) For the missing data problem, assume $e(\cdot, \cdot) \in \mathcal{R}(\Gamma)$, where

$$\mathcal{R}(\Gamma) = \Big\{ e(\cdot, \cdot) : \frac{1}{\Gamma} \leq \mathrm{OR}(e(\boldsymbol{x}, y_1), e(\boldsymbol{x}, y_2)) \leq \Gamma, \text{ for all } \boldsymbol{x} \in \mathscr{X}, y_1, y_2 \in \mathbb{R} \Big\}. \tag{33}$$

(2) For the observational studies problem, assume $e_a(\cdot, \cdot) \in \mathcal{R}(\Gamma)$, for $a = 0, 1$.

The proof of Proposition 1 indicates that not all $e_0(\boldsymbol{X}, Y)$ are compatible with the observed data. To compare the marginal sensitivity model $\mathcal{E}$ with Rosenbaum's model $\mathcal{R}$, we introduce the following concept:

Definition 5 (Compatibility of Sensitivity Model)**.** A sensitivity model $e_0(\boldsymbol{x}, y) = \mathbb{P}_0(A = 1 | \boldsymbol{X} = \boldsymbol{x}, Y = y)$ is called *compatible* if the RHS of (34) integrates to 1, for all $\boldsymbol{x} \in \mathscr{X}$, that is, $e_0(\cdot, \cdot) \in \mathcal{C}$, where

$$\mathcal{C} = \Big\{ e(\cdot, \cdot) : \int \mathrm{OR}(e_0(\boldsymbol{x}), e(\boldsymbol{x}, y)) \, \mathrm{d}\mathbb{P}_0(y | A = 1, \boldsymbol{X} = \boldsymbol{x}) = 1, \text{ for all } \boldsymbol{x} \in \mathscr{X} \Big\}.$$

Then Rosenbaum's sensitivity model and the marginal sensitivity model are related in the following way:

Proposition 7. *For any $\Lambda \geq 1$, $\mathcal{E}(\sqrt{\Lambda}) \subseteq \mathcal{R}(\Lambda)$ and $\mathcal{R}(\Lambda) \cap \mathcal{C} \subseteq \mathcal{E}(\Lambda) \cap \mathcal{C}$.*

As mentioned in the Introduction, Rosenbaum and his coauthors obtained point estimate and confidence interval of the causal effect under the collection of sensitivity models $\mathcal{R}(\Gamma)$. To this end, it is often assumed that the causal effect is additive and constant across the individuals, that is, $Y_i(1) - Y_i(0) \equiv \Delta$, for all $i \in [n]$. Then to determine if an effect Δ should be included in the $(1 - \alpha)$-confidence interval, one just needs to test the Fisher null $H_0 : Y_i(0) = Y_i(1)$ for all $i \in [n]$, using $Y - \Delta A$ as the outcome (Hodges and Lehmann, 1963) and under Rosenbaum's sensitivity model. We refer the reader to Rosenbaum (2002c, Chapter 4) for an overview of this approach. Our approach is different from existing methods targeting Rosenbaum's sensitivity model in many ways, sometimes markedly:

- *Population*: Most if not all existing methods for Rosenbaum's model treat the observed samples as the population, whereas we treat the observations as i.i.d. samples from a much larger super-population.
- *Design*: Existing methods usually require the data are paired or grouped. Statistical theory assumes the matching is *exact*, which is usually not strictly enforced in practice. Our approach is based on the IPW estimator and does not require exact matching.
- *Sensitivity Model*: We consider a different but closely related sensitivity model. Rosenbaum's sensitivity model is most natural for matched designs, whereas the marginal model is most natural when using IPW estimators. We also consider a parametric extension of the marginal sensitivity model.
- *Statistical Inference*: Most existing methods are based on randomization tests of Fisher's sharp null hypothesis, utilizing the randomness in treatment assignment. Our approach takes a point estimation perspective by trying to estimate the average treatment effect directly. The distinction can be best understood by comparing to the distinction between hypothesis testing and point estimation, or in Ding (2017)'s

terminology, the subtle difference between Neyman's null (the *average* causal effect is zero) and Fisher's null (the individual causal effects are *all* zero).

- *Effect Heterogeneity*: Constructing confidence intervals under Rosenbaum's sensitivity model usually require the causal effect is homogeneous, apart from Rosenbaum (2002a) who considered the "attributable effect" of a treatment. Some very recent advancements aim to remove this requirement in randomization inference. Fogarty et al. (2017) and Fogarty (2017) considered estimating the sample ATE in observational studies with a matched pairs design. Our approach inherently allows the causal effect to be heterogeneous.
- *Applicability to Missing Data Problems*: Our approach can be easily applied to missing data problems.

7.3. **Partially Identified Parameter.** Our framework is also related to a literature in econometrics on partially identified parameters (Imbens and Manski, 2004, Vansteelandt et al., 2006, Chernozhukov et al., 2007, Aronow and Lee, 2012, Miratrix et al., 2017). See Richardson et al. (2014) for a recent review. The mean response μ or the ATE Δ can be regarded as partially identified under the marginal sensitivity model. In fact, we have adopted the terminology "partially identified region" for the set $\{\mu^{(h)} : e^{(h)} \in \mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)\}$.

The main distinction is that existing methods in this literature usually require estimates of the boundaries of the partially identified region with known asymptotic distributions. In Section 4.3.1 we have shown that this is inherently difficult for sensitivity analysis. Our work opens the door for inference of partially identified parameters when it is difficult to analyze the asymptotic behavior of the boundary estimates.

8. Numerical Example

We illustrate the methods proposed in Sections 4 and 6 by an observational study, in which we are interested in estimating the causal effect of fish consumption on the blood mercury level. We obtained 2512 survey responses from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 who were at least 18 years old, answered the questionnaire about seafood consumption, and had blood mercury measured. Among these individuals, 1 has missing education, 7 have missing smoking, and 175 have missing income. We removed the individuals with missing education or smoking and imputed the missing income using the median income (we also added a binary indicator for missing income). Then we defined high fish consumption as more than 12 servings of fish or shellfish in the previous month, and low fish consumption as 0 or 1 servings of fish. In the end we were left with 234 treated individuals (high consumption), 873 controls (low consumption), and 8 covariates: gender, age, income, whether income is missing, race, education, ever smoked, and number of cigarettes smoked last month. The outcome variable is $\log_2$ of total blood mercury (in ug/L). This dataset was also analyzed by Zhao et al. (2017) and is publicly available in the `R` package `CrossScreening` on CRAN.

We used the percentile bootstrap to conduct sensitivity analyses for four estimators: $\hat{\Delta}$, $\hat{\Delta}_{\text{SAIPW}}$, $\hat{\Delta}_1$, and $\hat{\Delta}_{1,\text{SAIPW}}$. The propensity score $\mathbb{P}(A = 1|\boldsymbol{X})$ is estimated by a logistic regression and the outcome means $\mathbb{E}[Y(1)|\boldsymbol{X}]$ and $\mathbb{E}[Y(0)|\boldsymbol{X}]$ are estimated by linear regressions using all 8 covariates. We used $B = 1000$ bootstrap samples to obtain 90% confidence intervals of Δ and Δ_1 under $\mathcal{E}_{\boldsymbol{\beta}_0}(\Lambda)$ for 5 values of $\lambda = \log \Lambda = 0, 0.5, 1, 2, 3$.

We compared the results with Rosenbaum's sensitivity analysis as implemented in the `senmwCI` function (default options) in the `R` package `sensitivitymw` (Rosenbaum, 2015). We used the 234 matched pairs created by Zhao et al. (2017) as the basis of the sensitivity analysis.

