Title

Assessing and planning for unmeasured confounding in weighted observational studies

Permalink

https://escholarship.org/uc/item/84j1b24v

Author

Soriano, Daniel William

Publication Date

2023

Peer reviewed|Thesis/dissertation

Assessing and planning for unmeasured confounding in weighted observational studies

by

Daniel W Soriano

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Avi Feller, Co-chair
Professor Samuel Pimentel, Co-chair
Professor Peng Ding

Summer 2023

Assessing and planning for unmeasured confounding in weighted observational studies

Abstract

Assessing and planning for unmeasured confounding in weighted observational studies

by

Daniel W Soriano

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Avi Feller, Co-chair

Professor Samuel Pimentel, Co-chair

The ability to compare similar groups is central to causal inference. If two groups are the same except that one group received a treatment and the other group did not, we can attribute the difference in an outcome of interest to the treatment (Cochran, 1965). For this reason, randomized experiments are often considered to be the "gold standard" for estimating causal effects: when the treatment is randomly assigned, the treatment and control groups are comparable on average. In many settings, it might be unethical or otherwise infeasible for a researcher to randomly assign treatment. In these cases, researchers must rely on observational data to investigate their causal hypotheses.

Aside from their greater feasibility in many instances, there are a few possible benefits of observational studies compared to randomized experiments. Observational studies typically consist of larger, naturally occurring samples that more closely resemble a target population. However, there is no guarantee that the treatment and control groups are comparable in an observational study since units can select into a group. For example, in a study evaluating the effectiveness of a medication on a health outcome of interest, patients that are sicker to begin with might be more likely to take the treatment, biasing direct comparison of treatment and control groups. A common strategy to attempt to mitigate this bias is to adjust for observed covariates so that the adjusted treatment and control groups are comparable in terms of these covariates.

These methods that attempt to adjust for observed covariates rely on the key assumption that there are no unmeasured confounders that simultaneously impact the treatment and outcome, often referred to as *ignorability* or *unconfoundedness*. However, this assumption is not verifiable from observed data and never exactly holds in most real-world settings. Since we are still interested in studying causal relationships from observational data, the ignorability assumption is at the core of this thesis. First, we develop a framework to evaluate

how robust causal effect estimates are to violations of the ignorability assumption. Then, we investigate how to design observational studies to improve robustness to unmeasured confounding, rather than selecting designs that are optimal under the ignorability assumption. Chapter 1 briefly reviews these topics, and the following chapters detail our proposed frameworks.

Chapter 2 focuses on assessing the robustness of weighted observational studies to violations of the ignorability assumption. We develop a sensitivity analysis framework for a broad class of weighting estimators that allows for specified levels of unmeasured confounding, resulting in a range of possible effect estimates, rather than a single point estimate. We prove that the percentile bootstrap procedure can yield valid confidence intervals for causal effects under our sensitivity analysis framework. We also propose an amplification — a mapping from a one-dimensional sensitivity analysis to a higher dimensional sensitivity analysis — to enhance the interpretability of our sensitivity analysis's results, aiding researchers in reasoning about plausible levels of confounding in particular observational studies. We illustrate our sensitivity analysis procedure through real data examples.

Chapter 3 builds on Chapter 2 by focusing on how to design observational studies such that they are robust to unmeasured confounding, rather than optimal under ignorability. Specifically, we introduce a measure called design sensitivity for weighting estimators, which describes the asymptotic power of a sensitivity analysis. By comparing design sensitivities, we assess of how different design decisions impact sensitivity to unmeasured confounding. While sensitivity analysis is conducted post-hoc as a secondary analysis, design sensitivity enables researchers to plan ahead and optimize for robustness at the design stage. We illustrate our proposed framework on data evaluating the drivers of support for the 2016 Colombian peace agreement.

To my family.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **v**

1 Weighting, sensitivity analysis, and design sensitivity for causal inference **1**

1.1 Weighting 1
1.2 Sensitivity analysis 2
1.3 Design sensitivity 4

2 Interpretable Sensitivity Analysis for Balancing Weights **5**

2.1 Introduction 5
2.2 Background, notation, and review 7
2.3 Sensitivity analysis for balancing weights estimators 10
2.4 Amplifying, interpreting, and calibrating sensitivity parameters 13
2.5 Numerical examples 15
2.6 Discussion 22

3 Design Sensitivity and Its Implications for Weighted Observational Studies **23**

3.1 Introduction 23
3.2 Background 24
3.3 Design Sensitivity for Weighted Estimators 28
3.4 Design Choices that Impact Design Sensitivity 32
3.5 Empirical Application: Colombia FARC Peace Agreement 36
3.6 Conclusion 38

Bibliography **40**

A Supplementary materials for Chapter 2 **46**

A.1 Proofs 46
A.2 Simulation for sample splitting 51
A.3 Average treatment effect on the treated 52

B Supplementary materials for Chapter 3 54
B.1 Proofs . . . 54
B.2 Using a Planning Sample to Estimate Design Sensitivity . . . 67
B.3 Detailed Simulation Results . . . 75

List of Figures

2.1 Sensitivity analysis results with the LaLonde data 17
2.2 Sensitivity analysis results with the fish diet data 19
2.3 Comparison of confidence interval width after sensitivity analysis for three approaches in the fish consumption example 21

3.1 Magnitude of the design sensitivities for both MSM and VBM varying different aspects of the data generating process. 32
3.2 Design sensitivities under augmentation and trimming for the variance-based model and the marginal sensitivity model 37

A.1 Bootstrap distributions of estimates of μ_0 with the full data and with sample splitting 51

B.1 Distribution of estimated design sensitivity measures across 100 different planning samples 70

List of Tables

B.1 Estimating Design Sensitivity using a Planning Sample 68
B.2 Estimating Design Sensitivity using a Planning Sample and Simulated Outcomes 69
B.3 Estimating Design Sensitivity for Augmentation using a Proxy Outcome Model 72
B.4 Design sensitivity estimated using full FARC data 73
B.5 Estimated power for analysis sample using FARC data 74
B.6 Drivers of design sensitivity 78
B.7 Impact of trimming on design sensitivity for varying treatment effect heterogeneity 79
B.8 Impact of augmentation on design sensitivity under outcome model misspecification 80

Acknowledgments

I'd like to thank my advisors Avi Feller and Sam Pimentel, who have both been instrumental to my intellectual and professional development. Avi and Sam were extremely patient and encouraging as I got started with research during my first year. As I progressed through my PhD, Avi and Sam provided invaluable guidance and support, for which I will be forever grateful. I hope to incorporate lessons from their excellent mentorship to future scenarios in which I have the opportunity to mentor future researchers.

I am thankful to Peng Ding, along with Avi and Sam, for fostering a welcoming and stimulating environment in the causal inference reading group. Attending the reading group during my first year drew me to pursue causal inference as my research focus. In addition to serving on my committee, Peng's thorough and precise teaching in his causal inference and linear models courses helped me establish solid foundations in those key areas. It is rare that I go a week without consulting Peng's indispensable notes on causal inference. Without Peter Bickel, I don't know if Chapter 2 would have been possible. I am extremely grateful for his patience and generosity with his time, in addition to serving on my qualifying exam committee.

Several fellow PhD students played key roles in my research career. Eli Ben-Michael's mentorship was essential for my research and professional development. Melody Huang was a tremendous co-author and made the process to produce Chapter 3 far more enjoyable. I am grateful to Licong Lin for his generosity in helping with Chapter 3. I would also like to thank La Shana Porlaris, as well as the entire Statistics Department staff, who were instrumental in helping me stay on track during my PhD, in addition to making the department a fun and welcoming place to be.

I am extremely grateful to all of the SGSA presidents, both past and present, and graduate students who served on SGSA committees for making the graduate student experience in the department so special. I am especially thankful for Ella Hiesmayr, who perfectly complemented me when we served as SGSA co-presidents during our third years. Ella worked tirelessly to make the department a better place. I would like to thank Andy Shen, Corrine Elliott, Shuni Li, and Kellie Ottoboni for making office 345 a fun and relaxed environment. I would also like to thank my cohort, who made the grind of first-year core courses and late nights in Evans more tolerable. Finally, my parents, brother, and sister: without them I don't know where I'd be.

Chapter 1

Weighting, sensitivity analysis, and design sensitivity for causal inference

Consider an observational study with n units sampled identically and independently from a population with pre-treatment covariates $X_i \in \mathbb{R}^d$, binary treatment indicator $Z_i \in \{0, 1\}$, and outcome $Y_i \in \mathbb{R}$. We posit the existence of *potential outcomes*: the outcome had unit i received the treatment, $Y_i(1)$, and the outcome had unit i received the control, $Y_i(0)$ (Neyman, 1923; Rubin, 1974). Assuming stable treatment and no interference between units (Rubin, 1980b), the observed outcome is $Y_i = (1 - Z_i)Y_i(0) + Z_iY_i(1)$. The focus of this thesis is the key identification assumption that the potential outcomes are independent of the treatment given the pre-treatment covariates. Combined with the assumption that the treatment assignment is not deterministic conditional on X, these assumptions are commonly known as *strong ignorability* (Rosenbaum and Rubin, 1983b).

Assumption 1.1 (Ignorability). $Y(0), Y(1) \perp\!\!\!\perp Z \mid X$.

Assumption 1.2 (Overlap). The *propensity score* $\pi(x) \equiv P(Z = 1 \mid X = x)$ satisfies $0 < \pi(x) < 1$ for all $x \in \mathcal{X}$.

In this chapter, we focus on estimating the *Population Average Treatment Effect* (PATE):

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mu_1 - \mu_0, \tag{1.1}$$

where $\mu_1 = \mathbb{E}[Y(1)]$ and $\mu_0 = \mathbb{E}[Y(0)]$.

1.1 Weighting

A popular method to estimate treatment effects in observational studies is to weights units such that the covariate distributions in the treatment and control groups are similar. Rosenbaum and Rubin (1983b)'s influential work on the central role that the propensity score plays in observational studies motivates many weighting strategies. The first key result illuminates that the propensity score can be used for dimension reduction while still maintaining ignorability. If strong ignorability holds conditional on the pre-treatment covariates X, then the potential outcomes and treatment are also independent conditional on the propensity score $\pi(X)$. Therefore, if X is sufficient to remove confounding between Z and potential outcomes $Y(0), Y(1)$, then $\pi(X)$ is as well.

Under Assumptions 1.1 and 1.2, we can non-parametrically identify μ_1 using observed outcomes and propensity scores from treated units:

$$\mathbb{E}\left[\frac{ZY}{\pi(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{ZY}{\pi(X)} \mid X\right]\right] = \mathbb{E}[Y(1)]. \tag{1.2}$$

A similar result holds for μ_0. This motivates the inverse propensity score weighting (IPW) estimator

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{\pi}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)\,Y_i}{1-\hat{\pi}(X_i)}, \tag{1.3}$$

where $\hat{\pi}(X_i)$ is an estimate of the propensity score. Since the IPW estimator is not invariant to location transformations of the outcome and can be unstable when estimated propensity scores are near 0 or 1, an alternative estimator is the stabilized IPW (SIPW) estimator with normalized weights

$$\hat{\tau}_{\text{SIPW}} = \frac{\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{\pi}(X_i)}}{\sum_{i=1}^{n}\frac{Z_i}{\hat{\pi}(X_i)}} - \frac{\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{\pi}(X_i)}}{\sum_{i=1}^{n}\frac{1-Z_i}{1-\hat{\pi}(X_i)}}. \tag{1.4}$$

IPW estimators are a subset of more general weighting estimators that replace inverse propensity scores with more general weights:

$$\hat{\tau}_W = \frac{\sum_{i=1}^{n}\hat{w}_i Z_i Y_i}{\sum_{i=1}^{n}\hat{w}_i Z_i} - \frac{\sum_{i=1}^{n}\tilde{w}_i\,(1-Z_i)\,Y_i}{\sum_{i=1}^{n}\tilde{w}_i\,(1-Z_i)}. \tag{1.5}$$

An increasingly popular alternative approach to estimating weights is to solve a constrained optimization problem to obtain weights that satisfy constraints on covariate balance. Hence, rather than estimating propensity scores and using them to form weights, this class of balancing weights estimators directly targets covariate balance. See Ben-Michael et al. (2021) for a recent review.

1.2 Sensitivity analysis

In randomized experiments, random treatment assignment yields balanced treatment and control groups on average. On the other hand, researchers employ techniques such as weighting and matching to adjust for observed covariates X in observational studies to form treatment and control groups that are comparable in terms of the observed covariates. Since unobserved covariates U are not included in the estimation procedure, there is no guarantee that the treatment and control groups are comparable in terms of U. Therefore, researchers typically rely on the ignorability assumption (1.1) to estimate causal effects from observational studies.

Unfortunately, the ignorability assumption is almost never exactly true in observational studies, leading to serious concerns about the reliability of their findings. In order to test how robust a study's results are to violations of ignorability, researchers can run a sensitivity analysis. A sensitivity analysis typically relaxes the ignorability assumption, allowing for a specified magnitude of bias from unmeasured confounding, and examines how the causal effect estimates change. The sensitivity of an observational study is the degree of violation of ignorability needed to alter the study's conclusions. If a large amount of confounding is needed, then the study is robust, enhancing its reliability. Conversely, a sensitive study would require only a small deviation from ignorability.

Sensitivity analysis dates back to Cornfield et al. (1959), who conducted a formal sensitivity analysis of an observational study examining the effect of smoking on lung cancer. They determined that significant bias from unmeasured confounding would be required to change the conclusion that smoking causes lung cancer. Ding and VanderWeele (2016) build off of Cornfield et al. (1959)'s framework to deliver stronger conclusions while making no assumptions about the structure of the unmeasured confounder or confounders. Among other more recent sensitivity analysis frameworks, Rosenbaum (2002)'s sensitivity model stipulates that two units, j and k, with the same observed covariates, $X_j = X_k$, must have treatment odds that differ by at most a multiplier of sensitivity analysis parameter $\Gamma \geq 1$:

$$\frac{1}{\Gamma} \leq \frac{\pi_j / (1 - \pi_j)}{\pi_k / (1 - \pi_k)} \leq \Gamma, \tag{1.6}$$

where $\pi_i = \Pr(Z_i = 1 \mid Y_i(0), Y_i(1), X_i, U_i)$. $\Gamma = 1$ represents no unmeasured confounding. As Γ increases, the sensitivity model allows for larger deviations from ignorability.