TABLE 1. Results of different sensitivity analyses under the collection of parametric sensitivity models $\mathcal{E}_{\beta_0}(\Lambda)$. Five methods were considered: the SIPW and SAIPW estimators of ATE and ATT as described in Sections 4 and 6, and Rosenbaum's sensitivity analysis based on matched pairs. Eight sensitivity parameters were used: $\lambda = \log \Lambda$ or $\gamma = \log \Gamma = 0, 0.5, 1, 2, 3$. The running time of these five methods were also reported. For the first four methods, we used $B = 1000$ bootstrap samples and used 3 cores in parallel to obtain the confidence intervals. For matching, we do not include the computational cost of generating the matches.

Estimand	Method	Point	90% CI	Point	90% CI	Point	90% CI
		$\Lambda = e^0 = 1$		$\Lambda = e^{0.5} = 1.65$		$\Lambda = e^1 = 2.72$	
ATE	SIPW	(1.86,1.86)	(1.63,2.06)	(1.33,2.37)	(1.11,2.55)	(0.83,2.84)	(0.61,2.99)
	SAIPW	(1.80,1.80)	(1.55,2.04)	(1.34,2.28)	(1.16,2.53)	(0.89,2.76)	(0.73,2.99)
ATT	SIPW	(2.09,2.09)	(1.91,2.29)	(1.59,2.55)	(1.38,2.72)	(1.04,2.95)	(0.80,3.12)
	SAIPW	(2.12,2.12)	(1.93,2.31)	(1.64,2.56)	(1.45,2.74)	(1.15,2.95)	(0.95,3.16)
		$\Gamma = e^0 = 1$		$\Gamma = e^{0.5} = 1.65$		$\Gamma = e^1 = 2.72$	
Constant	Matching	(2.08,2.08)	(1.90,2.25)	(1.75,2.41)	(1.57,2.59)	(1.45,2.74)	(1.25,2.94)

Estimand	Method	Point	90% CI	Point	90% CI	Running time (seconds)
		$\Lambda = e^2 = 7.39$		$\Lambda = e^3 = 20.09$		
ATE	SIPW	(-0.10,3.78)	(-0.30,4.01)	(-0.91,4.78)	(-1.15,4.99)	32.8
	SAIPW	(0.12,3.61)	(-0.02,3.83)	(-0.55,4.31)	(-0.76,4.55)	47.7
ATT	SIPW	(-0.05,3.43)	(-0.43,3.58)	(-1.07,3.53)	(-1.36,3.68)	18.0
	SAIPW	(0.10,3.67)	(-0.20,3.92)	(-0.91,4.15)	(-1.36,4.47)	29.9
		$\Gamma = e^0 = 1$		$\Gamma = e^{0.5} = 1.65$		
Constant	Matching	(0.87,3.36)	(0.58,3.65)	(0.28,3.97)	(-0.23,4.48)	1.8

As mentioned previously, Rosenbaum's sensitivity analysis assumes constant treatment effect (CTE) to construct confidence intervals.

The results of the five sensitivity analyses are reported in Table 1. Alternatively, one can report the results by plotting the confidence intervals against Λ (Figure 1). Overall, the confidence intervals constructed by the percentile bootstrap were slightly wider than those constructed by Rosenbaum's sensitivity analysis under the same Λ, while the percentile bootstrap intervals under $\sqrt{\Lambda}$ were shorter than Rosenbaum's under Λ. This observation is not surprising given Proposition 7. Augmentation by outcome regressions (SAIPW estimators) helped to reduce the width of confidence intervals when the estimand is ATE, but did not reduce the width when the estimand is ATT. The IPW analyses suggested that the ATE/ATT is significantly positive for at least $\Lambda = 2.72$, while the matching analysis found the effect is significantly positive for at least $\Gamma = 7.39$.

Lastly we want to comment on the computational costs reported in the last column of Table 1. The IPW analyses are slower than the matching analyses (assuming the matches are given), but the running time is quite acceptable given we have over a thousand observations. The reason for the apparent advantage of matching is that we do not include the time for generating the matches. The matching analysis uses analytic approximations to conduct sensitivity analysis, hence it is faster than the IPW analyses which use the bootstrap. Since the time complexity of the IPW analyses scales almost linearly with the sample size, we expect the running times for larger studies will still be acceptable.

Acknowledgement: The authors thank Colin Fogarty for pointing out the difference between Rosenbaum's sensitivity model and the marginal sensitivity model.

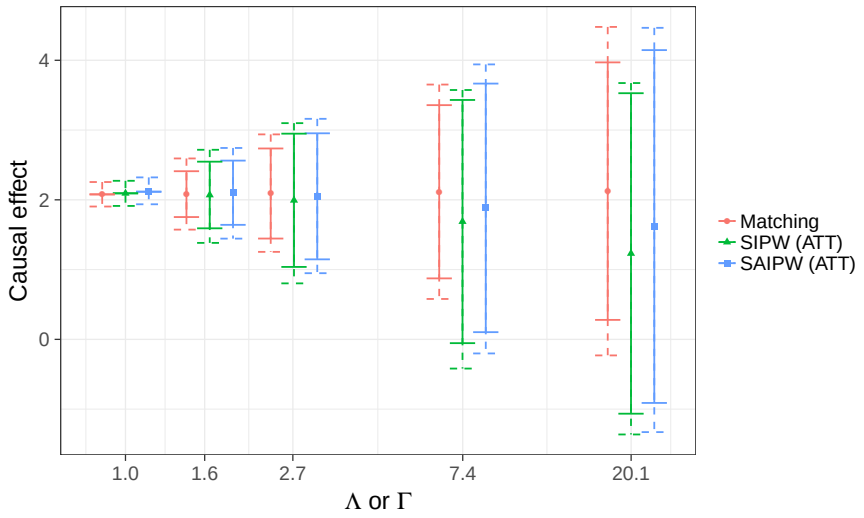

FIGURE 1. Graphical illustration of the sensitivity analysis results for three methods (matching, SIPW for ATT, and SAIPW for ATT). The solid error bars are the range of point estimates and the dashed error bars (together with the solid bars) are the confidence intervals. The circles/triangles/squares are the mid-points of the solid bars.

REFERENCES

Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.

Joseph G Altonji, Todd E Elder, and Christopher R Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1): 151–184, 2005.

Peter M Aronow and Donald KK Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240, 2012.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

Peter C Austin. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30):5642–5655, 2016.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Jolene Birmingham, Andrea Rotnitzky, and Garrett M Fitzmaurice. Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):275–297, 2003.

Babette A Brumback, Miguel A Hernán, Sebastien JPA Haneuse, and James M Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine*, 23(5):749–767, 2004.

Abraham Charnes and William W Cooper. Programming with linear fractional functionals. *Naval Research Logistics*, 9(3-4):181–186, 1962.

Victor Chernozhukov, Han Hong, and Elie Tamer. Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284, 2007.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, to appear, 2017.

Joel E Cohen. Generalized minimax and maximin inequalities for order statistics and quantile functions. *Proceedings of the American Mathematical Society*, 141(7), 2013.

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203, 1959.

Michael J Daniels and Joseph W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, 2008.

George B Dantzig. Maximization of a linear function of variables subject to linear inequalities. In T Koopmans, editor, *Activity Analysis of Production and Allocation*, pages 339–347. Wiley, New York, 1951.

Peng Ding. A paradox from randomization-based causal inference. *Statistical Science*, 32(3):331–345, 2017.

Peng Ding and Tyler J VanderWeele. Sensitivity analysis without assumptions. *Epidemiology*, 27(3): 368, 2016.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Ronald A Fisher. Cigarettes, cancer and statistics. *Centennial Review of Arts & Science*, 2:151–166, 1958.

Colin B Fogarty. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *arXiv preprint arXiv:1609.02112*, 2017.

Colin B Fogarty, Pixu Shi, Mark E Mikkelsen, and Dylan S Small. Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies. *Journal of the American Statistical Association*, 112(517):321–331, 2017.

Joseph L Gastwirth, Abba M Krieger, and Paul R Rosenbaum. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):907–920, 1998.