Most pertinent to this dissertation is the growing literature on sensitivity analysis for weighting estimators. We highlight two closely related sensitivity models that relax the ignorability assumption, the *marginal sensitivity model* and the *variance-based sensitivity model*. For simplicity, consider estimating $\tau_{\text{att}} = \mathbb{E}[Y(1) - Y(0) \mid Z = 1]$, let w be the population inverse propensity score weights that condition on X alone, and define the ideal weights w^* to be the population-level inverse propensity score weights in (X, U). Originally introduced by Tan (2006) and later studied by Zhao et al. (2019); Soriano et al. (2021); Dorn and Guo (2021), the marginal sensitivity model $\nu_{msm}(\Lambda, w)$ constrains the worst-case error from omitting a confounder, positing that the ratio between any ideal weight w^* and corresponding w may not exceed $\Lambda \geq 1$:

$$\nu_{msm}(\Lambda, w) := \left\{ w^* : \Lambda^{-1} \leq \frac{w^*}{w} \leq \Lambda \right\}. \tag{1.7}$$

The variance-based sensitivity model $\nu_{vbm}(R^2, w)$ constrains the variance in w^* not explained by w (Huang and Pimentel, 2022). More formally, for some $R^2 \in [0, 1)$, the variance-based sensitivity model is defined as follows:

$$\nu_{vbm}(R^2, w) := \left\{ w^* : 1 \leq \frac{\mathrm{var}(w^* \mid Z = 0)}{\mathrm{var}(w \mid Z = 0)} \leq \frac{1}{1 - R^2} \right\}. \tag{1.8}$$

We study both sensitivity models in depth in the chapters that follow.

1.3 Design sensitivity

In randomized experiments, a desirable feature of a statistical test is for it to have high power; i.e., to detect a true treatment effect with high probability. Similarly, we strive to maximize the power of a sensitivity analysis in observational studies. The power of a sensitivity analysis is the probability that the sensitivity analysis rejects the null hypothesis of no treatment effect for a given sensitivity model and value of the sensitivity parameter under a *favorable* situation, meaning (1) the study is free of unmeasured bias, and in addition (2) there is a treatment effect large enough to be of interest.

While it may seem paradoxical to evaluate a sensitivity analysis when no unmeasured confounding is present, in practice, researchers do not know whether the favorable situation is true from observable data and must conduct a sensitivity analysis anyway. Therefore, a natural goal when in the favorable situation is to conclude that treatment appears to have an effect that is highly insensitive to unmeasured biases.

The need to repeatedly compute power for each value of n and the sensitivity parameter makes it difficult to learn general principles about the behavior of the power. As $n \to \infty$, there is a value of the sensitivity parameter such that the power of a sensitivity analysis goes to 1 for all values of the sensitivity parameter less than that value and to 0 for all greater than that value. This value is called the *design sensitivity*.

While sensitivity analysis is conducted post-hoc as a secondary analysis, design sensitivity enables researchers to plan ahead and tailor studies to improve robustness to unmeasured confounding. Originally introduced by Rosenbaum (2004), design sensitivity has been studied extensively for matched studies (Heller et al., 2009; Rosenbaum et al., 2010; Hsu et al., 2013). Implications include the value of reducing heterogeneity within matched pairs (Rosenbaum, 2005) and choosing test statistics (Rosenbaum, 2011; Howard and Pimentel, 2021) and treatment doses (Rosenbaum, 2004) carefully. Chapter 3 focuses on developing design sensitivity for weighting and exploring its implications for important design choices.

Chapter 2

Interpretable Sensitivity Analysis for Balancing Weights

Assessing sensitivity to unmeasured confounding is an important step in observational studies, which typically estimate effects under the assumption that all confounders are measured. In this chapter, we develop a sensitivity analysis framework for balancing weights estimators, an increasingly popular approach that solves an optimization problem to obtain weights that directly minimizes covariate imbalance. In particular, we adapt a sensitivity analysis framework using the percentile bootstrap for a broad class of balancing weights estimators. We prove that the percentile bootstrap procedure can, with only minor modifications, yield valid confidence intervals for causal effects under restrictions on the level of unmeasured confounding. We also propose an amplification — a mapping from a one-dimensional sensitivity analysis to a higher dimensional sensitivity analysis — to allow for interpretable sensitivity parameters in the balancing weights framework. We illustrate our method through extensive real data examples.

2.1 Introduction

Observational studies can be an important source of evidence about causal effects across the medical and social sciences. Observational studies may be feasible in cases where randomized trials are not, or at least substantially less onerous to conduct at scale, but they raise challenges for analysis that are not present in randomized studies. As one example, consider evaluating the degree to which diets rich in fish elevate blood mercury relative to diets containing little fish. High levels of mercury in the blood can pose health risks; for instance, infants whose mothers had high mercury levels may be at increased risk for adverse neurodevelopmental events (Mahaffey et al., 2004). Consumption of fish or shellfish has been identified as a major source of mercury in the blood (Björnberg et al., 2003). These effects could be measured by randomly assigning subjects to high- and low-fish diets over long periods of time and comparing their blood mercury, but such experiments may be difficult to conduct and suffer from problems with compliance. Observational data describing blood mercury levels for subjects who choose to eat large or small amounts of fish are more readily available, but direct comparisons between groups are subject to confounding if the high-fish-diet and low-fish-diet subjects are systematically different in other ways. Similarly, measuring the impact of job training programs on wages using randomized experiments is expensive and difficult, but observational studies suffer from substantial confounding (LaLonde, 1986).

In observational studies for both examples just described, some confounding may be apparent in the form of obvious differences in observed variables between comparison groups, and analysis often proceeds under a key assumption that all confounders are measured, sometimes known as *ignorability* or *unconfoundedness*. However, this assumption is not verifiable from observed data, and it is often easy to suggest unmeasured factors that may contribute at least a limited amount of confounding. For example, in the case of job training programs, one might wonder if individuals who choose to participate in job training may have higher intrinsic motivation to succeed than those who choose not to. A sensitivity analysis seeks to determine the magnitude of unobserved confounding required to alter a study's findings. If a large amount of confounding is needed, then the study is robust, enhancing its reliability. Assessing sensitivity to unmeasured confounding is a critical part of the workflow for causal inference in observational studies.

In this chapter, we develop a sensitivity analysis framework for *balancing weights estimators*. Building on classical methods from survey calibration, these estimators find weights that minimize covariate imbalance between a weighted average of the observed units and a given distribution, such as by re-weighting control units to have a similar covariate distribution to the treated units. Balancing weights have become increasingly common within causal inference, with better finite sample properties than traditional inverse propensity score weighting (IPW). See Section 2.2 for additional details and Ben-Michael et al. (2021) for a recent review.

Our proposed sensitivity analysis framework adapts the percentile bootstrap sensitivity analysis that Zhao et al. (2019) develop for traditional IPW. Specifically, for a given sensitivity parameter, we compute the upper and lower bounds of our estimator for each bootstrap sample, and then form a confidence interval using percentiles across bootstrap samples. We prove that this approach yields valid confidence intervals for our proposed sensitivity analysis procedure over a broad class of balancing weights estimators.

To make a sensitivity analysis more interpretable, Rosenbaum and Silber (2009) introduce an *amplification* of a sensitivity analysis, which is a mapping from each point in a low-dimensional sensitivity analysis to a set of points in a higher-dimensional sensitivity analysis that all have the same possible inferences. We propose a new amplification that expresses the bias from confounding in terms of: (1) the imbalance in an unobserved covariate; and (2) the strength of the relationship between the outcome and the unobserved covariate. Researchers can then relate the results of our amplification to estimates from observed covariates. We demonstrate this approach via a numerical illustration and via several applications.

2.2 Background, notation, and review

Setup and review of marginal sensitivity model

We consider an observational study setting with independently and identically distributed data (Y_i, X_i, Z_i), $i \in \{1, \dots, n\}$, drawn from some joint distribution $P(\cdot)$ with outcome $Y_i \in \mathbb{R}$, covariates $X_i \in \mathcal{X}$, and treatment assignment $Z_i \in \{0, 1\}$. We posit the existence of *potential outcomes*: the outcome had unit i received the treatment, $Y_i(1)$, and the outcome had unit i received the control, $Y_i(0)$ (Neyman, 1923; Rubin, 1974). We assume stable treatment and no interference between units (Rubin, 1980b), so the observed outcome is $Y_i = (1 - Z_i)Y_i(0) + Z_iY_i(1)$. An estimand of interest is the *Population Average Treatment Effect* (PATE):

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mu_1 - \mu_0, \tag{2.1}$$

where $\mu_1 = \mathbb{E}[Y(1)]$ and $\mu_0 = \mathbb{E}[Y(0)]$. To simplify the exposition, we will focus on estimating μ_1; estimating μ_0 is symmetric. We consider an alternative estimand, the *Population Average Treatment Effect on the Treated* (PATT) in Section 2.5 and Appendix A.3.

A common set of identification assumptions in this setting, known as *strong ignorability*, assumes that conditioning on the covariates X sufficiently removes confounding between treatment Z and the potential outcomes $Y(0), Y(1)$, and that treatment assignment is not deterministic given X (Rosenbaum and Rubin, 1983b).

Assumption 2.1 (Ignorability)**.** $Y(0), Y(1) \perp\!\!\!\perp Z \mid X$.

Assumption 2.2 (Overlap)**.** The *propensity score* $\pi(x) \equiv P(Z = 1 \mid X = x)$ satisfies $0 < \pi(x) < 1$ for all $x \in \mathcal{X}$.

Under Assumptions 2.1 and 2.2, we can non-parametrically identify μ_1, solely with the outcomes from units receiving treatment,

$$\mu_1 = \mathbb{E}\left[\frac{ZY}{\pi(X)}\right]. \tag{2.2}$$

In an observational setting, the researcher does not know the *true* treatment assignment mechanism, $\pi(x, y) \equiv P(Z = 1 \mid X = x, Y(1) = y)$, which in general can depend on *both* the covariates X and the potential outcomes $Y(1)$ and $Y(0)$. A rich literature assesses the sensitivity of estimates to violations of the ignorability assumption. This approach dates back at least to Cornfield et al. (1959), who conducted a formal sensitivity analysis of the effect of smoking on lung cancer. More recent examples of sensitivity analysis include Rosenbaum and Rubin (1983a), Rosenbaum (2002), VanderWeele and Ding (2017), Franks et al. (2019), Tudball et al. (2019), Cinelli and Hazlett (2020), Fogarty (2020), Huang (2022), and Huang and Pimentel (2022). See Hong et al. (2020) for a recent discussion of weighting-based sensitivity methods.

We adopt the marginal sensitivity model proposed originally by Tan (2006) and further developed by Zhao et al. (2019) and Dorn and Guo (2021) for traditional IPW weights. Following these authors, we split the problem into two parts: sensitivity for the mean of the treated potential outcomes and sensitivity for the mean of the control potential outcomes; without loss of generality, we consider the mean for the treated potential outcomes. Since unbiased estimation of $\mathbb{E}[Y(1)]$ requires knowledge only of $\pi(x, y) = P(Z = 1 \mid X = x, Y(1) = y)$ rather than the full propensity score that also conditions on $Y(0)$, we can rewrite Assumption 2.1 as $\pi(x, y) = \pi(x)$. For details on combining sensitivity analyses for $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ into a single sensitivity analysis for the ATE, see Section 5 from Zhao et al. (2019).

The marginal sensitivity model relaxes the ignorability assumption so that the odds ratio between the two conditional probabilities $\pi(x)$ and $\pi(x, y)$ is bounded.

Assumption 2.3 (Marginal sensitivity model)**.** For $\Lambda \geq 1$, the true propensity score satisfies

$$\pi(x, y) \in \mathcal{E}(\Lambda) = \left\{\pi(x, y) \in (0, 1) : \Lambda^{-1} \leq \mathrm{OR}(\pi(x), \pi(x, y)) \leq \Lambda\right\},$$

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

Here, Λ is a sensitivity parameter, quantifying the difference between the true propensity score $\pi(x, y)$ and the probability of treatment given $X = x$, $\pi(x)$; when $\Lambda = 1$, the two probabilities are equivalent, and Assumption 2.1 holds. If, for example, $\Lambda = 2$, Assumption 2.3 constrains the odds ratio between $\pi(x)$ and $\pi(x, y)$ to be between $\frac{1}{2}$ and 2.

Again following Zhao et al. (2019), we will consider an equivalent characterization of the set $\mathcal{E}(\Lambda)$ in terms of the log odds ratio $h(x, y) = \log \mathrm{OR}(\pi(x), \pi(x, y))$:

$$\mathcal{H}(\Lambda) = \{h : \mathcal{X} \times \mathbb{R} \to \mathbb{R} : \|h\|_\infty \leq \log \Lambda\}, \tag{2.3}$$

where $\|h\|_\infty = \sup_{x \in \mathcal{X}, y \in \mathbb{R}} |h(x, y)|$ is the supremum norm. Rearranging the definition of $h(x, y)$ to be $\log \frac{\pi(x,y)}{1-\pi(x,y)} = \log \frac{\pi(x)}{1-\pi(x)} - h\,(x, y)$ and applying the inverse logit transformation, we can write the true propensity score under a particular sensitivity model h as

$$\pi^{(h)}(x, y) = \left[1 + \left(\frac{1}{\pi(x)} - 1\right) e^{h(x,y)}\right]^{-1}. \tag{2.4}$$

Zhao et al. (2019) refer to $\pi^{(h)}(x, y)$ as the *shifted propensity score*. Then, for a particular $h \in \mathcal{H}(\Lambda)$, we can write the *shifted estimand* as

$$\mu_1^{(h)} = \mathbb{E}\left[\frac{Z}{\pi^{(h)}(X, Y(1))}\right]^{-1} \mathbb{E}\left[\frac{ZY}{\pi^{(h)}(X, Y(1))}\right]. \tag{2.5}$$

[1] Zhao et al. (2019) introduce an extension to the marginal sensitivity model that they call the parametric marginal sensitivity model. The parametric marginal sensitivity model replaces $\pi(x)$ with the best parametric approximation to $\pi(x)$, $\pi_\beta(x)$, and compares $\pi(x, y)$ to $\pi_\beta(x)$ so that the sensitivity analysis addresses both model misspecification and unobserved confounding.

Under the marginal sensitivity model in Assumption 2.3, we then have a non-parametric partial identification bound, $\inf_{h \in \mathcal{H}(\Lambda)} \mu_1^{(h)} \leq \mu_1 \leq \sup_{h \in \mathcal{H}(\Lambda)} \mu_1^{(h)}$.

The bound just given depends on population quantities that must be estimated, and in practice it is important to take sampling uncertainty into account. Zhao et al. (2019) use the percentile bootstrap to build confidence intervals that cover this partial identification set, under the assumption that the weights are constructed using IPW.

We go beyond Zhao et al. (2019)'s work in two important ways. In Section 2.3, we show that the percentile bootstrap strategy for constructing confidence intervals is valid for the broader class of balancing weights, not just IPW. This requires a different proof strategy than the one based on Z-estimation used by Zhao et al. (2019) in order to handle balancing weights estimators that achieve approximate (rather than exact) balance on covariates, such as the stable balancing weights of Zubizarreta (2015). In Section 2.4 we then introduce an amplification that allows us to better interpret and calibrate marginal sensitivity analyses.

Weighting estimators under strong ignorability

We estimate μ_1 via a weighted average of treated units' outcomes using weights $\hat{\gamma}(X)$,

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{Z_i \hat{\gamma}(X_i)}{\sum_{i=1}^{n} Z_i \hat{\gamma}(X_i)} Y_i. \tag{2.6}$$

Under strong ignorability (Assumptions 2.1 and 2.2), traditional Inverse Propensity Score Weighting (IPW) first models the propensity score, $\hat{\pi}(x)$, directly and then sets weights to be $\hat{\gamma}(X_i) = \frac{1}{\hat{\pi}(X_i)}$. Thus, $\hat{\mu}_1$ is a plug-in version of Equation (2.2). This approach can perform poorly in moderate to high dimensions or when there is poor overlap and either $\pi(x)$ or $\hat{\pi}(x)$ is near 0 or 1 (Kang et al., 2007).

Balancing weights, by contrast, directly optimize for covariate balance; recent proposals include Hainmueller (2012); Zubizarreta (2015); Athey et al. (2018); Wang and Zubizarreta (2019); Hirshberg et al. (2019); Tan (2020) and have a long history in survey calibration for non-response (Deville and Särndal, 1992; Deville et al., 1993). See Chattopadhyay et al. (2020) and Ben-Michael et al. (2021) for recent reviews.

Most balancing weights estimators attempt to control the imbalance between the weighted treated sample and the full sample in some transformation of the covariates $\phi : \mathcal{X} \to \mathbb{R}^d$. For example, Zubizarreta (2015) proposes *stable balancing weights* (SBW) that find weights $\hat{\gamma}(X)$ that solve

$$\begin{aligned} \min_{\gamma(X) \in \mathbb{R}^{n_1}} \quad & \int Z\gamma(X)^2 \, dP_n \\ \text{subject to} \quad & \left\| \int Z\gamma(X)\phi(X) - \phi(X) \, dP_n \right\|_\infty \leq \lambda \quad \gamma(X) \geq 0, \end{aligned} \tag{2.7}$$

where P_n is the empirical distribution corresponding to a sample of size n from joint distribution $P(\cdot)$. These are the weights of minimum variance that guarantee *approximate balance*:

that the worst imbalance in ϕ, the transformed covariates, is less than some hyper-parameter λ. There are many other choices of both the penalty on the weights and the measure of imbalance.[2] For instance, in low dimensions, setting $\lambda = 0$ guarantees *exact balance* on the covariates $\phi(X_i)$. Here we focus on the more common case in which achieving exact balance is infeasible; in that case, the particular choice of penalty function is less important.

The balancing weights procedure is connected to the modeled IPW approach above through the Lagrangian dual formulation of optimization problem (2.7). The imbalance in the d transformations of the covariates induces a set of Lagrange multipliers $\beta \in \mathbb{R}^d$, and the Lagrangian dual is

$$\min_{\beta \in \mathbb{R}^d} \underbrace{\int Z\left[\beta \cdot \phi(X)\right]_+^2 - \beta \cdot \phi(X)\, dP_n}_{\text{balancing loss}} + \underbrace{\lambda \|\beta\|_1}_{\text{regularization}}, \tag{2.8}$$

where $[x]_+ = \max\{0, x\}$. The weights are recovered from the dual solution as $\hat{\gamma}(X_i) = \left[\hat{\beta} \cdot \phi(X_i)\right]_+$. As Zhao (2019) and Wang and Zubizarreta (2019) show, this is a regularized M-estimator of the propensity score when it is of the form $\frac{1}{\pi(x)} = [\beta^* \cdot \phi(x)]_+$ for some true β^*. Therefore, we can view $\beta^* \cdot \phi(x)$ as a natural parameter for the propensity score; different penalty functions will induce different link functions, see Wang and Zubizarreta (2019). Similarly, different measures of balance will induce different forms of *regularization* on the propensity score parameters. In the succeeding sections, we will use this dual connection to show that the percentile bootstrap sensitivity procedure proposed by Zhao et al. (2019) for traditional IPW estimators in the marginal sensitivity model is valid with balancing weights estimators.

2.3 Sensitivity analysis for balancing weights estimators

We now outline our procedure for extending the percentile bootstrap sensitivity analysis to balancing weights. We introduce the shifted balancing weights estimator, detail the bootstrap sampling procedure, and describe how to efficiently compute the confidence intervals. Key to constructing the confidence intervals for the partial identification set will be to construct intervals for each sensitivity model h in the collection of sensitivity models $\mathcal{H}(\Lambda)$ in Equation (2.3). Each h represents a particular deviation from ignorability that remains in the set defined by the marginal sensitivity model. We show that the percentile bootstrap yields valid confidence intervals for each sensitivity model in $\mathcal{H}(\Lambda)$, resulting in a valid interval for the partial identification set. While the procedure for constructing confidence intervals given the weights computed in each bootstrap sample is the same as that in Zhao et al. (2019),

[2] Other possibilities include soft balance penalties rather than hard constraints (e.g. Ben-Michael et al., 2020; Keele et al., 2020) and non-parametric measures of balance (e.g. Hirshberg et al., 2019).

our result allows for the weights to be constructed by more general methods. We provide guidance for interpreting our sensitivity analysis procedure in Section 2.4.

To construct the confidence intervals, we first consider the case where we know the log odds function $h(x, y) \in \mathcal{H}(\Lambda)$. With h, we can shift the balancing weights estimator for the shifted estimand $\mu_1^{(h)}$ as

$$\hat{\mu}_1^{(h)} = \left(\sum_{Z_i=1} \hat{\gamma}^{(h)}(X_i, Y_i(1)) \right)^{-1} \sum_{Z_i=1} \hat{\gamma}^{(h)}(X_i, Y_i(1)) Y_i, \tag{2.9}$$

where $\hat{\gamma}^{(h)}(X_i, Y_i(1)) = 1 + (\hat{\gamma}(X_i) - 1)e^{h(X_i, Y_i(1))}$ for $i \in \{i : Z_i = 1\}$ are the shifted balancing weights. Note that there is no requirement for the shifted balancing weights to balance the transformed covariates ϕ. We then take B bootstrap samples of size n without conditioning on treatment assignment — so the number of units in the treatment and control groups may vary from sample to sample — and re-estimate the weights in each sample by solving the balancing weights optimization problem (2.7) using the bootstrapped data.

Then, for every $h \in H(\Lambda)$, we can construct a confidence interval for $\mu_1^{(h)}$ using the percentile bootstrap as

$$\left[L^{(h)}, U^{(h)} \right] = \left[Q_{\frac{\alpha}{2}} \left(\hat{\mu}_{1,b}^{*(h)} \right), Q_{1-\frac{\alpha}{2}} \left(\hat{\mu}_{1,b}^{*(h)} \right) \right]. \tag{2.10}$$

$Q_\alpha(\hat{\mu}_{1,b}^{*(h)})$ is the α-percentile of $\hat{\mu}_{1,b}^{*(h)}$ in the bootstrap distribution made up of the B bootstrap samples and $\hat{\mu}_{1,b}^{*(h)}$ is the shifted balancing weights estimator (2.9) using bootstrap sample $b \in \{1, \ldots, B\}$. Note, the * in $\hat{\mu}_{1,b}^{*(h)}$ indicates that it is an estimate from bootstrap data and b is used as an index for the B bootstrap samples. The following theorem states that $[L^{(h)}, U^{(h)}]$ is an asymptotically valid confidence interval for $\mu_1^{(h)}$ with at least $(1-\alpha)$-coverage under high-level assumptions in Appendix A.1 on how well the balancing weights estimate the propensity scores.

Theorem 2.1. Under Assumption A.1 in Appendix A.1, for every $h \in H(\Lambda)$,

$$\limsup_{n \to \infty} \mathbb{P}_0(\mu_1^{(h)} < L^{(h)}) \leq \frac{\alpha}{2}$$

and

$$\limsup_{n \to \infty} \mathbb{P}_0(\mu_1^{(h)} > U^{(h)}) \leq \frac{\alpha}{2},$$

where $\mathbb{P}_0$ denotes the probability under the joint distribution of the data $P(\cdot)$. The probability statements apply under both the conditions on the inverse probabilities and the outcomes in Assumption A.1 and the marginal sensitivity model (2.3).

Since each of the confidence intervals $[L^{(h)}, U^{(h)}]$ are valid, we can use the Union Method to combine them into a single valid confidence interval $[L^{\text{union}}, U^{\text{union}}]$ for μ_1 under Assumption 2.3, where

$$L^{\text{union}} = \inf_{h \in \mathcal{H}(\Lambda)} L^{(h)}, \quad U^{\text{union}} = \sup_{h \in \mathcal{H}(\Lambda)} U^{(h)}. \tag{2.11}$$

Finding $[L^{\text{union}}, U^{\text{union}}]$ would require conducting a grid search over the space of log-odds functions $\mathcal{H}(\Lambda)$ and computing percentile bootstrap confidence intervals at each point; this is computationally infeasible. Instead, we can obtain a confidence interval $[L, U]$ for μ_1 by using generalized minimax and maximin inequalities as

$$[L, U] = \left[Q_{\frac{\alpha}{2}} \left(\inf_{h \in \mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(h)} \right), Q_{1-\frac{\alpha}{2}} \left(\sup_{h \in \mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(h)} \right) \right]. \tag{2.12}$$

Zhao et al. (2019) show that this interval will be conservative, in the sense of being too wide, since $L \leq L^{\text{union}}$ and $U \geq U^{\text{union}}$. In fact, Dorn and Guo (2021) show this can be overly conservative; see Sections 2.5 and 2.6 for further discussion.

The extrema of the point estimates can be solved efficiently using Proposition 2 from Zhao et al. (2019) by the following linear fractional programming problem:

$$\begin{aligned} \min_{r \in \mathbb{R}^{n_1}} / \max \quad & \hat{\mu}_1^{(h)} = \frac{\sum_{i=1}^{n} Z_i \left(1 + r_i \left[\hat{\gamma}(X_i) - 1\right]\right) Y_i}{\sum_{i=1}^{n} Z_i \left(1 + r_i \left[\hat{\gamma}(X_i) - 1\right]\right)} \\ \text{subject to} \quad & r_i \in [\Lambda^{-1}, \Lambda], \text{ for all } i \in \{1, \dots, n\}, \end{aligned} \tag{2.13}$$

where $r_i = \text{OR}\{\pi(X_i), \pi(X_i, Y_i(1))\}$ are the decision variables. The procedure to obtain confidence interval $[L, U]$ is then:

Step 1. Obtain B bootstrap samples of the data of size n without conditioning on treatment assignment.

Step 2. For each bootstrap sample $b = 1, \dots, B$, re-estimate the weights and compute the extrema $\inf_{h \in \mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(h)}$ and $\sup_{h \in \mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(h)}$ under the collection of sensitivity models $\mathcal{H}(\Lambda)$ by solving (2.13).

Step 3. Obtain valid confidence intervals for sensitivity analysis:

$$L = Q_{\frac{\alpha}{2}} \left(\inf_{h \in \mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(h)} \right), \quad U = Q_{1-\frac{\alpha}{2}} \left(\sup_{h \in \mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(h)} \right). \tag{2.14}$$

Replacing $\hat{\gamma}(X_i)$ in Equation (2.13) with the inverse of propensity scores estimated by a generalized linear model recovers the procedure from Zhao et al. (2019). As in Zhao et al.

(2019), the added computational cost for additional values of Λ is minimal since they do not require a researcher to draw additional bootstrap samples nor re-estimate the weights.

Finally, a researcher must compute a sensitivity value for a given study; see Rosenbaum (2002) for extensive discussion. Suppose the confidence interval for PATE under ignorability ($\Lambda = 1$) does not contain zero, indicating a statistically significant effect. As Λ increases, allowing for stronger violations of ignorability, the confidence interval will widen and eventually cross zero. Of particular interest then is the minimum value of Λ for which the confidence interval contains zero; we denote this value as Λ^*.[3] Thus, we can interpret Λ^* as a necessary difference in the odds ratio between the probability of treatment with and without conditioning on the treated potential outcome for which we no longer observe a significant treatment effect. This represents the degree of confounding required to change a study's causal conclusions, with larger values of Λ^* representing more robust estimates.

Sensitivity analysis may also be useful in cases where the confidence interval under $\Lambda = 1$ is very small and includes zero, indicating no large effect in any direction or bioequivalence in the sense discussed by Brown et al. (1995). In this setting, a researcher may obtain a sensitivity value Λ^* by defining a minimal effect size $\iota > 0$ of practical interest and repeating the sensitivity analysis for larger and larger values of Λ until the confidence interval includes either $-\iota$ or ι, revealing the degree of confounding needed to mask a practically important effect. For examples of such sensitivity analyses, see Pimentel et al. (2015); Pimentel and Kelz (2020).