Peter B Gilbert, Ronald J Bosch, and Michael G Hudgens. Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541, 2003.

Peter B Gilbert, Bryan E Shepherd, and Michael G Hudgens. Sensitivity analysis of per-protocol time-to-event treatment efficacy in randomized clinical trials. *Journal of the American Statistical Association*, 108(503):789–800, 2013.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Joseph L Jr Hodges and Erich L Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, pages 598–611, 1963.

Paul W Holland. Statistics and causal inference. *Journal of American Statistical Association*, 81: 945–960, 1986.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Michael G Hudgens and M Elizabeth Halloran. Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association*, 101(473):51–64, 2006.

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539, 2007.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2006.

Roderick J Little, Ralph D'agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367

(14):1355–1360, 2012.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

Lawrence C McCandless, Paul Gustafson, and Adrian Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26(11):2331–2347, 2007.

Luke W Miratrix, Stefan Wager, and Jose R Zubizarreta. Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika*, to appear, 2017.

Amy Richardson, Michael G Hudgens, Peter B Gilbert, and Jason P Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical Science*, 29(4):596, 2014.

James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.

James M Robins. Association, causation, and marginal structural models. *Synthese*, 121(1):151–179, 1999.

James M Robins. Comment on "covariance adjustment in randomized experiments and observational studies". *Statistical Science*, 17(3):309–321, 2002.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

Paul R Rosenbaum. Attributing effects to treatment in matched observational studies. *Journal of the American statistical Association*, 97(457):183–192, 2002a.

Paul R Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002b.

Paul R Rosenbaum. *Observational Studies*. Springer New York, 2002c.

Paul R Rosenbaum. Two R packages for sensitivity analysis in observational studies. *Observational Studies*, 1(1):1–17, 2015.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B Rubin. Comment on "Randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94 (448):1096–1120, 1999.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine*, 330:1029–1035, 1994.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer, 2011.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Tyler J VanderWeele and Onyebuchi A Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 22(1):42–52, 2011.

Stijn Vansteelandt, Els Goetghebeur, Michael G Kenward, and Geert Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3):953–979, 2006.

J. A. Wellner and Yihui Zhan. Bootstrapping z–estimators. *manuscript*, pages 1–37, 1996.

Jon A. Wellner. *Empirical Processes: Theory and Applications*. Delft Technical University Lecture Notes, 2005.

Walter C Willett. Vitamin A and lung cancer. *Nutrition Reviews*, 48(5):201–211, 1990.

Qingyuan Zhao, Dylan S Small, and Paul R Rosenbaum. Cross-screening in observational studies that test many hypotheses. *Journal of American Statistical Association*, 2017.

Appendix A. Proofs

A.1. **Proof of Proposition 1.** The complete data density can be factorized as

$$\mathbb{P}_0(A, \boldsymbol{X}, Y) = \mathbb{P}_0(A, \boldsymbol{X}) \cdot \mathbb{P}_0(Y|A, \boldsymbol{X}).$$

The first term in the RHS above is identifiable from the data, because A and $\boldsymbol{X}$ are always observed. For the second term, note that $\mathbb{P}_0(Y|A=1, \boldsymbol{X})$ is identifiable because Y is observed if $A=1$. For $\mathbb{P}_0(Y|A=0, \boldsymbol{X})$, using Bayes rule, we get

$$\begin{aligned}
&\mathbb{P}_0(Y|A=0, \boldsymbol{X}) \\
=&\frac{\mathbb{P}_0(A=0|Y, \boldsymbol{X}) \cdot \mathbb{P}_0(Y|\boldsymbol{X})}{\mathbb{P}_0(A=0|\boldsymbol{X})} \\
=&\frac{\mathbb{P}_0(A=0|Y, \boldsymbol{X}) \cdot \{\mathbb{P}_0(Y|A=0, \boldsymbol{X})\mathbb{P}_0(A=0|\boldsymbol{X}) + \mathbb{P}_0(Y|A=1, \boldsymbol{X})\mathbb{P}_0(A=1|\boldsymbol{X})\}}{\mathbb{P}_0(A=0|\boldsymbol{X})}.
\end{aligned}$$

Notice that if the denominator $\mathbb{P}_0(A=0|\boldsymbol{X}=\boldsymbol{x})=0$, then we don't need to consider the conditional distribution of Y. By simple algebra, we have

$$\mathbb{P}_0(Y|A=0, \boldsymbol{X}) = \mathrm{OR}(e_0(\boldsymbol{X}), e_0(\boldsymbol{X}, Y)) \cdot \mathbb{P}_0(Y|A=1, \boldsymbol{X}), \tag{34}$$

where $\mathrm{OR}(p_1, p_2) := [p_1/(1-p_1)]/[p_2/(1-p_2)]$ is the *odds ratio* of $p_1, p_2 \in (0,1)$. The only terms in the above equation that are not directly identifiable from the data are $\mathbb{P}_0(Y|A=0, \boldsymbol{X})$ and $e_0(\boldsymbol{X}, Y)$. We can arbitrarily specify $e_0(\boldsymbol{x}, y)$ as long as the RHS of (34) integrates to 1. This gives us $\mathbb{P}_0(Y|A=0, \boldsymbol{X})$ and hence the complete data distribution. It is easy to see that different choices of $e_0(\boldsymbol{x}, y)$ can generate the same density for the observed data $(A, \boldsymbol{X}, AY)$ but different densities for the complete data $(A, \boldsymbol{X}, Y)$.

A.2. **Proof of Proposition 2.** By definition, under the sensitivity model $\mathcal{H}(\gamma)$, the true data generating distribution F_0 satisfies $h_0(\boldsymbol{x}, y) \in \mathcal{H}_\gamma$. This implies that

$$\mathbb{P}_0(\mu \in [L, U]) = \mathbb{P}_0\big(\mu^{(h_0)} \in [L, U]\big) \geq \mathbb{P}_0\big(\mu^{(h_0)} \in [L_{h_0}, U_{h_0}]\big).$$

The last inequality is true because $[L_{h_0}, U_{h_0}]$ is a $(1-\alpha)$-confidence interval for $\mu^{(h_0)}$. Now, taking limit on both sides gives

$$\liminf_{n\to\infty} \mathbb{P}_0(\mu \in [L, U]) \geq \liminf_{n\to\infty} \mathbb{P}_0\big(\mu^{(h_0)} \in [L_{h_0}, U_{h_0}]\big) \geq 1-\alpha.$$

For the second part of the proposition, let $h_{\min} = \arg\min_{h\in\mathcal{H}(\gamma)} \mu^{(h)}$. By applying assumption (14) to the intervals $[L^{(h)}, \infty)$, for $h \in \mathcal{H}(\gamma)$, gives $\limsup_{n\to\infty} \mathbb{P}_0\big(\mu^{(h_{\min})} < L\big) \leq \alpha'$.

Similarly, defining $h_{\max} = \arg\max_{h\in\mathcal{H}(\gamma)} \mu^{(h)}$, we have $\limsup_{n\to\infty} \mathbb{P}_0\big(\mu^{(h_{\max})} > U\big) \le \alpha - \alpha'$. Then using the union bound,

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}_0\left(\{\mu^{(h)} : e^{(h)} \in \mathcal{E}_{\boldsymbol{\beta}_0}(\Gamma)\} \not\subseteq [L,U]\right) \\
&\qquad \le \limsup_{n\to\infty} \mathbb{P}_0\left(\left[\mu^{(h_{\min})}, \mu^{(h_{\max})}\right] \not\subseteq [L,U]\right) \\
&\qquad \le \limsup_{n\to\infty} \mathbb{P}_0\left(\mu^{(h_{\min})} < L\right) + \limsup_{n\to\infty} \mathbb{P}_0\left(\mu^{(h_{\max})} > U\right) \\
&\qquad \le \alpha,
\end{aligned}$$

completing the proof.