2.4 Amplifying, interpreting, and calibrating sensitivity parameters

In this section, we provide guidance for interpreting the main sensitivity parameter Λ^* by "amplifying" the sensitivity analyses into a constraint on the product of: (1) the level of remaining imbalance in confounders after weighting; and (2) the strength of the relationship between the confounders and the treated potential outcome.

In order for a confounder to bias causal effect estimates, it must be associated with both the treatment and the outcome. An "amplification" enhances a sensitivity analysis's interpretability by allowing a researcher to instead interpret the results of the sensitivity analysis in terms of two parameters: one controlling the confounder's relationship with the treatment and the other controlling its relationship with the outcome (Rosenbaum and Silber, 2009). Under the marginal sensitivity model in Assumption 2.3, the parameter Λ controls how far the propensity score conditioned on only observed covariates $\pi(x)$ can be from an oracle propensity score that includes the treated potential outcome $\pi(x, y)$. This odds ratio bound can be difficult to reason about in applied analyses. To aid interpretation,

[3] Similar to the robustness value with $q = 1$ from Cinelli and Hazlett (2020), researchers can also consider the minimum value of Λ for which the point estimate interval contains zero. The point estimate interval can be computed by solving (2.13) using the full observed data for a particular value of Λ.

we propose an amplification that expresses the results of our procedure in terms of the imbalance in confounders and the strength of the relationship between the confounders and the treated potential outcome.

For our amplification, we will use $U \in \mathbb{R}$ to represent a latent unmeasured confounding variable, standardized to have mean zero and variance 1.[4] We then consider a working model for the conditional expectation of the treated potential outcome, decomposing it into a term involving the observed covariates X and a linear term for the unmeasured confounder U:

$$\mathbb{E}[Y(1) \mid X = x, U = u] = f(x) + \beta_u \cdot u. \tag{2.15}$$

This model merely serves as a guide to interpretation, rather than being a true relationship that we are assuming in the primary causal analysis, and is in fact general. As one extreme case, we can consider a situation in which $f(x) = E[Y(1)]$ and the unmeasured confounder U is a standardized version of the treated potential outcome itself, $U = \frac{Y(1)-\mathbb{E}[Y(1)]}{\text{sd}(Y(1))}$; in this case β_u is simply equal to the standard deviation of $Y(1)$. More generally, if some of the variation in $Y(1)$ can be explained by observed covariates and by pure additive noise uncorrelated with treatment, β_u describes the amount of additional systematic variation contributed by unobserved confounders. Specifically, β_u is the difference in expected $Y(1)$ associated with a one-standard-deviation difference in U while holding covariates fixed. If one is concerned about multiple unobserved confounders, one may also view U as the one-dimensional function of these confounders that best explains the variance in $Y(1)$'s conditional expectation under model (2.15).

With this model in place, we can decompose the difference between the true expected value of treated potential outcomes μ_1 and the IPW estimand — i.e., the bias — into (i) the strength of the unmeasured confounder U in predicting $Y(1)$ beyond the observed covariates, β_u, and (ii) the imbalance in U, δ_u:

$$\mathbb{E}[Y(1)] - \mathbb{E}\left[\frac{ZY}{\pi(X)}\right] = \beta_u \cdot \underbrace{\left(\mathbb{E}\left[U\right] - \mathbb{E}\left[\frac{ZU}{\pi(X)}\right]\right)}_{\delta_u}.$$

Note that here we have used the property that $\mathbb{E}[f(X)] = \mathbb{E}[Zf(X)/\pi(X)]$ for all functions f.

Now, we can use the partial identification of μ_1 under the marginal sensitivity model in Assumption 2.3 to find upper and lower bounds for this product under the sensitivity value Λ^*,

$$\inf_{h \in \mathcal{H}(\Lambda^*)} \mu_1^{(h)} - \mathbb{E}\left[\frac{ZY}{\pi(X)}\right] \leq \beta_u \cdot \delta_u \leq \sup_{h \in \mathcal{H}(\Lambda^*)} \mu_1^{(h)} - \mathbb{E}\left[\frac{ZY}{\pi(X)}\right].$$

These are population-level bounds for the highest and lowest possible bias $\beta_u \cdot \delta_u$. To construct finite-sample versions of these bounds, we bound the bias as the maximum of the absolute

[4] Dorn and Guo (2021) similarly consider a general unobserved confounder U, of which $U = Y(1)$ is a special case.

values of the highest and lowest possible differences in the estimated values,

$$|\beta_u \cdot \delta_u| \leq \max\left\{\left|\inf_{h \in \mathcal{H}(\Lambda^*)} \hat{\mu}_1^{(h)} - \hat{\mu}_1\right|, \left|\sup_{h \in \mathcal{H}(\Lambda^*)} \hat{\mu}_1^{(h)} - \hat{\mu}_1\right|\right\}. \tag{2.16}$$

Recall that $\hat{\mu}_1$ (2.6) is a weighted average of treated units' outcomes using weights $\hat{\gamma}(X)$.

The constrained relationship between the β_u and δ_u allows us to reason about potential unobserved confounders. To understand this relationship, we compute a curve that maps the value of the bias to different combinations of δ_u and β_u for enhanced interpretation. For example, $(\delta_u, \beta_u) = (1.5, 2)$ and $(\delta_u, \beta_u) = (1, 3)$ are both consistent with a bias of 3. Reading off this curve allows the researchers to see that for an unmeasured confounder with any given strength in predicting the treated potential outcome beyond the observed covariates, there must be *at least* some level of imbalance after weighting to induce bias. To explain a given amount of unmeasured confounding bias, an unmeasured confounder strongly predictive of potential outcomes (after controlling for observed covariates) need only be mildly imbalanced after weighting. Conversely, an unmeasured confounder with weak predictive strength must be highly imbalanced even after the observed covariates are approximately balanced by the estimated weights. In Section 2.5, we illustrate our sensitivity analysis procedure and how our amplification can produce more interpretable results.

2.5 Numerical examples

We now illustrate the sensitivity analysis and amplification procedures using two real data examples. We consider the situation in which a researcher uses balancing weights to estimate the Population Average Treatment Effect on the Treated (PATT) of a treatment on an outcome of interest; see Appendix A.3 for an overview of the PATT in our setting. Based on domain knowledge, the researcher believes that the set of observed covariates includes most factors associated with the treatment assignment and the outcome, while leaving open the possibility that there remain relevant unobserved covariates.

To start, we compute Λ^*, which represents the confounding required to alter a study's causal conclusions. In order to compute Λ^*, we compute confidence intervals for a grid of values of Λ, starting with $\Lambda = 1$ and then considering larger values of Λ. If the confidence interval corresponding to $\Lambda = 1$ contains zero, then the effect estimate is not significant, even under ignorability. If the confidence interval for $\Lambda = 1$ does not contain zero, increasing the value of Λ causes the confidence intervals to widen and eventually cross zero for some value of Λ. We set Λ^* equal to the minimum value of Λ for which the confidence interval includes zero. Since the the percentile bootstrap procedure induces randomness, this value of Λ^* is computed with Monte Carlo error.

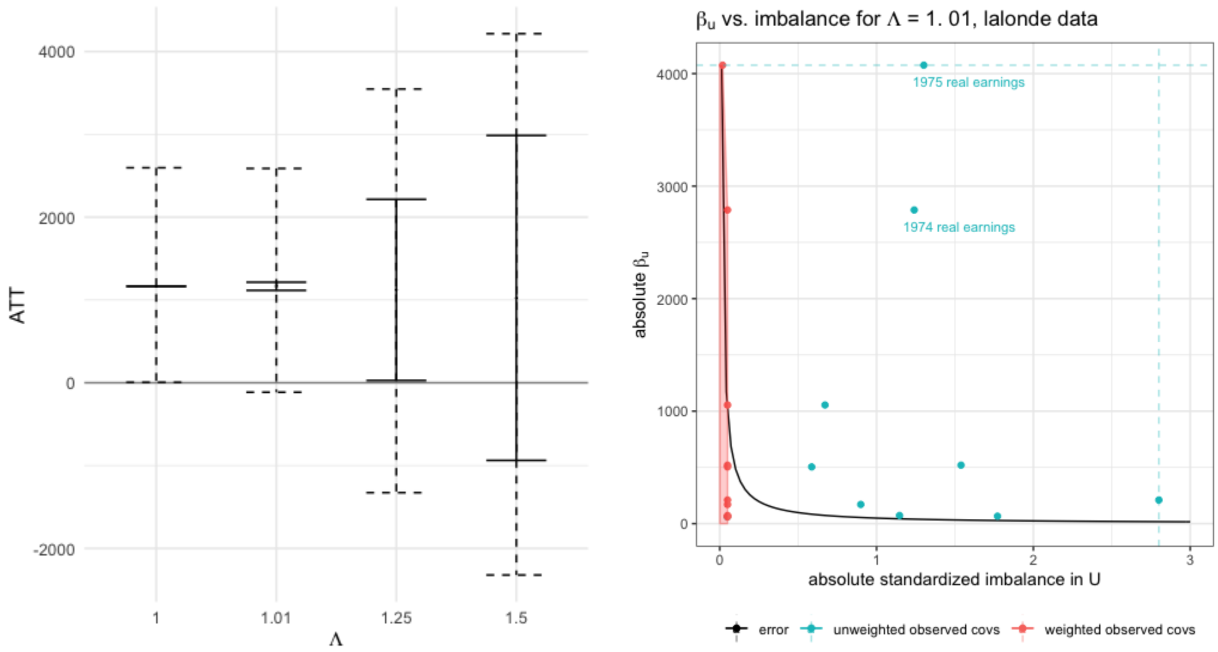
We fix the bias equal to the maximum absolute value of the upper and lower bounds on the bias in Equation (2.16). This value is the maximum absolute value of bias possible under the balancing weights sensitivity model with $\Lambda = \Lambda^*$ and is therefore a level of bias required to overturn the study's causal conclusion. We create contour plots with curves that map the particular value of bias to varying values of δ_u and β_u, allowing the bias to be alternatively interpreted in terms of two sensitivity analysis parameters. Veitch and Zaveri (2020) use the term "Austen plot" to describe similar plots. We include standardized observed covariates on the contour plots, which serve as guides for reasoning about potential unobserved covariates. Our proposed calibration process using observed covariates is intended to provide a broad sense of plausible parameter values, rather than an attempt to obtain precise estimates as a part of a formal benchmarking exercise. See Section 2.6 for further discussion. Blue points correspond to observed covariates with imbalance prior to weighting, while red points represent post-weighting imbalance. In the PATT setting, the imbalance prior to weighting in a standardized covariate X can be computed as $\frac{1}{\sum_{i=1}^{n} Z_i} \sum_{i=1}^{n} Z_i X_i - \frac{1}{\sum_{i=1}^{n}(1-Z_i)} \sum_{i=1}^{n}(1 - Z_i)X_i$, while the post-weighting imbalance is $\frac{1}{\sum_{i=1}^{n} Z_i} \sum_{i=1}^{n} Z_i X_i - \sum_{i=1}^{n} \frac{(1-Z_i)\hat{\gamma}(X_i)}{\sum_{i=1}^{n}(1-Z_i)\hat{\gamma}(X_i)} X_i$. We view the post-weighting imbalance corresponding to the red points as a best-case scenario for potential unobserved covariates — in general, we expect to achieve better balance in terms of the observed covariates that we directly target than unobserved covariates. Conversely, the pre-weighting imbalance represented by the blue points may be more in line with our expectations for unobserved covariates.

LaLonde job training experiment

We re-examine data analyzed by LaLonde (1986) from the National Supported Work Demonstration Program (NSW), a randomized job training program. Specifically, we use the subset of data from Dehejia and Wahba (1999) to form a treatment group and observational data from the Current Population Survey–Social Security Administration file (CPS1) to form a control group. We consider estimating the effect of the job training program on 1978 real earnings. The covariates for each individual include their age, years of education, race, marital status, whether or not they graduated high school, and earnings and employment status in 1974 and 1975. In total, there are 185 treated units and 15,992 control units.

First, we use stable balancing weights in Equation (2.7) to estimate $\widehat{\text{PATT}} = \$1,165$ (estimated with $\phi(x) = x$ and $\lambda = 0.05$), which is in line with Wang and Zubizarreta (2019)'s estimate using slightly different approximate balancing weights. We then compute $\Lambda^* = 1.01$, which indicates that even a slight difference between the estimated and oracle weights can render the PATT estimate statistically insignificant. Figure 2.1a shows how the range of point estimates and the 95% confidence interval widen as Λ increases, with the confidence interval including zero for Λ^*. The range of point estimates is obtained by computing the extrema of the point estimates for a particular Λ.

Figure 2.1b shows the contour plot for the LaLonde data, which adds concrete detail to our interpretation of Λ^*. The black contour line, representing all combinations of β_u and δ_u for which $\Lambda^* = 1.01$, lies below all of the blue points, suggesting that an unobserved



(a) Point estimate and confidence intervals.

(b) Contour plot illustrating amplification of the sensitivity analysis with comparison to observed variables.

Figure 2.1: Sensitivity analysis results with the LaLonde data
(a): Solid intervals are point estimate intervals and dotted intervals are 95% confidence intervals.
(b): Each location in the plot represents a possible unobserved confounder with parameters (δ_u, β_u) in the amplification. The contour line gives all such pairs that result in Λ equal to the observed sensitivity threshold $\Lambda^* = 1.01$. Plotted points represent observed covariates, with y-coordinates given by absolute multiple regression coefficients in an ordinary least-squares regression of the outcome on standardized covariates among the control group, equivalent to β_u if the covariate in question were the only omitted confounder, and with x-coordinates given by treated-control differences in standardized covariates both before weighting (these points are blue) and after weighting (these points are red). The red shaded region groups locations associated with unobserved confounders no stronger than the observed covariates after weighting, in the sense that some convex combination of post-weighting covariate locations is at least as far from the origin.

confounder similar even to one of the very weakest observed confounders would be sufficient to reverse the study results. Furthermore, the black contour line intersects the shaded red region containing post-weighting imbalance, suggesting that even closely-balanced variables like those explicitly accounted for in the weighting algorithm could be sufficient to explain the observed effect. All of this strongly substantiates the idea that our study result could be due to very mild unobserved confounding and should not be trusted as a reliable qualitative statement about the true impact of this job training program. In fact, since several red points lie above the contour line, our finding may even be plausibly explained by residual imbalance in these observed covariates after weighting, whether or not unobserved confounders are present.