A.3. **Proof of Lemma 1.** For $1 \le b \le N$, where $N = n^n$ is the total number of possible bootstrap resamples, denote by $\hat{\hat{\mu}}_b^{(h)}$ the estimate (13) in the b-th resample. Therefore, for every $h \in \mathcal{H}(\gamma)$,

$$\hat{\hat{\mu}}_b^{(h)} \ge \inf_{h\in\mathcal{H}(\gamma)} \hat{\hat{\mu}}_b^{(h)}, \quad \text{for all } 1 \le b \le N. \tag{35}$$

Note that both sides of (35) above are sequences indexed by $b \in [N]$, and since the inequality is true entry by entry, it is also true for any order statistic of the sequences, namely for any $0 < \alpha < 1$,

$$Q_{\frac{\alpha}{2}}\left(\left(\hat{\hat{\mu}}_b^{(h)}\right)_{b\in[N]}\right) \ge Q_{\frac{\alpha}{2}}\left(\left(\inf_{h\in\mathcal{H}(\gamma)} \hat{\hat{\mu}}_b^{(h)}\right)_{b\in[N]}\right) = L, \tag{36}$$

by definition (17). Since (36) is true for any $h \in \mathcal{H}(\gamma)$, taking infimum on the LHS above gives

$$\inf_{h\in\mathcal{H}(\gamma)} L^{(h)} = \inf_{h\in\mathcal{H}(\gamma)} Q_{\frac{\alpha}{2}}\left(\left(\hat{\hat{\mu}}_b^{(h)}\right)_{b\in[N]}\right) \ge L.$$

The lower bound on U can be proved similarly.

A.4. **Proof of Proposition 5.** We prove the first claim by contradiction. Suppose there exists two indices $s_1 < s_2$ such that $A_{s_1} = A_{s_2} = 1$, $Y_{s_1} > Y_{s_2}$ but $z_{s_1} < z_{s_2}$. Then consider the following perturbation,

$$z_i' = \begin{cases} z_i, & s \ne s_1, s_2, \\ z_{s_1} + \varepsilon e^{\hat{g}(\boldsymbol{X}_{s_1})}, & i = s_1, \\ z_{s_2} - \varepsilon e^{\hat{g}(\boldsymbol{X}_{s_2})}, & s = s_2. \end{cases}$$

When $\varepsilon > 0$ is sufficiently small, $(z_i')_{i=1}^m$ is still feasible but the objective becomes larger, which contradicts the assumption that $(z_i)_{i=1}^m$ is the maximizer.

Next, we prove the second claim. It is well known that if a linear programming is feasible and bounded, then there is at least one vertex (also called the basic feasible) solution (Dantzig, 1951). Notice that in (20), there are $m+1$ optimization variables and 1 equality constraint. It is also easy to verify that $t = 0$ is not feasible. Therefore, among the $2m$ inequality constraints, $\frac{1}{\Gamma} t \le \bar{z}_i \le \Gamma t$, $1 \le s \le m$, there exists a solution of (20) such that m equalities hold. This implies that $z_i = \frac{1}{t}\bar{z}_i$ is either Γ or $\frac{1}{\Gamma}$, for all $1 \le s \le m$. Then, using the first part of the proposition, there exists a M such that $z_i = \Gamma$, if $Y_i \ge M$, and $z_i = \frac{1}{\Gamma}$, if $Y_i < M$.

A.5. **Proof of Proposition 7.** To begin with suppose that $e \in \mathcal{E}(\sqrt{\Gamma})$. Then, for all $\boldsymbol{x} \in \mathscr{X}$, and $y_1, y_2 \in \mathbb{R}$,

$$\begin{aligned}|\log \mathrm{OR}(e(\boldsymbol{x}, y_1), e(\boldsymbol{x}, y_2))| &\leq |\log \mathrm{OR}(e(\boldsymbol{x}, y_1), e_0(\boldsymbol{x}))| + |\log \mathrm{OR}(e_0(x), e(\boldsymbol{x}, y_2))| \\ &\leq \log \Gamma,\end{aligned}$$

which implies $e \in \mathcal{R}(\Gamma)$.

Now, suppose the function $e \in \mathcal{R}(\Gamma) \cap \mathcal{C}$. By (33),

$$\frac{1}{\Gamma} \leq \mathrm{OR}\left(\inf_{y\in\mathbb{R}} e(\boldsymbol{x}, y), \sup_{y\in\mathbb{R}} e(\boldsymbol{x}, y)\right) \leq \Gamma, \text{ for all } \boldsymbol{x} \in \mathscr{X}.$$

Notice that $\inf_y e(\boldsymbol{x}, y) \leq e_0(\boldsymbol{x}) \leq \sup_y e(\boldsymbol{x}, y)$, because $e(\boldsymbol{x}, y)$ marginalizes to $e_0(\boldsymbol{x})$ and $e(\boldsymbol{x}, y)$ is compatible. Thus, $\mathrm{OR}(e(\boldsymbol{x}, y), e_0(\boldsymbol{x}))$ must be between $1/\Gamma$ and Γ, which implies $e \in \mathcal{E}(\Gamma) \cap \mathcal{C}$.

Appendix B. Proof of Theorem 3

To begin with, define $\kappa^{(h)} := \mathbb{E}_0\left[A\left(1 + e^{h(\boldsymbol{X},Y) - \mathrm{logit}(\mathbb{P}_{\boldsymbol{\beta}_0}(A=1|\boldsymbol{X}))}\right)\right]$. Recall from (9), that

$$\mu^{(h)} = \frac{1}{\kappa^{(h)}}\mathbb{E}_0\left[AY\left(1 + e^{h(\boldsymbol{X},Y) - \mathrm{logit}(\mathbb{P}_{\boldsymbol{\beta}_0}(A=1|\boldsymbol{X}))}\right)\right],$$

for $h \in \mathcal{H}(\gamma)$. We begin by showing that the estimates of the parameters $(\mu^{(h)}, \kappa^{(h)}, \boldsymbol{\beta}_0')'$ can be derived using the framework of Z-estimation.

B.1. **The Z-Estimation Framework.** Given a vector $\boldsymbol{\upsilon} = (\nu, \kappa, \boldsymbol{\beta}')' \in \Theta \subset \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}^d$, where Θ is the compact parameter space, define the function $Q : \{0,1\} \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d+2}$ as follows: For $\boldsymbol{t} = (a, \boldsymbol{x}', y)' \in \{0,1\} \times \mathbb{R} \times \mathbb{R}^d$,

$$Q(\boldsymbol{t}|\boldsymbol{\upsilon}) = \begin{pmatrix} Q_1(\boldsymbol{t}|\boldsymbol{\upsilon}) \\ Q_2(\boldsymbol{t}|\boldsymbol{\upsilon}) \\ Q_3(\boldsymbol{t}|\boldsymbol{\upsilon}) \end{pmatrix} := \begin{pmatrix} \left(a - \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1+e^{\boldsymbol{\beta}'\boldsymbol{x}}}\right)\boldsymbol{x} \\ \kappa - a\left(1 + e^{h(\boldsymbol{x},y) - \boldsymbol{\beta}'\boldsymbol{x}}\right) \\ \kappa\nu - ay\left(1 + e^{h(\boldsymbol{x},y) - \boldsymbol{\beta}'\boldsymbol{x}}\right) \end{pmatrix}. \tag{37}$$

Next, define $\Phi(\boldsymbol{\upsilon}) = \int Q(\boldsymbol{t}|\boldsymbol{\upsilon})\mathrm{d}\mathbb{P}_0(\boldsymbol{t})$, where $\boldsymbol{T} = (A, \boldsymbol{X}', AY)' \sim \mathbb{P}_0$, the true distribution generating the data. Note that $\Phi(\boldsymbol{\upsilon}_0) = 0$, where $\boldsymbol{\upsilon}_0 = (\mu^{(h)}, \kappa^{(h)}, \boldsymbol{\beta}_0')$ is the true parameter value. The Z-estimates $\hat{\boldsymbol{\upsilon}} = (\hat{\mu}^{(h)}, \hat{\kappa}^{(h)}, \hat{\boldsymbol{\beta}}')'$ are obtained by solving the equations

$$\Phi_n(\hat{\boldsymbol{\upsilon}}) := \frac{1}{n}\sum_{i=1}^n Q(\boldsymbol{T}_i|\hat{\boldsymbol{\upsilon}}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \left(A_i - \frac{e^{\hat{\boldsymbol{\beta}}'\boldsymbol{X}_i}}{1+e^{\hat{\boldsymbol{\beta}}'\boldsymbol{X}_i}}\right)\boldsymbol{X}_i \\ \hat{\kappa}^{(h)} - \frac{1}{n}\sum_{i=1}^n A_i\left(1 + e^{h(\boldsymbol{X}_i,Y_i) - \hat{\boldsymbol{\beta}}'\boldsymbol{X}_i}\right) \\ \hat{\kappa}^{(h)}\hat{\mu}^{(h)} - \frac{1}{n}\sum_{i=1}^n A_iY_i\left(1 + e^{h(\boldsymbol{X}_i,Y_i) - \hat{\boldsymbol{\beta}}'\boldsymbol{X}_i}\right) \end{pmatrix} = 0. \tag{38}$$

It is easy to see that the Z estimate $\hat{\mu}^{(h)}$ is exactly the SIPW estimate (13) for $\mu^{(h)}$ and $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ for the logistic regression model.

To formally define the bootstrap estimates, let $\mathbb{P}_n$ be the empirical measure of the sample $\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_n$, where $\boldsymbol{T}_i = (A_i, \boldsymbol{X}_i', A_iY_i)$, and $\hat{\boldsymbol{T}}_1, \hat{\boldsymbol{T}}_2, \ldots, \hat{\boldsymbol{T}}_n$ be i.i.d. samples from the empirical measure. The bootstrap empirical distribution and the bootstrap empirical process are

$$\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{\boldsymbol{T}}_i} \quad \text{and} \quad \hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n),$$

respectively. Noting that $\hat{\Phi}_n(\boldsymbol{\upsilon}) = \int Q(\boldsymbol{t}|\boldsymbol{\upsilon})\mathrm{d}\hat{\mathbb{P}}_n(\boldsymbol{t})$, the bootstrap Z-estimates $\hat{\hat{\boldsymbol{\upsilon}}}$ are obtained from the equations:

$$\hat{\Phi}_n(\hat{\hat{\boldsymbol{\upsilon}}}) := \frac{1}{n}\sum_{i=1}^{n} Q(\hat{\boldsymbol{T}}_i|\hat{\hat{\boldsymbol{\upsilon}}}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \left(\hat{A}_i - \frac{e^{\hat{\hat{\boldsymbol{\beta}}}'\hat{\boldsymbol{X}}_i}}{1+e^{\hat{\hat{\boldsymbol{\beta}}}'\hat{\boldsymbol{X}}_i}}\right)\hat{\boldsymbol{X}}_i \\ \hat{\hat{\kappa}} - \frac{1}{n}\sum_{i=1}^n \hat{A}_i\left(1+e^{h(\hat{\boldsymbol{X}}_i,\hat{Y}_i)-\hat{\hat{\boldsymbol{\beta}}}'\hat{\boldsymbol{X}}_i}\right) \\ \hat{\hat{\kappa}}\hat{\hat{\mu}}^{(h)} - \frac{1}{n}\sum_{i=1}^n \hat{A}_i\hat{Y}_i\left(1+e^{h(\hat{\boldsymbol{X}}_i,\hat{Y}_i)-\hat{\hat{\boldsymbol{\beta}}}'\hat{\boldsymbol{X}}_i}\right) \end{pmatrix} = 0. \tag{39}$$

Now, invoking the asymptotic theory of bootstrap for Z-estimators (Wellner and Zhan, 1996, Kosorok, 2006, Chapter 10), we can derive validity of the bootstrap confidence intervals discussed in Section 4.3.2. To this end, we need the following assumption.

Assumption 4. *The parameter space Θ is compact and the true parameter $\boldsymbol{\upsilon}_0$ is in the interior of Θ. Moreover, the joint distribution of $(Y, \boldsymbol{X})$ satisfies:*

(1) $\mathbb{E}[Y^4] < \infty$.

(2) $\det(\mathbb{E}\left(\frac{e^{\boldsymbol{\beta}_0'\boldsymbol{X}}}{(1+e^{\boldsymbol{\beta}_0'\boldsymbol{X}})^2}\boldsymbol{X}\boldsymbol{X}'\right)) > 0$.

(3) *For every compact subset $S \subset \mathbb{R}^d$, $\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in S} e^{\boldsymbol{\beta}'\boldsymbol{X}}\right] < \infty$.*

Note that all the assumptions are trivially satisfied if we assume that the supports of $\boldsymbol{X}$ and Y are bounded and $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}']$ is positive definite. The assumptions allow for more general distributions, for example, distributions with $\mathbb{E}[e^{t||\boldsymbol{X}||}] < \infty$, for all $t \in \mathbb{R}$. Under this assumption we can derive the asymptotic limiting distribution of the bootstrap estimates (recall (38) and (39)).

Theorem 8. *Suppose the joint distribution of $(Y, \boldsymbol{X})$ satisfies Assumption 4. Then, for $h \in \mathcal{H}(\gamma)$ be fixed, under $\mathbb{P}_0$,*

$$\sqrt{n}\,(\hat{\hat{\boldsymbol{\upsilon}}} - \boldsymbol{\upsilon}) \xrightarrow{D} N(\boldsymbol{0}, \dot{\Phi}_0^{-1}\Sigma\dot{\Phi}_0), \quad \textit{and} \quad \sqrt{n}\left(\hat{\hat{\boldsymbol{\upsilon}}} - \boldsymbol{\upsilon}\right) \xrightarrow{D} N(\boldsymbol{0}, \dot{\Phi}_0^{-1}\Sigma\dot{\Phi}_0),$$

where

$$\dot{\Phi}_0 = \mathbb{E}\nabla_{\boldsymbol{\upsilon}=\boldsymbol{\upsilon}_0} Q(\boldsymbol{T}|\boldsymbol{\upsilon}) = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} & -\mathbb{E}\left(\frac{e^{\boldsymbol{\beta}_0'\boldsymbol{X}}}{(1+e^{\boldsymbol{\beta}_0'\boldsymbol{X}})^2}\right)\boldsymbol{X}\boldsymbol{X}' \\ 0 & 1 & \mathbb{E}A\boldsymbol{X}'e^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}} \\ \kappa^{(h)} & \mu^{(h)} & \mathbb{E}AY\boldsymbol{X}'e^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}} \end{pmatrix}, \tag{40}$$

and $\Sigma := \mathbb{E}[Q(\boldsymbol{T}, \boldsymbol{\upsilon}_0)Q(\boldsymbol{T}, \boldsymbol{\upsilon}_0)']$.

The limiting distribution of $\hat{\mu}^{(h)}$ (recall (13) and (38)) and $\hat{\hat{\mu}}^{(h)}$ (defined through (39)) is an immediate consequence of the above theorem:

Corollary 9. *Suppose the joint distribution of $(Y, \boldsymbol{X})$ satisfies Assumption 4. Then, for $h \in \mathcal{H}(\gamma)$ be fixed, under $\mathbb{P}_0$,*

$$\sqrt{n}\left(\hat{\mu}^{(h)} - \mu^{(h)}\right) \xrightarrow{D} N(0, (\sigma^{(h)})^2), \quad \textit{and} \quad \sqrt{n}\left(\hat{\hat{\mu}}^{(h)} - \hat{\mu}^{(h)}\right) \xrightarrow{D} N(0, (\sigma^{(h)})^2),$$

where $(\sigma^{(h)})^2 = (\dot{\Phi}_0^{-1}\Sigma\dot{\Phi}_0)_{11}$ is the first diagonal element of $\dot{\Phi}_0^{-1}\Sigma\dot{\Phi}_0$.