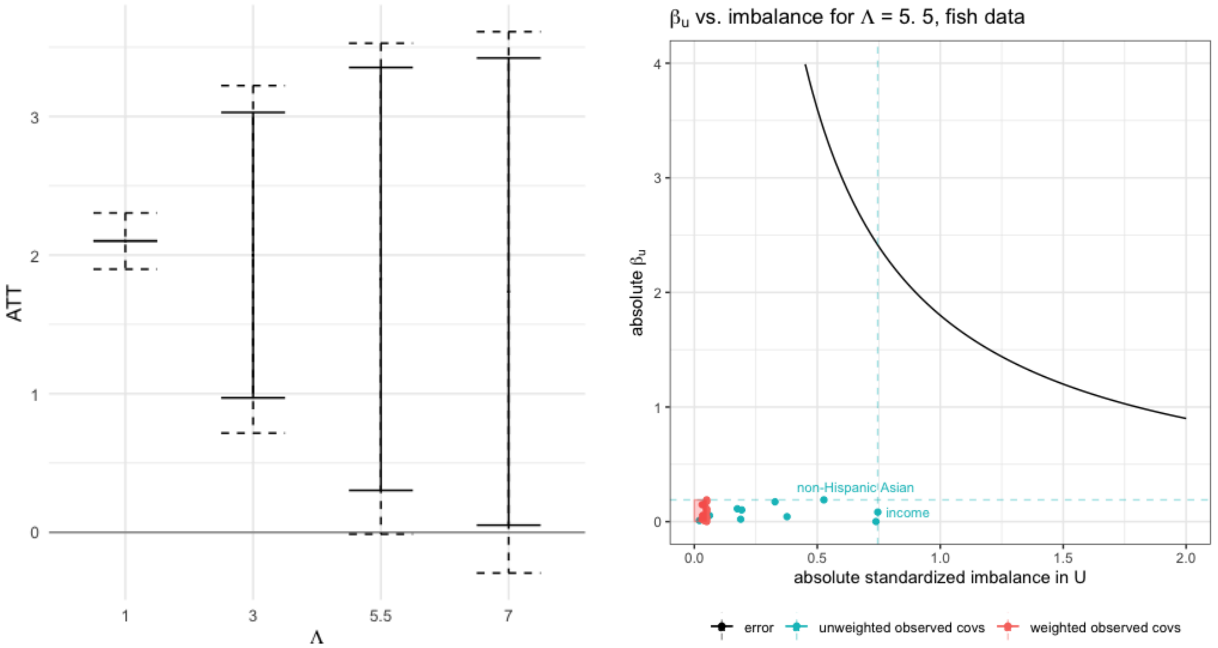
Note that visual comparisons of the curve with the blue points and the red region should never be taken at face value as binary statements about whether a study is robust to unmeasured confounding. Instead, one must always account for the context of the individual variables involved. For instance, the intersection of the curve with the red region occurs only in the upper region of the plot, because two of the variables, real earnings in 1974 and 1975 (both time-lagged versions of the study outcome), are highly correlated with the outcomes. It is not necessarily plausible that an unobserved confounder would exhibit such high outcome correlation, so intersection with the red region is perhaps less worrying than in a setting where all the observed variables are general demographic measures less directly tied to the observed outcome. In addition, it is important to include all potentially important observed covariates on the plot lest the red shaded region appear misleadingly small.

Fish consumption and blood mercury levels

We now examine data analyzed by Zhao et al. (2018) and Zhao et al. (2019) from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 containing information about fish consumption and blood mercury levels. We evaluate the sensitivity of estimating the effect of fish consumption on blood mercury levels using balancing weights. There are 234 treated units (consumption of greater than 12 servings of fish or shellfish in the past month) and 873 control units (zero or one servings). The outcome of interest is $\log_2$(total blood mercury), measured in micrograms per liter; the covariates include gender, age, income, whether income is missing and imputed, race/ethnicity, education, smoking history, and the number of cigarettes smoked in the previous month.

To start, the stable balancing weights (2.7) estimate of the PATT is an increase of 2.1 in $\log_2$(total blood mercury), estimated with $\phi(x) = x$ and $\lambda = 0.05$; Λ^* is approximately equal to 5.5 for the fish consumption data. We display the sensitivity analysis results for multiple values of Λ in Figure 2.2a. We observe that the confidence interval corresponding to no confounding ($\Lambda = 1$) is far from zero and that the confidence interval for $\Lambda^* = 5.5$ just begins to cross zero.

The contour plot (Figure 2.2b) for the fish data indicates that the causal effect estimate is robust to all but extremely strong unobserved confounders. Here the bias curve is far above the intersection of the dotted lines that represents the maximum strength and pre-weighting



(a) Point estimate and confidence intervals.

(b) Contour plot illustrating amplification of the sensitivity analysis, with comparison to observed variables.

Figure 2.2: Sensitivity analysis results with the fish diet data
(a): Solid intervals are point estimate intervals and dotted intervals are 95% confidence intervals.
(b): Each location in the plot represents a possible unobserved confounder with parameters (δ_u, β_u) in the amplification. The contour line gives all such pairs that result in Λ equal to the observed sensitivity threshold $\Lambda^* = 5.5$. Plotted points represent observed covariates, with y-coordinates given by absolute multiple regression coefficients in an ordinary least-squares regression of the outcome on standardized covariates among the control group, equivalent to β_u if the covariate in question were the only omitted confounder, and with x-coordinates given by treated-control differences in standardized covariates both before weighting (these points are blue) and after weighting (these points are red). The red shaded region groups locations associated with unobserved confounders no stronger than the observed covariates after weighting, in the sense that some convex combination of post-weighting covariate locations is at least as far from the origin.

imbalance among the observed covariates. Thus, confounding significantly stronger than the observed covariates would be required to alter the causal conclusion. In particular, consider the most imbalanced pre-treatment confounder, income. The large vertical gap between the associated blue dot (and indeed any of the blue dots) and the contour line suggests that an unobserved confounder sufficient to alter the study's conclusion would not only have to be as imbalanced as income prior to treatment, but would simultaneously have to be a full order of magnitude more predictive of blood mercury than any other variable measured in the study. In fact, in order to change the study's conclusion, an unmeasured confounder as imbalanced as income would have to have an approximately 29 times higher β_u than income. While the contour plot itself cannot rule out the possibility that such an unmeasured confounder might exist, it imposes stringent requirements for alternative theories behind the apparent causal effect.

The LaLonde data results in Figure 2.1 and the fish consumption data results in Figure 2.2 illustrate two extremes for possible outcomes of the sensitivity analysis. In our experience, more intermediate results frequently arise also; for example, the contour line might pass above some observed covariates but below others. In this case especially, it is important to remember that sensitivity analysis is not designed to provide a binary judgment about whether a study's effect is real or not; instead, the contour plot gives a sense for the types of unobserved confounder that might be problematic and the types that can be safely ignored.

Finally, in Figure 2.3 we compare the results of our sensitivity analysis in the fish consumption data to the results of the approaches described by Zhao et al. (2019) and Dorn and Guo (2021). As discussed above, Zhao et al. (2019) use IPW weights and otherwise conduct the sensitivity analysis in an identical manner. Dorn and Guo (2021) also use IPW weights but alter the sensitivity analysis by adding a constraint to the population version of the maximization problem in (2.13) that enforces balance on certain conditional quantiles of the observed outcomes. This is designed to ensure that that true propensity scores implied by the sensitivity model balance the observed data properly in large samples (the set of shifted balancing weights over which we take extrema need not do so). Figure 2.3 gives the expanded confidence intervals for the ATT from each approach at three values of Λ. All three approaches are qualitatively similar in each case. However, our approach based on stabilized balancing weights outperforms Zhao et al. (2019)'s IPW approach at each Λ-value investigated, achieving strictly shorter intervals. This suggests that the ability of balancing weights to achieve more precise inference than IPW in moderate samples, previously documented for settings with no unobserved confounding (Ben-Michael et al., 2021), seems to extend to sensitivity analysis as well. The approach of Dorn and Guo (2021) achieves narrower intervals than either of the other approaches; however, we note that Dorn and Guo (2021)'s added constraint relies on quantile regression and hence requires the outcome to be continuous, unlike the other two approaches. Additionally, the authors find that the quantile balancing confidence intervals can result in under-coverage when the quantiles are correctly specified, which could suggest a setting in which our proposed sensitivity analysis procedure's wider intervals could be advantageous. As such the combination of stabilized balancing weights and sensitivity analysis appears to offer an attractive mix of generality

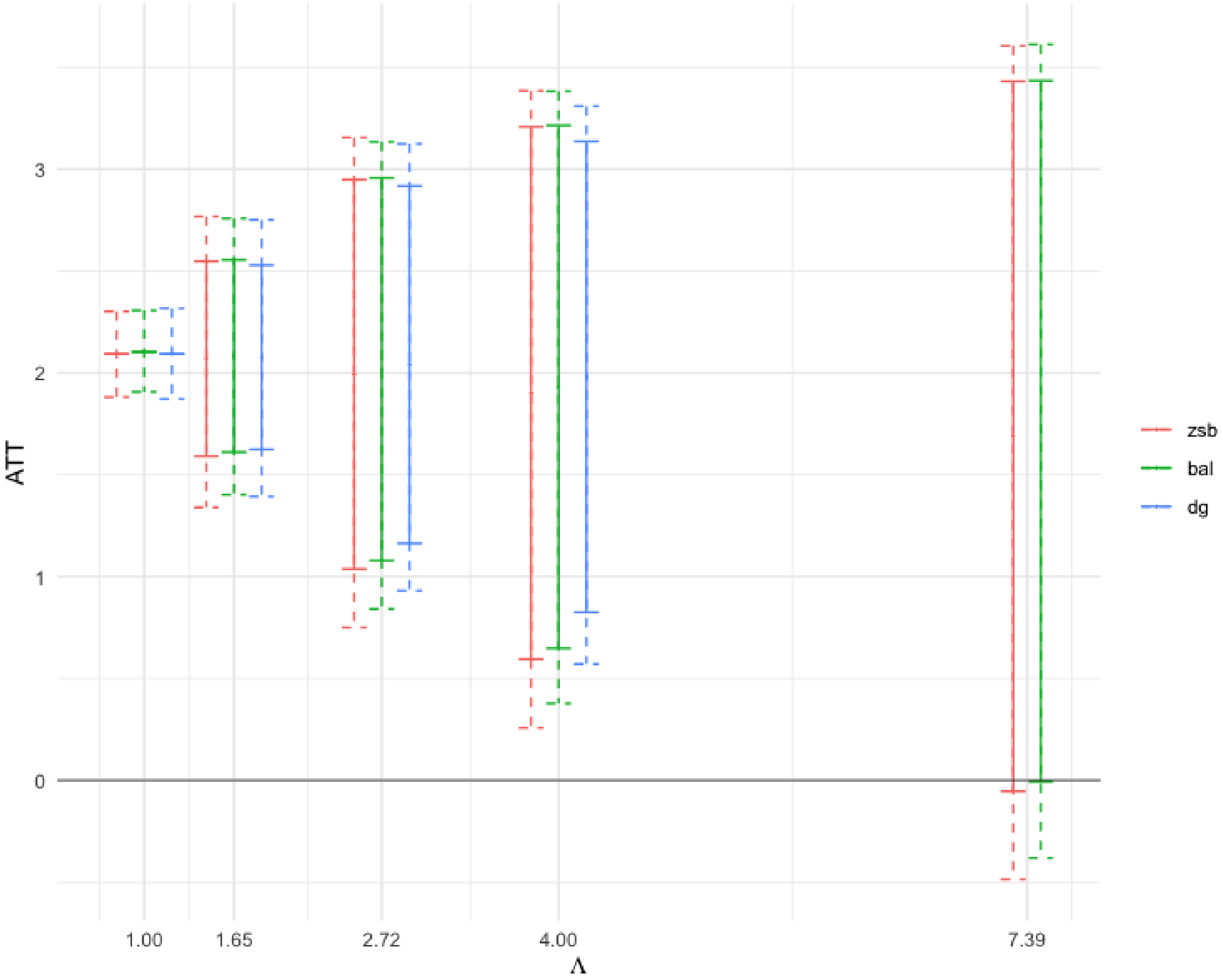


Figure 2.3: Comparison of confidence interval width after sensitivity analysis for three approaches in the fish consumption example. We compare the intervals constructed using stabilized weights followed by our proposed sensitivity analysis (labeled "bal") against those obtained by fitting IPW weights and conducting sensitivity analysis as described in Zhao et al. (2019) ("zsb" in the plot), and against those obtained by IPW and the approach of Dorn and Guo (2021) ("dg"), at several values of Λ. The Dorn & Guo bounds could not be computed at $\Lambda = 7.39$ due to numerical problems encountered in fitting the required quantile regression. All three approaches give similar results, but the balancing weights approach consistently outperforms Zhao et al. (2019)'s approach, while Dorn and Guo (2021)'s approach in turn produces narrower intervals than the stabilized weights approach for all values $\Lambda > 1$ investigated. Note that the results reported here for the Zhao et al. (2019) approach differ slightly from the results reported for their analysis of this dataset because we focus on the ATT rather than the ATE.

and precision compared to existing competitors.

2.6 Discussion

Balancing weights estimation is a popular approach for estimating treatment effects by weighting units to balance covariates. In this chapter, we develop a framework for assessing the sensitivity of these estimators to unmeasured confounding. We then propose an amplification for enhanced interpretation and illustrate our method through real data examples.

We briefly outline potential directions for future work. First, as discussed in Section 2.5, Dorn and Guo (2021) show that the intervals obtained from solving the linear programming problem (2.13) can be overly conservative, and resolve this issue by adding constraints that require balance on certain conditional quantiles of the outcome. It seems likely that such constraints would offer benefits for balancing weights estimators as well. We leave a thorough investigation to future work.