The rest of the section is organized as follows: In Section B.2 we prove Theorem 8, and in Section B.3 we complete the proof of Theorem 3 using Corollary 9.

B.2. **Proof of Theorem 8.** We begin by verifying that the matrices $\dot{\Phi}_0^{-1}$ and Σ in Theorem 8 are well-defined. To begin with, a direct computation gives

$$|\det(\dot{\Phi}_0)| = |\det\begin{pmatrix} -\mathbb{E}\left(\frac{e^{\boldsymbol{\beta}_0'\boldsymbol{X}}}{(1+e^{\boldsymbol{\beta}_0'\boldsymbol{X}})^2}\right)\boldsymbol{X}\boldsymbol{X}' & \mathbf{0} & \mathbf{0} \\ \mathbb{E}A\boldsymbol{X}'e^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}} & 0 & 1 \\ \mathbb{E}AY\boldsymbol{X}'e^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}} & \kappa^{(h)} & \mu^{(h)} \end{pmatrix}| = \kappa^{(h)}|\det\mathbb{E}\left(\frac{e^{\boldsymbol{\beta}_0'\boldsymbol{X}}}{(1+e^{\boldsymbol{\beta}_0'\boldsymbol{X}})^2}\boldsymbol{X}\boldsymbol{X}'\right)| > 0,$$

by Assumption 4 (2). Therefore, $\dot{\Phi}_0$ is invertible. Also, note that $\Sigma < \infty$, which follows by direct multiplication and Assumption 4 (1).

We can now proceed to prove Theorem 8. This will be done by invoking (Kosorok, 2006, Theorem 10.16), which gives conditions are asymptotic normality of bootstrapped Z-estimators. This entails verifying the following three conditions:

(A) The class of functions $\{\boldsymbol{t} \to Q(\boldsymbol{t}|\boldsymbol{\upsilon}) : \boldsymbol{\upsilon} \in \Theta\}$ is $\mathbb{P}_0$-Glivenko-Cantelli (proved in Section B.2.1).
(B) $||\Phi(\boldsymbol{\upsilon})||_1$ is strictly positive outside every open neighborhood of $\boldsymbol{\upsilon}_0$ (proved in Section B.2.2).
(C) The class of functions $\{\boldsymbol{t} \to Q(\boldsymbol{t}|\boldsymbol{\upsilon}) : \boldsymbol{\upsilon} \in \Theta\}$ is $\mathbb{P}_0$-Donsker and $\mathbb{E}[(Q(\boldsymbol{T}|\boldsymbol{\upsilon}_n) - Q(\boldsymbol{T}|\boldsymbol{\upsilon}_0))^2] \to 0$, whenever $||\boldsymbol{\upsilon}_n - \boldsymbol{\upsilon}_0||_1 \to 0$ (proved in Section B.2.3).

B.2.1. ***Proof of*** *(A).* Define the envelope function $B(\boldsymbol{t}) := \sup_{\boldsymbol{\upsilon}\in\Theta}||Q(\boldsymbol{t}|\boldsymbol{\upsilon})||_1$. Then using the compactness of Θ, $||h||_\infty \leq \gamma$, and $|a| \leq 1$,

$$\begin{aligned} ||Q(\boldsymbol{t}|\boldsymbol{\upsilon})||_1 &\leq ||Q_1(\boldsymbol{t}|\boldsymbol{\upsilon})||_1 + |Q_2(\boldsymbol{t}|\boldsymbol{\upsilon})| + |Q_3(\boldsymbol{t}|\boldsymbol{\upsilon})| \\ &\leq ||\boldsymbol{x}||_1 + |y| + e^{\gamma}e^{-\boldsymbol{\beta}'\boldsymbol{x}}(1+|y|) + M, \end{aligned} \tag{41}$$

for some absolute constant M. Therefore, by Assumption 4, $\mathbb{E}B(\boldsymbol{T}) < \infty$, and by (Wellner, 2005, Lemma 6.1), the class of functions $\{\boldsymbol{t} \to Q(\boldsymbol{t}|\boldsymbol{\upsilon}) : \boldsymbol{\upsilon} \in \Theta\}$ is $\mathbb{P}_0$-Glivenko-Cantelli.

B.2.2. ***Proof of*** *(B).* To begin with note that by Assumption 4 (2),

$$\mathbb{E}\left(\frac{e^{\boldsymbol{\beta}'\boldsymbol{X}}}{(1+e^{\boldsymbol{\beta}'\boldsymbol{X}})}\boldsymbol{X}\right) = 0,$$

has a unique root $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, since its gradient is positive definite at $\boldsymbol{\beta}_0$, and non-negative definite everywhere. Then, fixing $\varepsilon > 0$, we have

$$||\Phi(\boldsymbol{\upsilon})||_1 \geq \left|\left|\mathbb{E}\left[\frac{e^{\boldsymbol{\beta}_0'\boldsymbol{X}}}{1+e^{\boldsymbol{\beta}_0'\boldsymbol{X}}}\boldsymbol{X} - \frac{e^{\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\boldsymbol{\beta}'\boldsymbol{X}}}\boldsymbol{X}\right]\right|\right|_1 > 0, \tag{42}$$

whenever $||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1 > \frac{\varepsilon}{M}$, where M is a constant to be chosen later.

Next, assume that $||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1 \leq \frac{\varepsilon}{M}$. This implies $||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_\infty \leq \frac{\varepsilon}{M}$ and

$$\begin{aligned} \left|\mathbb{E}\left[Ae^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}} - Ae^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}'\boldsymbol{X}}\right]\right| &\leq \mathbb{E}\left[\left|Ae^{h(\boldsymbol{X},Y)}\right| \cdot \left|e^{-\boldsymbol{\beta}_0'\boldsymbol{X}} - e^{-\boldsymbol{\beta}'\boldsymbol{X}}\right|\right] \\ &\leq e^{\gamma}\mathbb{E}\left|e^{-\boldsymbol{\beta}_0'\boldsymbol{X}} - e^{-\boldsymbol{\beta}'\boldsymbol{X}}\right| \\ &\leq e^{\gamma}\mathbb{E}\left|(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{X}e^{-t_*}\right| \quad (\text{for some } t_* \in [\boldsymbol{\beta}_0'\boldsymbol{X}, \boldsymbol{\beta}'\boldsymbol{X}]) \\ &\leq e^{\gamma}||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_\infty\mathbb{E}\left[||\boldsymbol{X}||_1 e^{-t_*}\right] \\ &\leq e^{\gamma}||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_\infty\left(\mathbb{E}[||\boldsymbol{X}||_1^2\mathbb{E}[e^{-2t_*}]\right)^{\frac{1}{2}} \end{aligned}$$

$$\leq ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_\infty \left(e^{2\gamma}\mathbb{E}[||\boldsymbol{X}||_1^2]\cdot\mathbb{E}\left[\sup_{\boldsymbol{\beta}_*:||\boldsymbol{\beta}_*-\boldsymbol{\beta}_0||_1\leq\frac{\varepsilon}{M}} e^{-2\boldsymbol{\beta}_*\boldsymbol{X}}\right]\right)^{\frac{1}{2}}$$
$$\leq K_1(\gamma)\cdot\frac{\varepsilon}{M}\leq\frac{\varepsilon}{64K}. \tag{43}$$

by choosing $M\geq 64K\cdot K_1(\gamma)$, where

$$K_1(\gamma)^2 := e^{2\gamma}\mathbb{E}[||\boldsymbol{X}||_1^2]\cdot\mathbb{E}\left[\sup_{\boldsymbol{\beta}_*:||\boldsymbol{\beta}_*-\boldsymbol{\beta}_0||_1\leq\frac{\varepsilon}{M}} e^{-2\boldsymbol{\beta}_*\boldsymbol{X}}\right]<\infty,$$

by Assumption 4, and $K:=\sup_{\boldsymbol{\upsilon}\in\Theta}|\nu|\in(0,\infty)$ by the compactness of Θ. Therefore, whenever $||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_1\leq\frac{\varepsilon}{M}$ and $|\kappa-\kappa^{(h)}|>\frac{\varepsilon}{4K}$,

$$||\Phi(\boldsymbol{\upsilon})||_1\geq\left|\kappa-\kappa^{(h)}+\mathbb{E}\left[Ae^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}}-Ae^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}'\boldsymbol{X}}\right]\right|>0. \tag{44}$$