Second, we could extend our framework to include augmented balancing weights estimators, which use an outcome model to correct for bias due to inexact balance. Additionally, we could extend our sensitivity analysis framework to balancing weights in panel data settings. For example, we could adapt this framework to variants of the synthetic control method (Abadie and Gardeazabal, 2003; Ben-Michael et al., 2018), extending proposals for sensitivity analysis from Firpo and Possebom (2018).

Additionally, Cinelli and Hazlett (2020) point out that informal benchmarking procedures can be misleading if used to perform an exact calibration of sensitivity analysis parameters based on observed data. The authors argue that this occurs because the estimates of the observed covariates' relationships with the outcomes may be impacted by unmeasured confounding. They propose a formal benchmarking procedure to bound the strength of unmeasured confounders based on observed covariates. Adapting Cinelli and Hazlett (2020)'s formal benchmarking procedure to our setting could be a topic of future research.

Finally, we could use our framework to provide guidance in the design stage of balancing weights estimators. When estimating treatment effects using balancing weights, researchers must make decisions including the specific dispersion function of the weights, the particular imbalance measure, and, in many cases, an acceptable level of imbalance. We could extend our sensitivity analysis procedure to help make these decisions to improve robustness and power in the presence of unmeasured confounding. For example, we could provide insight into the trade-off between achieving better (marginal) balance on a few covariates or worse balance on a richer set of covariates.

Chapter 3

Design Sensitivity and Its Implications for Weighted Observational Studies

Sensitivity to unmeasured confounding is not typically a primary consideration in designing treated-control comparisons in observational studies. We introduce a framework allowing researchers to optimize robustness to omitted variable bias at the design stage using a measure called design sensitivity. Design sensitivity, which describes the asymptotic power of a sensitivity analysis, allows transparent assessment of the impact of different estimation strategies on sensitivity. We apply this general framework to two commonly-used sensitivity models, the marginal sensitivity model and the variance-based sensitivity model. By comparing design sensitivities, we interrogate how key features of weighted designs, including choices about trimming of weights and model augmentation, impact robustness to unmeasured confounding, and how these impacts may differ for the two different sensitivity models. We illustrate the proposed framework on a study examining drivers of support for the 2016 Colombian peace agreement.

3.1 Introduction

Increasingly, observational studies are being used to answer causal questions in the social and biomedical sciences. Estimating causal effects in observational settings often requires an assumption that unmeasured confounding is absent. In practice, this assumption is not testable and often untenable. Recent literature has introduced sensitivity analyses to assess the potential impact of an unobserved confounder on a study's results (Zhao et al., 2019; Soriano et al., 2021; Jin et al., 2022; Ishikawa and He, 2023). However, sensitivity analysis remains underutilized in practice, and sensitivity to unmeasured confounding is not typically a primary consideration in designing the treated-control comparison.

Given that findings from observational studies cannot be viewed as reliable unless a sensitivity analysis demonstrates some robustness to unmeasured confounding, it seems natural to approach the design of an observational study with unmeasured confounding in mind. We introduce a measure called *design sensitivity* for weighted observational studies. It describes the asymptotic power of a sensitivity analysis and can be computed prior to carrying out a study. Computing and comparing design sensitivities across possible weighted designs allows researchers to optimize for robustness to unmeasured confounding rather than treating sensitivity analysis as a post hoc secondary analysis. Under our framework, design sensitivity can be constructed for a wide variety of sensitivity models with only mild regularity conditions required.

We illustrate design sensitivity in two common sensitivity models: the marginal sensitivity model and the variance-based sensitivity model. By comparing design sensitivities, we interrogate how key features of weighted designs impact robustness to unmeasured confounding, and how these impacts can differ for the different sensitivity models.

3.2 Background

Set-Up, Notation, and Assumptions

We consider an observational study of n units sampled identically and independently from an infinite population. Define Z as a treatment indicator, where $Z = 1$ if a unit is in the treatment group, and 0 otherwise. Furthermore, define $Y(1)$ and $Y(0)$ as the potential outcomes under treatment and control, respectively. Throughout, we will make the stable unit treatment value assumption (SUTVA)–i.e., no interference or spillovers, such that the observed outcomes Y_i can be written as $Y = Y(1) \cdot Z + Y(0) \cdot (1 - Z)$ (Rubin, 1980a).

In randomized trials, researchers determine the probability of assignment to treatment or control for each individual, but in observational studies, propensities for treatment may co-vary with unobserved potential outcomes. To permit unbiased estimation of treatment effects, researchers must measure all background covariates describing common variation in treatment and potential outcomes, as formalized in the following assumption.

Assumption 3.1 (Conditional Ignorability of Treatment Assignment)**.** For some vector of pre-treatment covariates $\widetilde{X} \in \widetilde{\mathcal{X}}$:

$$Y(1), Y(0) \perp\!\!\!\perp Z \mid \widetilde{X}.$$

This assumption, also known as *selection on observables*, requires that given pre-treatment covariates $\widetilde{X}$, treatment is 'as-if' random. In addition to conditional ignorability, treatment effect estimation generally requires overlap, meaning that all units have a non-zero probability of being treated.

Assumption 3.2 (Overlap)**.** For all units $i \in 1, ..., n$ and any $x \in \widetilde{\mathcal{X}}$, $0 < \Pr(Z = 1 \mid \widetilde{X} = x) < 1$.

Our estimand of interest is the *average treatment effect across the treated* (i.e., ATT):

$$\tau := \mathbb{E}\left[Y(1) - Y(0) \mid Z = 1\right],$$

where the expectation is taken with respect to the population. The proposed framework can be extended for settings in which researchers are interested in the average treatment effect (ATE), as well as other common missingness settings such as external validity and survey non-response (Huang, 2022; Hartman and Huang, 2022).

A common approach to estimating the ATT is by using weighted estimators:

$$\hat{\tau}_W := \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Y_i Z_i - \frac{\sum_{i=1}^n \hat{w}_i Y_i (1 - Z_i)}{\sum_{i=1}^n \hat{w}_i (1 - Z_i)}. \tag{3.1}$$

The weights $\hat{w}_i$ are chosen so that the re-weighted distribution of pre-treatment covariates $\widetilde{X}$ across the control units matches the distribution of $\widetilde{X}$ across the treated units; for example, the population-level inverse propensity score weights guarantee this. Weights must be estimated. A common approach is to fit a propensity score model to estimate a unit's probability of treatment given the pre-treatment covariates and use the fitted values to construct estimated inverse weights. An alternative approach uses balancing weights, which solve an optimization problem that selects weights to balance sample moments, thereby bypassing a need to estimate a propensity model. For examples of such balancing weights and additional methods and theory, see Hainmueller (2012); Zubizarreta (2015); Ben-Michael et al. (2021). As shown by Chattopadhyay and Zubizarreta (2021), certain regression estimators can also be represented in the form of weighting estimators if weights are allowed to take on negative values. We note that the theoretical framework introduced in Section 3.3 can be easily extended to accommodate negative weights, though we focus on more familiar positive-weights settings for ease of exposition.

When the full set of covariates $\widetilde{X}$ in Assumption 3.1 are observed and the weights are correctly specified, the weighted estimator is consistent and unbiased for the ATT. However, in practice, it is impossible to know whether or not all confounders have been measured. Omitted confounders lead to biased estimates. In what follows, we consider the setting in which the full vector of covariates is defined as $\widetilde{X} := \{X, U\}$, where $X \in \mathcal{X}$ is observed and measured across all units, but $U \in \mathcal{U}$ is unobserved. As such, the estimated weights $\hat{w}_i$ are functions of X alone. We assume that the estimated weights $\hat{w}$ converge in probability to the population weights $w := \Pr(Z = 1 \mid X) / \Pr(Z = 0 \mid X)$ that condition on X alone, and define τ_W as the large-sample probability limit of $\widehat{\tau}_W$. We define the ideal weights $w^* := \Pr(Z = 1 \mid X, U) / \Pr(Z = 0 \mid X, U)$ as the population-level inverse propensity score weights in (X, U). Were researchers to use the ideal weights w^*, they would consistently recover the ATT.

We note that the framework just presented accommodates balancing weights that converge asymptotically to inverse propensity score weights (Ben-Michael et al., 2021). Researchers may also relax the assumption of correct specification in settings where specification concerns can be formulated as an omitted variable problem (Huang, 2022; Hartman and Huang, 2022).

Review: Sensitivity Analyses for Weighted Estimators

Sensitivity analyses allow researchers to assess their study findings' robustness to varying degrees of violation of underlying assumptions. We define a *sensitivity model* as a set of ideal weight vectors w^* over which we are interested in conducting a worst-case analysis.

Typically a sensitivity model consists of all w^* within a local neighborhood of the population weights w based on $\mathbf{X}$ alone, where the neighborhood is described by a specified error structure and indexed by a parameter that can be chosen to make the neighborhood larger or smaller.

The ideal weights w^* give a mapping $w^*(x,u)$ from $\widetilde{\mathcal{X}}$ to a real-valued weight. Formally a sensitivity model $\nu(\Gamma, w)$ is a set of such mappings $\{w^*(x,u) : f_\nu(w^*(X,U), w(X)) \leq \Gamma\}$, where the function f_ν measures dissimilarity between two probability distributions (in this case, the distributions induced by the random covariate vector $\mathcal{X}$ under $w(X)$ and $w^*(X,U)$). $\Gamma \in \mathbb{R}$ is a parameter constraining the overall dissimilarity allowed. Larger Γ values allow for larger deviations from w and hence more unobserved confounding. For any particular choice of ν and Γ, we define the interval of possible values for the true ATT:

$$\left[L_{\nu(\Gamma,w)}, U_{\nu(\Gamma,w)}\right] := \left[\inf_{\tilde{w}\in\nu(\Gamma,w)} \tau(\tilde{w}), \sup_{\tilde{w}\in\nu(\Gamma,w)} \tau(\tilde{w})\right]. \tag{3.2}$$

Following Zhao et al. (2019), we will refer to interval (3.2) as the *partially identified region.*

When assessing whether unobserved confounding is sufficiently strong to overturn a research conclusion, there are two sources of error for which we must account. The interval $\left[L_{\nu(\Gamma,w)}, U_{\nu(\Gamma,w)}\right]$ describes the first source of error, the bias arising from the omitted confounders. With an infinite number of samples from the population, this interval would be known and would represent the only source of error. However, in practice, we work with estimated weights $\hat{w}$ instead of the population weights w, which produce a noisy approximation $\left[\hat{L}_{\nu(\Gamma,w)}, \hat{U}_{\nu(\Gamma,w)}\right]$ to the partially identified region and the resulting sampling variability provides another source of error. Therefore, it is typically necessary to construct a bias-aware confidence interval $CI_{\nu(\Gamma,w)}(\alpha) \supseteq \left[\hat{L}_{\nu(\Gamma,w)}, \hat{U}_{\nu(\Gamma,w)}\right]$ that contains any true parameter $\tau(\widetilde{w}) \in \left[L_{\nu(\Gamma,w)}, U_{\nu(\Gamma,w)}\right]$ with probability at least $1-\alpha$.

For a study with a nominally significant result, a *sensitivity analysis* is conducted by searching over values of Γ, repeating the test for the hardest-to-reject value of w^* in $\nu(\Gamma, w)$, and finding the largest value Γ^* for which it is still possible to reject the null. If Γ^* is small, then the initial finding is sensitive to a small amount of unobserved confounding. If Γ^* is large, only a strong unobserved confounder could explain the results under a true null hypothesis.

Many different approaches to constructing sensitivity models using different specifications of f_ν have been proposed, including restricting various L^p-norms of the ratio w^*/w and its image under convex functions (e.g., Zhao et al., 2019; Zhang and Zhao, 2022; Huang, 2022; Jin et al., 2022). Given the rich and developing literature on different sensitivity models, one key contribution of our proposed method is the flexibility to be applied to *any* sensitivity model that meets a set of relatively weak regularity conditions. However, for illustrative purposes, we will discuss two common sensitivity models: the variance-based sensitivity model of Huang and Pimentel (2022) and the marginal sensitivity model of Tan (2006) and Zhao et al. (2019).

Example: The Variance-based Sensitivity Model

The variance-based sensitivity model $\nu_{vbm}(R^2, w)$ constrains the variance in w^* not explained by w (Huang and Pimentel, 2022). More formally, for some $R^2 \in [0, 1)$, the variance-based sensitivity model is defined as follows:

$$\nu_{vbm}(R^2, w) := \left\{ w^* : 1 \leq \frac{\text{var}(w^* \mid Z = 0)}{\text{var}(w \mid Z = 0)} \leq \frac{1}{1 - R^2} \right\}. \tag{3.3}$$

The optimal bias bound for a set $\nu_{vbm}(R^2, w)$ is defined as:

$$\begin{aligned} &\max_{\tilde{w} \in \nu_{vbm}(R^2, w)} \text{Bias}(\hat{\tau}_W \mid \tilde{w}) \\ &= \sqrt{1 - \text{cor}(w, Y \mid Z = 0)^2} \sqrt{\frac{R^2}{1 - R^2} \text{var}(Y \mid Z = 0) \text{var}(w \mid Z = 0)}. \end{aligned} \tag{3.4}$$

For a fixed R^2 value, researchers can estimate the other quantities in Equation (3.4) using observed sample analogues. The optimal bias bound then defines the range of potential point estimates $\left[L_{\nu_{vbm}(R^2,w)}, U_{\nu_{vbm}(R^2,w)}\right]$ as:

$$\left[\tau_W - \max_{\tilde{w} \in \nu_{vbm}(R^2, w)} \text{Bias}(\tau_W \mid \tilde{w}), \quad \tau_W + \max_{\tilde{w} \in \nu_{vbm}(R^2, w)} \text{Bias}(\tau_W \mid \tilde{w}) \right],$$

where the maximum bias is directly calculated using Equation (3.4). In essence, the variance-based sensitivity model constrains a weighted L_2 distance between the ideal weights w^* and the weights w (Huang and Pimentel, 2022), so it is especially relevant in settings in which researchers are comfortable reasoning about the average degree of unobserved confounding across subjects. A percentile bootstrap approach proposed originally in Zhao et al. (2019) is used in concert with the bias bound to account for sampling variability, creating confidence intervals for the ATT that remain valid even in the presence of confounding under the sensitivity model.