Finally, assume that $||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_1\leq\frac{\varepsilon}{M}$ and $|\kappa-\kappa^{(h)}|\leq\frac{\varepsilon}{4K}$, but $|\nu-\mu^{(h)}|>\frac{\varepsilon}{2\kappa^{(h)}}$. Then, as in (44),

$$\left|\mathbb{E}\left[AYe^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}}-AYe^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}'\boldsymbol{X}}\right]\right|\leq K_2(\gamma)\cdot\frac{\varepsilon}{M}\leq\frac{\varepsilon}{64K}, \tag{45}$$

by choosing $M\geq 64K\cdot K_2(\gamma)$, where

$$K_2(\gamma)^2 := e^{2\gamma}\mathbb{E}[||Y\boldsymbol{X}||_1^2]\cdot\mathbb{E}\left[\sup_{\boldsymbol{\beta}_*:||\boldsymbol{\beta}_*-\boldsymbol{\beta}_0||_1\leq\frac{\varepsilon}{M}} e^{-2\boldsymbol{\beta}_*\boldsymbol{X}}\right]<\infty,$$

by Assumption 4. Using (45) and $|\kappa^{(h)}\nu-\kappa\nu|\leq\frac{\varepsilon}{4}$ now gives

$$||\Phi(\boldsymbol{\upsilon})||_1=\left|\kappa\nu-\kappa^{(h)}\mu^{(h)}+\mathbb{E}\left[AYe^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}_0'\boldsymbol{X}}-AYe^{h(\boldsymbol{X},Y)-\boldsymbol{\beta}'\boldsymbol{X}}\right]\right|>0. \tag{46}$$

Combining (42), (44), and (46), we get, for all $\delta>0$, $\inf\{||\Phi(\boldsymbol{\upsilon})||^2:||\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_0||_1>\delta\}>0$, as required.

B.2.3. ***Proof of*** *(C).* We begin by showing the class of functions $\{\boldsymbol{t}\to Q(\boldsymbol{t}|\boldsymbol{\upsilon}):\boldsymbol{\upsilon}\in\Theta\}$ is $\mathbb{P}_0$-Donsker. To this end, let $\boldsymbol{\upsilon}_1=(\nu_1,\kappa_1,\boldsymbol{\beta}_1')'$ and $\boldsymbol{\upsilon}_2=(\nu_2,\kappa_2,\boldsymbol{\beta}_2')'$ be two points in the parameter space Θ, and recall the definition of $Q(\cdot|\boldsymbol{\upsilon})$ from (37). Then, by the mean-value theorem, there exists $t_*\in\mathbb{R}$ such that $t_*\in[\boldsymbol{\beta}_1'\boldsymbol{x},\boldsymbol{\beta}_2'\boldsymbol{x}]$ such that[2]

$$||Q_1(\boldsymbol{t}|\boldsymbol{\upsilon}_2)-Q_1(\boldsymbol{t}|\boldsymbol{\upsilon}_1)||_1\leq\left|\left|\left(\frac{e^{\boldsymbol{\beta}_2'\boldsymbol{x}}}{1+e^{t_*}}-\frac{e^{\boldsymbol{\beta}_1'\boldsymbol{x}}}{1+e^{\boldsymbol{\beta}_1'\boldsymbol{x}}}\right)\boldsymbol{x}\right|\right|_1\leq\left|\left|\left(\frac{e^{t_*}}{(1+e^{t_*})^2}\right)(\boldsymbol{\beta}_2-\boldsymbol{\beta}_1)'\boldsymbol{x}\boldsymbol{x}\right|\right|_1$$
$$\leq|(\boldsymbol{\beta}_2-\boldsymbol{\beta}_1)'\boldsymbol{x}|\,||\boldsymbol{x}||_1$$
$$\leq||\boldsymbol{\beta}_2-\boldsymbol{\beta}_1'||_2||\boldsymbol{x}||_2||\boldsymbol{x}||_1$$
$$\lesssim_d M_1(\boldsymbol{x})||\boldsymbol{\beta}_2-\boldsymbol{\beta}_1'||_1, \tag{47}$$

where $M_1(\boldsymbol{x})=||\boldsymbol{x}||_1^2$. Next, observe that

$$|Q_2(\boldsymbol{t}|\boldsymbol{\upsilon}_2)-Q_2(\boldsymbol{t}|\boldsymbol{\upsilon}_1)|\leq|\kappa_2-\kappa_1|+e^{\gamma}|e^{-\boldsymbol{\beta}_2'\boldsymbol{x}}-e^{-\boldsymbol{\beta}_1'\boldsymbol{x}}|$$
$$\lesssim_\gamma|\kappa_2-\kappa_1|+\left|(\boldsymbol{\beta}_2-\boldsymbol{\beta}_1)'\boldsymbol{x}e^{-t_*}\right| \qquad \text{(for some } t_*\in[\boldsymbol{\beta}_1'\boldsymbol{x},\boldsymbol{\beta}_2'\boldsymbol{x}])$$

[2]For $x,y\in\mathbb{R}$, $x\lesssim y$ means $x\leq Cy$, for some constant $C>0$. We will use subscripts $\lesssim_\square$, to denote that the constant may depend on the subscripted parameters.

$$\lesssim_\gamma |\kappa_2 - \kappa_1| + ||\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||_2 \cdot ||\boldsymbol{x}||_2 \sup_{\boldsymbol{\beta}\in\Theta_0} e^{-\boldsymbol{\beta}'\boldsymbol{x}}$$
$$\lesssim_{\gamma,d} M_2(\boldsymbol{x}) \left(|\kappa_2 - \kappa_1| + ||\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||_1\right), \tag{48}$$

where $M_2(\boldsymbol{x}) = 1 + ||\boldsymbol{x}||_1 \sup_{\boldsymbol{\beta}\in\Theta_0} e^{-\boldsymbol{\beta}'\boldsymbol{x}}$ and Θ_0 is the projection of the parameter space to the last d coordinates. Finally,

$$\begin{aligned}|Q_3(\boldsymbol{t}|\boldsymbol{\upsilon}_2) - Q_2(\boldsymbol{t}|\boldsymbol{\upsilon}_1)| &\le |\kappa_2\nu_2 - \kappa_1\nu_1| + e^\gamma |y e^{-\boldsymbol{\beta}_2'\boldsymbol{x}} - y e^{-\boldsymbol{\beta}_1'\boldsymbol{x}}| \\ &\lesssim_\gamma |\kappa_2 - \kappa_1| + |\nu_2 - \nu_1| + ||\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1'||_2 \cdot ||y\boldsymbol{x}||_2 \sup_{\boldsymbol{\beta}\in\Theta_0} e^{-\boldsymbol{\beta}'\boldsymbol{x}} \\ &\lesssim_{\gamma,d} M_3(\boldsymbol{x}, y) \left(|\kappa_2 - \kappa_1| + |\nu_2 - \nu_1| + ||\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||_1\right)\end{aligned}$$

where $M_3(\boldsymbol{x}, y) = 1 + ||y\boldsymbol{x}||_1 \sup_{\boldsymbol{\beta}\in\Theta_0} e^{-\boldsymbol{\beta}'\boldsymbol{x}}$.

Therefore, defining $M(\boldsymbol{x}, y) = M_1(\boldsymbol{x}) + M_2(\boldsymbol{x}) + M_3(\boldsymbol{x}, y)$ and combining (47), (48), and (49) gives

$$||Q(\boldsymbol{t}|\boldsymbol{\upsilon}_2) - Q(\boldsymbol{t}|\boldsymbol{\upsilon}_1)||_1 = \sum_{b=1}^{3} ||Q_b(\boldsymbol{t}|\boldsymbol{\upsilon}_2) - Q_b(\boldsymbol{t}|\boldsymbol{\upsilon}_1)||_1 \lesssim_{\gamma,d} M(\boldsymbol{x}, y)||\boldsymbol{\upsilon}_2 - \boldsymbol{\upsilon}_1||_1 \tag{49}$$

Since $\mathbb{E}[M(\boldsymbol{X}, Y)^2] < \infty$, by Assumption 4, this implies that the class $\{\boldsymbol{t} \to Q(\boldsymbol{t}|\boldsymbol{\upsilon}) : \boldsymbol{\upsilon} \in \Theta\}$ is $\mathbb{P}$-Donsker.

The inequalities in (47), (48), and (49), combined with Assumption 4 also implies that $\mathbb{E}[(Q(\boldsymbol{T}|\boldsymbol{\upsilon}_n) - Q(\boldsymbol{T}|\boldsymbol{\upsilon}_0))^2] \to 0$, coordinate-wise, whenever $||\boldsymbol{\upsilon}_n - \boldsymbol{\upsilon}_0||_1 \to 0$.

B.3. **Completing the Proof of Theorem 3.** To begin with note that the α-th bootstrap quantile $Q_\alpha(\hat{\hat{\mu}}^{(h)}) = \hat{\mu}^{(h)} + \frac{\hat{z}_{\alpha,n}^{(h)}}{\sqrt{n}}$, where $\hat{z}_{\alpha,n}^{(h)} := \inf\{t : \hat{\mathbb{P}}_n(\sqrt{n}(\hat{\hat{\mu}}^{(h)} - \hat{\mu}^{(h)}) \le t) \ge \alpha\}$. By Corollary 9, as $n \to \infty$, $\hat{z}_{\alpha,n}^{(h)} = z_\alpha^{(h)} + O_P(1/\sqrt{n})$, where $z_\alpha^{(h)}$ is α-th quantile of $N(0, (\sigma^{(h)})^2)$. Then, recalling (15), as $n \to \infty$,

$$\mathbb{P}_0\left(\mu^{(h)} < L_B^{(h)}\right) = \mathbb{P}_0\left(\sqrt{n}\left(\hat{\mu}^{(h)} - \mu^{(h)}\right) \le z_{\frac{\alpha}{2},n}^{(h)}\right) \to \frac{\alpha}{2},$$

by Corollary 9. The limit of $\mathbb{P}_0\left(\mu^{(h)} > U_B^{(h)}\right)$ follows similarly.

Appendix C. Proof of Corollary 6

The proof of Corollary 6 is similar to the proof of Theorem 4. We begin by defining

$$\kappa_0^{(h)} := \mathbb{E}\left[\frac{1 - A}{e_0^{(h)}(\boldsymbol{X}, Y)}\right], \quad \kappa_1^{(h)} := \mathbb{E}\left[\frac{A}{e_1^{(h)}(\boldsymbol{X}, Y)}\right].$$

We now set up the SIPW estimate of the ATE (recall (26)) as a Z-estimation problem. To this end, given a vector $\boldsymbol{\upsilon} = (\nu_0, \nu_1, \kappa_0^{(h)}, \kappa_1^{(h)}, \boldsymbol{\beta}')' \in \Theta \subset \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d$, where Θ is the parameter space, define the function $Q : \{0, 1\} \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d+4}$ as follows: For $\boldsymbol{t} = (a, \boldsymbol{x}', y)' \in \{0, 1\} \times \mathbb{R} \times \mathbb{R}^d$,

$$Q(\boldsymbol{t}|\boldsymbol{\upsilon}) = \begin{pmatrix} Q_1(\boldsymbol{t}|\boldsymbol{\upsilon}) \\ Q_2(\boldsymbol{t}|\boldsymbol{\upsilon}) \\ Q_3(\boldsymbol{t}|\boldsymbol{\upsilon}) \\ Q_4(\boldsymbol{t}|\boldsymbol{\upsilon}) \\ Q_5(\boldsymbol{t}|\boldsymbol{\upsilon}) \end{pmatrix} := \begin{pmatrix} \left(a - \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1+e^{\boldsymbol{\beta}'\boldsymbol{x}}}\right)\boldsymbol{x} \\ \kappa_1^{(h)} - a\left(1 + e^{h_1(\boldsymbol{x},y)-\boldsymbol{\beta}'\boldsymbol{x}}\right) \\ \kappa_1^{(h)}\nu_1 - ay\left(1 + e^{h_1(\boldsymbol{x},y)-\boldsymbol{\beta}'\boldsymbol{x}}\right) \\ \kappa_0^{(h)} - (1-a)\left(1 + e^{-h_0(\boldsymbol{x},y)+\boldsymbol{\beta}'\boldsymbol{x}}\right) \\ \kappa_0^{(h)}\nu_0 - (1-a)y\left(1 + e^{-h_0(\boldsymbol{x},y)+\boldsymbol{\beta}'\boldsymbol{x}}\right) \end{pmatrix}. \tag{50}$$

Next, define $\Phi(\boldsymbol{\upsilon}) = \int Q(\boldsymbol{t}|\boldsymbol{\upsilon})\mathrm{d}\mathbb{P}_0(\boldsymbol{t})$, where $\boldsymbol{T} = (A, \boldsymbol{X}', Y)' \sim \mathbb{P}_0$, the true distribution generating the data. Note that $\Phi(\boldsymbol{\upsilon}_0) = 0$, where $\boldsymbol{\upsilon}_0 = (\mu^{(h_0)}(0), \mu^{(h_1)}(1), \kappa_0^{(h)}, \kappa_1^{(h)}, \boldsymbol{\beta}_0')$ is the true parameter value. Then, as in (38), Z-estimates $\hat{\boldsymbol{\upsilon}} = (\hat{\mu}^{(h_0)}(0), \hat{\mu}^{(h_1)}(1), \hat{\kappa}_0^{(h_0)}, \hat{\kappa}_1^{(h_1)}, \hat{\boldsymbol{\beta}}')'$ are obtained by solving the equations

$$\Phi_n(\hat{\boldsymbol{\upsilon}}) := \frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{T}_i|\hat{\boldsymbol{\upsilon}}) = 0.$$

It is easy to see that the Z estimate $\hat{\mu}^{(h_1)}(1) - \hat{\mu}^{(h_0)}(0)$ is exactly the SIPW estimate (26) for $\Delta^{(h_0,h_1)}$ and $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ for the logistic regression model. Moreover, as in (39), the bootstrap Z-estimates $\hat{\hat{\boldsymbol{\upsilon}}}$ are obtained from the equations:

$$\hat{\Phi}_n(\hat{\hat{\boldsymbol{\upsilon}}}) := \frac{1}{n}\sum_{i=1}^{n} Q(\hat{\boldsymbol{T}}_i|\hat{\hat{\boldsymbol{\upsilon}}}) = 0.$$

Then, as in Theorem 8, the limiting joint normality of $(\hat{\mu}^{(h_0)}(0), \hat{\mu}^{(h_1)}(1))$, and hence the asymptotic normality of $\hat{\Delta}^{(h_0,h_1)} = \hat{\mu}^{(h_0)}(0) - \hat{\mu}^{(h_1)}(1)$, can be derived. This would imply the asymptotic validity of the confidence interval (27), as in Section B.3, completing the proof of Corollary 6. Details are omitted.