Example: The Marginal Sensitivity Model

The marginal sensitivity model $\nu_{msm}(\Lambda, w)$ constrains the worst-case error from omitting a confounder, positing that the ratio between any ideal weight w^* and corresponding w may not exceed $\Lambda \geq 1$ (Tan, 2006; Zhao et al., 2019).

$$\nu_{msm}(\Lambda, w) := \left\{ w^* : \Lambda^{-1} \leq \frac{w^*}{w} \leq \Lambda \right\}. \tag{3.5}$$

The extrema $[L_{\nu_{msm}(\Lambda,w)}, U_{\nu_{msm}(\Lambda,w)}]$ can be computed using linear programming (Zhao et al., 2019). In contrast to the variance-based sensitivity model, the marginal sensitivity model is most useful when the researcher is comfortable reasoning about the maximal degree of confounding for any given subject. The percentile bootstrap approach of Zhao et al. (2019) is again used to account for sampling variability. Dorn and Guo (2021) introduced alternative approaches to obtain sharp limiting sets of point estimates under the marginal sensitivity model; for further discussion of the implications of this work in our context, see Section 3.6.

From Sensitivity Analysis to Design Sensitivity

Sensitivity analyses provide valuable information and recent innovations have improved their interpretability and utility (Ding and VanderWeele, 2016; Cinelli and Hazlett, 2020; Soriano et al., 2021). However, they are underutilized in practice (VanderWeele and Ding, 2017; Hazlett and Parente, 2023). One fundamental drawback is that sensitivity analysis is conducted post-hoc, as a secondary analysis. If an estimated result is found to be easily overturned by a relatively weak confounder, there is little that the researcher can do. Returning to the analysis and altering the estimation approach to mitigate sensitivity to an omitted confounder after conducting sensitivity analysis may introduce bias; this practice violates the 'design principle,' which forbids consultation of in-sample outcomes during study design (Rubin, 2007).

We now provide a design tool, *design sensitivity* for weighted estimators, that enables researchers to plan ahead and tailor studies to improve robustness to unmeasured confounding. In brief, design sensitivity characterizes the power of a sensitivity analysis in large samples. Comparing design sensitivities across different estimation approaches and study specifications provides insight into the implications of those choices for robustness to unmeasured bias, much as power calculations provide insight into design choices' impacts on precision in randomized studies. Design sensitivity in this formal sense was introduced by Rosenbaum (2004) and has been explored extensively for matched studies (Heller et al., 2009; Rosenbaum et al., 2010; Hsu et al., 2013). Implications include the value of reducing heterogeneity within matched pairs (Rosenbaum, 2005) and choosing test statistics (Rosenbaum, 2011; Howard and Pimentel, 2021) and treatment doses (Rosenbaum, 2004) carefully. We construct design sensitivity for weighting estimators and explore its implications for important design choices about which types of weights to construct and whether to incorporate outcome modeling into the analysis.

3.3 Design Sensitivity for Weighted Estimators

Power of a Sensitivity Analysis and Design Sensitivity

Design sensitivity is closely related to familiar notions of statistical power. The power of a test is the probability of rejecting a null hypothesis when a specific alternative is instead true; the alternative is typically chosen to reflect a "favorable" situation in which the effect of interest is present and important identifying assumptions are met (Rosenbaum, 2010, §15). The power of a sensitivity analysis, for a null hypothesis of no treatment effect, a given sensitivity model $\nu(\Gamma, w)$, and a value Γ_0 of the sensitivity parameter, is the probability that $CI_{\nu(\Gamma_0,w)}$ excludes zero under a *favorable* situation, meaning (1) the study is free of unmeasured bias, and in addition (2) a null hypothesis of no treatment effect is false, with potential outcomes generated by a specific stochastic model. While it may seem paradoxical to evaluate a sensitivity analysis when no unmeasured confounding is present, in practice, researchers do not know whether unmeasured confounding is present and must conduct a sensitivity analysis anyway. Power helps determine when true effects can be detected, even when the test allows for some confounding. Power depends on the potential outcome distribution; as a result, researchers must specify a hypothesized model for the potential outcomes and treatment effects in order to compute power in advance of data analysis. When outcome data is available from a planning sample, it can be used to calibrate the hypothetical outcome distribution (see Appendix B.2).

We now provide a convenient normal approximation to the power of a sensitivity analysis. Throughout the chapter, we will assume without loss of generality that the specified alternative has a positive treatment effect ($\tau > 0$).

Theorem 3.1 (Power of a Sensitivity Analysis)**.** Let $\hat{\tau}_W$ be a standard weighted estimator (i.e., Equation (3.1)), and for a sensitivity model $\nu(\Gamma, w)$ let $\tau_{\nu(\Gamma,w)} := \inf_{\tilde{w}\in\nu(\Gamma,w)} \tau(\tilde{w})$ and $\xi_{\nu(\Gamma,w)} := \tau_W - \tau_{\nu(\Gamma,w)}$. Finally, let $k_\alpha := 1 - \Phi(\alpha)$. Then, under standard regularity conditions (see Assumption B.1), the power of a sensitivity analysis is given as follows:

$$\begin{aligned}\Pr\left(\frac{\sqrt{n}(\hat{\tau}_W - \xi_{\nu(\Gamma,w)})}{\sigma_{\nu(\Gamma,w)}} \geq k_\alpha\right) &= \Pr\left(\frac{\sqrt{n}(\hat{\tau}_W - \tau_W)}{\sigma_W} \geq \frac{k_\alpha \sigma_{\nu(\Gamma,w)} + \sqrt{n}(\xi_{\nu(\Gamma,w)} - \tau_W)}{\sigma_W}\right) \\ &\simeq 1 - \Phi\left(\frac{k_\alpha \sigma_{\nu(\Gamma,w)} + \sqrt{n}(\xi_{\nu(\Gamma,w)} - \tau_W)}{\sigma_W}\right), \qquad (3.6)\end{aligned}$$

where σ_W and $\sigma_{\nu(\Gamma,w)}$ represent the variance of τ_W and $\tau_{\nu(\Gamma,w)}$ respectively.

The $\simeq$ symbol indicates asymptotic equivalence, meaning that the quantities on the two sides converge to the same limit as $n \to \infty$. Expression (3.6) reveals helpful patterns for large n. The numerator of the fractional term includes a variance term that is stable across sample sizes, as well as a bias term that grows with n and will dominate the formula in large samples. The sign of the bias term $\xi_{\nu(\Gamma,w)} - \tau_W$ is thus highly consequential, determining whether asymptotic power will be very large or very small. For a given distribution of weights w, increasing Γ eventually leads to a shift from a high-power to a low-power regime. Design sensitivity characterizes this important phase transition.

Theorem 3.2 (Design Sensitivity)**.** For a sensitivity model $\nu(\Gamma, w)$ where $\sigma_{\nu(\Gamma,w)} < \infty$ and a given favorable situation, let the *design sensitivity* be any value $\widetilde{\Gamma}$ such that the following two conditions hold, where $\beta_{\nu(\Gamma,w)}$ is the power of the sensitivity analysis:

$$\beta_{\nu(\Gamma,w)} \longrightarrow 1 \quad \forall\, \Gamma < \widetilde{\Gamma} \qquad \text{and} \qquad \beta_{\nu(\Gamma,w)} \longrightarrow 0 \quad \forall\, \Gamma > \widetilde{\Gamma}.$$

Then $\widetilde{\Gamma}$ is given by solving the estimating equation $\xi_{\nu(\Gamma,w)} - \tau_W = 0$ for Γ.

The theorem is a direct consequence of Equation (3.6). In short, the design sensitivity delineates the maximum amount of unobserved confounding under which the effect can still be detected given a sufficiently large sample. It is often a more useful quantity than the power, which must typically be calculated separately for many values of Γ (since a single compelling value of Γ cannot often be identified in advance); design sensitivity does not require either a Γ-value or sample size to be specified. Through the $\xi_{\nu(\Gamma,w)}$ and τ_W terms, design sensitivity depends on the weights selected by the researcher, and evaluating design sensitivities provides a natural basis on which to compare different design specifications. As will be shown, analytical solutions for the design sensitivity in specific sensitivity models also highlight general principles about the factors that most influence a study's robustness to unobserved confounding.

Design Sensitivity in the Variance-based Sensitivity Model

In the variance-based sensitivity model, the bias term $\xi_{\nu_{vbm}(R^2,w)}$ defined in Theorem 3.1 is equal to the optimal bias bound from Equation (3.4). This follows immediately from the fact that $\xi_{\nu(\Gamma,w)}$ is generally defined as the difference between the estimates τ_W and the minimum value (i.e., $L_{\nu(\Gamma,w)}$). As such, $\xi_{\nu_{vbm}(R^2,w)} := \tau_W - L_{\nu_{vbm}(R^2,w)} = \tau_W - \big(\tau_W - \max_{\tilde{w}\in\nu_{vbm}(R^2,w)} \text{Bias}(\tau_W \mid \tilde{w})\big) = \max_{\tilde{w}\in\nu_{vbm}(R^2,w)} \text{Bias}(\tau_W \mid \tilde{w})$. Using Equation (3.4), we derive a closed form for the design sensitivity $\widetilde{R}^2$.

Theorem 3.3 (Design Sensitivity for Variance-Based Models)**.** Define $\tilde{R}^2$ as

$$\tilde{R}^2 := \frac{a^2}{1+a^2} \quad \text{where } a^2 = \frac{1}{1-\text{cor}(w, Y \mid Z=0)^2} \cdot \frac{\tau_W^2}{\text{var}(w \mid Z=0)\cdot \text{var}(Y \mid Z=0)},$$

where covariances and variances are computed under the favorable situation. Then, under mild regularity conditions (Assumption B.1), $\widetilde{R}^2$ is the design sensitivity, i.e.

$$\begin{aligned} &\Pr\Big(0 \notin \big[L_{\nu_{vbm}(R^2,w)}, U_{\nu_{vbm}(R^2,w)}\big]\Big) \to 1 \text{ as } n \to \infty \text{ for } R^2 < \tilde{R}^2 \\ \text{and} \quad &\Pr\Big(0 \notin \big[L_{\nu_{vbm}(R^2,w)}, U_{\nu_{vbm}(R^2,w)}\big]\Big) \to 0 \text{ as } n \to \infty \text{ for } R^2 > \tilde{R}^2. \end{aligned}$$

Theorem 3.3 highlights the drivers of design sensitivity of the variance-based sensitivity model. In particular, the variance of the estimated weights (i.e., $\text{var}(w \mid Z=0)$), the variance of the outcomes (i.e., $\text{var}(Y \mid Z=0)$), the correlation between the estimated weights and the outcomes (i.e., $\text{cor}(w, Y \mid Z=0)$), and the effect size will affect the size of the design sensitivity.

Theorem 3.3 suggests that design decisions that reduce the variance in the estimated weights or the variance in the outcomes, or increase the association between the weights and the outcomes, can help increase the design sensitivity, and by extension, improve robustness to unobserved confounding. Examples of such decisions include trimming and augmentation. It is important to note that not all of these design choices guarantee an improvement in design sensitivity because they may affect all three of the components highlighted above. The specific impact of each decision requires numerical assessment, and Section 3.4 uses design sensitivity calculations to determine when we expect improvements to design sensitivity under each design choice.

Design Sensitivity in the Marginal Sensitivity Model

In the marginal sensitivity model, unlike the variance-based sensitivity model, design sensitivity does not have a closed form. We characterize it via an estimating equation.

Theorem 3.4 (Design Sensitivity for the Marginal Sensitivity Model)**.** Define $\tilde{\Lambda}$ as any solution to the following estimating equation (where $F_{y|x}$ represents the population cdf of y given x under the favorable situation):

$$\begin{aligned}&\mathbb{E}[wY(0) \mid Z=0] + \tau \\ &\quad = \sup_{\theta \in [0,1]} \frac{\Lambda \mathbb{E}\left[wY(0)G_\theta(Y(0)) \mid Z=0\right] + \frac{1}{\Lambda}\mathbb{E}\left[wY(0)(1-G_\theta(Y(0))) \mid Z=0\right]}{\Lambda \mathbb{E}\left[wG_\theta(Y(0)) \mid Z=0\right] + \frac{1}{\Lambda}\mathbb{E}\left[w(1-G_\theta(Y(0))) \mid Z=0\right]},\end{aligned}$$

where $G_\theta(Y(0)) = \mathbb{1}\{Y(0) \geq F^{-1}_{Y(0)|Z=0}(1-\theta)\}$.

Then under mild regularity assumptions $\widetilde{\Lambda}$ is the design sensitivity, i.e.

$$\begin{aligned}&\Pr\left(0 \notin \left[L_{\nu_{msm}(\Lambda,w)}, U_{\nu_{msm}(\Lambda,w)}\right]\right) \to 1 \text{ as } n \to \infty \text{ for } \Lambda < \tilde{\Lambda} \\ \text{and} \quad &\Pr\left(0 \notin \left[L_{\nu_{msm}(\Lambda,w)}, U_{\nu_{msm}(\Lambda,w)}\right]\right) \to 0 \text{ as } n \to \infty \text{ for } \Lambda > \tilde{\Lambda}.\end{aligned}$$

The design sensitivity under the set of marginal sensitivity models is a function of the joint cumulative density function of the weights and the control outcomes, and depends on an optimal cutoff θ that determines the maximum bias that can occur for a fixed Λ. As shown by Zhao et al. (2019), the worst-case setting for the marginal sensitivity model requires observations with the smallest $Y(0)$ values to have the smallest weights possible under the model (scaling observed weights by Λ^{-1}), while setting the observations with largest $Y(0)$ to have the largest weights possible (scaling observed weights by Λ). θ represents the population version of the cutoff in the ordering at which small weights are replaced by large weights. The optimal value of θ depends on Λ, w and the alternative distribution for the outcomes (conditional on covariates) specified as part of defining the favorable situation. Results similar to Theorem 3.4 hold for trimmed and augmented weighting estimators (see Appendix B.1).

Because Theorem 3.4 does not give a closed expression for $\widetilde{\Lambda}$, we examine the drivers of design sensitivity for the marginal sensitivity model using numerical simulations.

Example 3.1 (Drivers of Design Sensitivity)**.** Define the treatment assignment process and the outcome model as follows:

$$P(Z=1 \mid X) \propto \frac{\exp(\beta_\pi X)}{1+\exp(\beta_\pi X)}, \qquad Y = \beta_y X + \tau Z + u,$$

where $X \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$ and $u \overset{iid}{\sim} N(0, \sigma_y^2)$. We sample from this process under different choices for parameters $\{\beta_\pi, \beta_y, \tau, \sigma_y\}$ and estimate design sensitivity under both the variance-based sensitivity model and the marginal sensitivity model (see Appendix B.3 for full details). In contrast to the variance-based sensitivity model, design sensitivity under the marginal sensitivity model is primarily driven by the effect size and the variance in the outcomes.

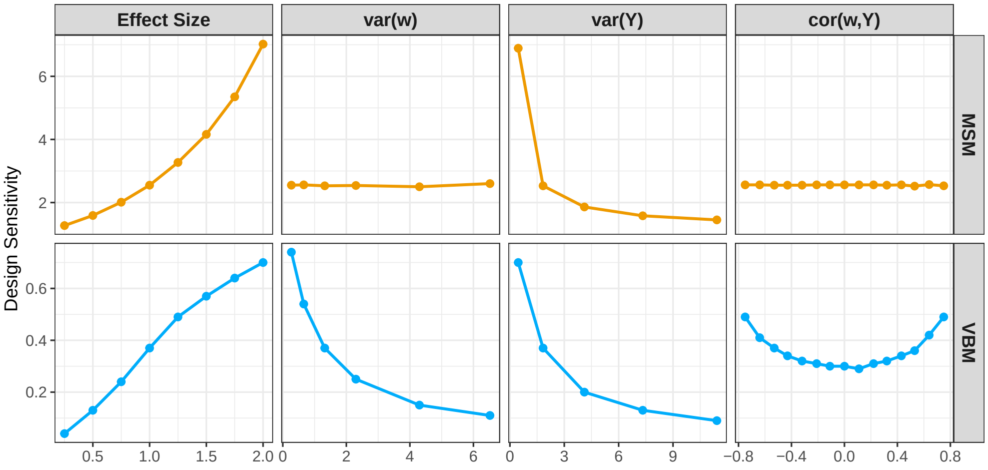


Figure 3.1: Magnitude of the design sensitivities for both MSM and VBM varying different aspects of the data generating process.

In the following section, we show how two specific design choices — augmentation of weighting designs with outcome models and trimming of estimated weights — influence design sensitivity, providing new perspective on the advantages they bring.

3.4 Design Choices that Impact Design Sensitivity

Augmentation using Outcome Models

Consider an augmented weighted estimator of the following form:

$$\hat{\tau}_W^{aug} = \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i Y_i - \left(\frac{\sum_{i=1}^n w_i (1-Z_i)(Y_i - \hat{g}(X_i))}{\sum_{i=1}^n w_i (1-Z_i)} + \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i \hat{g}(X_i) \right),$$

where $\hat{g}(X_i)$ represents an estimated outcome model. The augmented weighted estimator is doubly robust: as long as either the outcome model (i.e., $\hat{g}$), or the treatment assignment model (i.e., $\hat{w}_i$) is correctly specified, the augmented weighted estimator will consistently estimate the ATT (Tan, 2007; Bang and Robins, 2005; Kang et al., 2007). However, doubly robust estimation does not eliminate concerns about omitted variable bias. More specifically, if there is an omitted confounder that is relevant to both the treatment and outcome, then neither the outcome model nor the treatment assignment model will be correctly specified. Compounding these concerns, when one (or both) of the models is misspecified, the finite-sample performance of augmented weighted estimators may be inferior to standard weighted (or regression) estimators (Kang et al., 2007).

Our key result helps resolve these questions by demonstrating that augmentation can improve robustness to unobserved confounding in large samples even when the outcome model is misspecified, suggesting a clear advantage to augmented estimation distinct from double robustness.

Theorem 3.5 (Impact of Augmentation on Design Sensitivity).
Let $e := Y - \hat{g}(X)$ be the population residual from an arbitrary, fixed outcome model g used to augment a weighted estimate. Then, for the variance-based sensitivity model, the design sensitivity from an augmented weighted estimator will be greater than the design sensitivity for a standard weighted estimator if the following holds:

$$\text{var}(e \mid Z = 0) \leq \frac{1 - \text{cor}(w, Y \mid Z = 0)^2}{1 - \text{cor}(w, e \mid Z = 0)^2} \cdot \text{var}(Y \mid Z = 0). \tag{3.7}$$

While Theorem 3.5 assumes a fixed outcome model g, we may relax the assumption of a fixed outcome model by extending the Z-estimation framework, introduced in Appendix B.1 (Zhao et al., 2019).

If the correlation between the estimated weights and the outcomes and the correlation between the estimated weights and the residuals are roughly similar (i.e., $\text{cor}(w, Y \mid Z = 0) \approx \text{cor}(w, e \mid Z = 0)$), then Equation (3.7) simplifies to a simple comparison between the variance across the residuals and the variance of the outcomes (i.e., $\text{var}(e \mid Z = 0) \leq \text{var}(Y \mid Z = 0)$).

Theorem 3.5 highlights that the degree of improved robustness from augmentation depends directly on how much variation the estimated outcome model is able to explain in the outcomes across the control group. In other words, if $\text{var}(e \mid Z = 0)$ is relatively small, then we expect a larger improvement in design sensitivity from augmentation. Importantly, the gains to design sensitivity from augmentation are not dependent on any additional specification assumptions. Even if the outcome model is misspecified, if it successfully explains variance in the control outcomes (while maintaining similar outcome-weight correlations), then augmentation will improve the robustness of estimated effects.

Similar results hold for the marginal sensitivity model, although we cannot obtain closed-form criteria (see Appendix B.1 for more discussion). This is consistent with the results from Example 3.1, in which the design sensitivity of the marginal sensitivity model varied systematically with the variation in the outcomes. Highly variable outcomes lead to more extreme outcome values and worse worst-case bounds as in the matching context described by Rosenbaum (2005), so stabilizing outcomes improves robustness by limiting the extremity of worst-case settings.

Trimming

Another design decision that commonly arises in practice is trimming, or exclusion of units with extreme weights. Trimming implicitly redefines the estimand of interest to exclude units with extreme propensity scores, which can be helpful in cases in which researchers are worried about potential overlap or positivity violations. Under trimming, we consider a modified estimand:

$$\tau_{trim} := \mathbb{E}(Y(1) - Y(0) \mid Z = 1, X \in \mathcal{A}), \tag{3.8}$$

where $\mathcal{A} := \{x \in X \mid a \leq P(Z = 1 \mid x) \leq 1 - a\}$—i.e., the set of covariate values for which the probability of treatment, conditional on the observed covariates, is strictly bounded away from 0 and 1 (Crump et al., 2009). Equation (3.8) defines the estimand as a function of conditional probabilities in the *observable* covariates X. When researchers wish to consider trimming with respect to an oracle set $\mathcal{A}^*$, which also conditions on the omitted variable, the underlying procedure for estimating the bounds for both sensitivity models must be changed to account for trimming with respect to the ideal weights w^* instead of w. We defer the details of such a procedure for future work.

In practice, in the context of ATT estimation, we focus on trimming large weights by choosing a cutoff m for the weights w. The following theorem shows that for any degree of trimming, the relative improvement to design sensitivity for the variance-based sensitivity model depends on how successfully trimming reduces the variance of the weights compared to the reduction in the variance of Y and the change in the correlation between weights and outcomes. Design sensitivity for the marginal sensitivity model under trimming is described in Appendix B.1.

Theorem 3.6 (Impact of Trimming Weights on Design Sensitivity)**.** Let m be a cutoff above which weights are trimmed. Assume that the trimmed weights have mean 1 and that the treatment effect is constant. Then, for the variance-based sensitivity model, the design sensitivity from a trimmed estimator will be greater than the standard weighted estimator if the following holds:

$$\underbrace{\frac{\text{var}(w \mid w < m, Z = 0)}{\text{var}(w \mid Z = 0)}}_{(a)\text{Variance reduction in } w} \leq \underbrace{\frac{1 - \text{cor}(w, Y \mid Z = 0)^2}{1 - \text{cor}(w, Y \mid w < m, Z = 0)^2}}_{(b)\text{ Change in relationship between } w \text{ and } Y} \cdot \underbrace{\frac{\text{var}(Y \mid Z = 0)}{\text{var}(Y \mid w < m, Z = 0)}}_{(c)\text{ Variance reduction in } Y}. \tag{3.9}$$

Unlike augmentation, in which design sensitivity is improved so long as researchers are able to estimate an outcome model that explains some variation in the outcome, trimming provides weaker guarantees on improvements to design sensitivity. More specifically, Equation (3.9) provides a bound on the necessary variance reduction in the weights to improve design sensitivity. By construction, the variance of the trimmed weights (i.e., $\text{var}(w \mid w < m, Z = 0)$) will be no greater than the variance of the untrimmed weights (i.e., $\text{var}(w \mid Z = 0)$). If the right-hand side of Equation (3.9) were greater than or equal to 1, then this bound would be trivially met.

The magnitude of the bound depends on the correlation between the weights and the outcome. For intuition, consider the scenario in which the weights and the outcomes are highly correlated. Removing extreme weights will also remove extreme outcomes, such that the post-trimming outcome variance will be smaller than the initial variance. As a result, we expect $\text{var}(Y \mid w < m, Z = 0)$ to be less than $\text{var}(Y \mid Z = 0)$, thereby increasing term (c) in Equation (3.9). In cases when the weights are completely unrelated to the outcome, we expect $\text{var}(Y \mid Z = 0) \approx \text{var}(Y \mid w < m, Z = 0)$. In that case, Equation (3.9)-(c) will be approximately equal to 1.

An added complexity is that the bound is also dependent on potential changes in the linear relationship between w and Y (i.e., Equation (3.9)-(b)). As a result, if extreme values in the weights correspond to large values of the outcome Y, then by trimming, we may reduce the correlation between w and Y. In practice, $\text{cor}(w, Y \mid Z = 0)$ tends to be relatively low; as a result, we expect changes in the relationship between w and Y from trimming to be relatively small.

Theorem 3.6 assumed a constant treatment effect to simplify the criteria in Equation 3.9. Notably, this assumption is not necessary for the existence of design sensitivity for the trimmed estimator (see Appendix B.1 for more discussion). However, the existence of treatment effect heterogeneity introduces additional complexity for evaluating the impact of trimming on design sensitivity. If weights and treatment effects are positively correlated, trimming large weights will tend to exclude the subjects with the largest treatment effects. As such, trimming would reduce the treatment effect size and therefore also reduce design sensitivity. Conversely, trimming can increase the effect size and improve design sensitivity when the weights and treatment effect are negatively correlated. For more discussion of the connection between treatment effect heterogeneity and design sensitivity, see Rosenbaum (2007).

In practice, researchers may utilize Theorem 3.3 and Theorem 3.4 to estimate the design sensitivity under different trimming criteria. While Theorem 3.6 is formulated with respect to a trimmed estimator that directly omits units with extreme weights, the results easily extend in cases when researchers use a smooth trimmed estimator instead (see Appendix B.1 for details). By computing design sensitivity, researchers can directly assess whether or not trimming can help improve robustness to omitted variable bias, and whether or not these potential gains to robustness are worth the trade-off of using a different estimand of interest. For example, if researchers estimate the design sensitivity under trimming and find that for a small effect size, trimming a small portion of extreme weights results in a large improvement in design sensitivity, it may be helpful to perform trimming. In contrast, if researchers find that only at a large effect size or only by trimming large number of weights would trimming help with design sensitivity, it would not be worth performing trimming and altering the estimand of interest.

3.5 Empirical Application: Colombia FARC Peace Agreement

Background and Context

To illustrate design sensitivity, we re-analyze a study from Hazlett and Parente (2023). After decades of fighting, the Colombian government under President Juan Santos reached a historic peace deal with the Revolutionary Armed Forces of Colombia (FARC). However, in a 2016 referendum, the public narrowly voted to reject the peace deal. The FARC peace deal remains an important case study in understanding drivers of support for peace.

Following Hazlett and Parente (2023), we examine two prevalent hypotheses for drivers of support for peace: (1) exposure to violence, and (2) presidential support. We define the outcome of interest as the proportion of individuals at the municipality level who voted in favor of the peace deal. For exposure to violence, treatment is defined by whether any recorded deaths attributed to FARC occurred in a municipality. For presidential support, treatment is defined as whether or not President Santos won the popular vote in the municipality in the second round of presidential elections (which would have implied that he won that particular region). We estimate inverse propensity score weights by fitting a logistic regression using the available pre-treatment covariate data. This includes variables such as past incidents of FARC-related deaths, GDP per capita for a specific municipality, and the number of people who live in each municipality.

We vary the possible treatment effects and estimate the resulting design sensitivities. Figure 3.2 displays the results. Because we are examining the percentage of individuals who vote in favor of the peace deal in a municipality as the outcome, the range of possible treatment effects is restricted by the fact that the average treatment outcome cannot be outside the range 0 - 100%. In practice, researchers can estimate design sensitivity by calibrating to a chosen outcome distribution. For illustrative purposes, we calibrate the estimated design sensitivities using the true outcome distribution. However, the results can be estimated for any arbitrary distribution for the outcome. See Appendix B.2 for recommendations for using a planning sample to calibrate outcome distributions.

Illustration on the Variance-Based Sensitivity Models

To examine the potential impact of augmentation, we vary the amount of variation that can be explained in the control outcomes by a hypothetical outcome model and estimate the updated design sensitivity. Consistent with Theorem 3.5, we see that as the amount of variation explained in the outcomes increases, the amount of improvement in design sensitivity also increases. Notably, even in cases when the outcome model can explain 50% of the variation in the control outcomes, there is a relatively limited impact on the design sensitivity for the variance-based model.

To assess the impact of trimming, we estimate the design sensitivity of the weighted estimator, trimming at thresholds of 0.9 and 0.8 (trimming observations with estimated

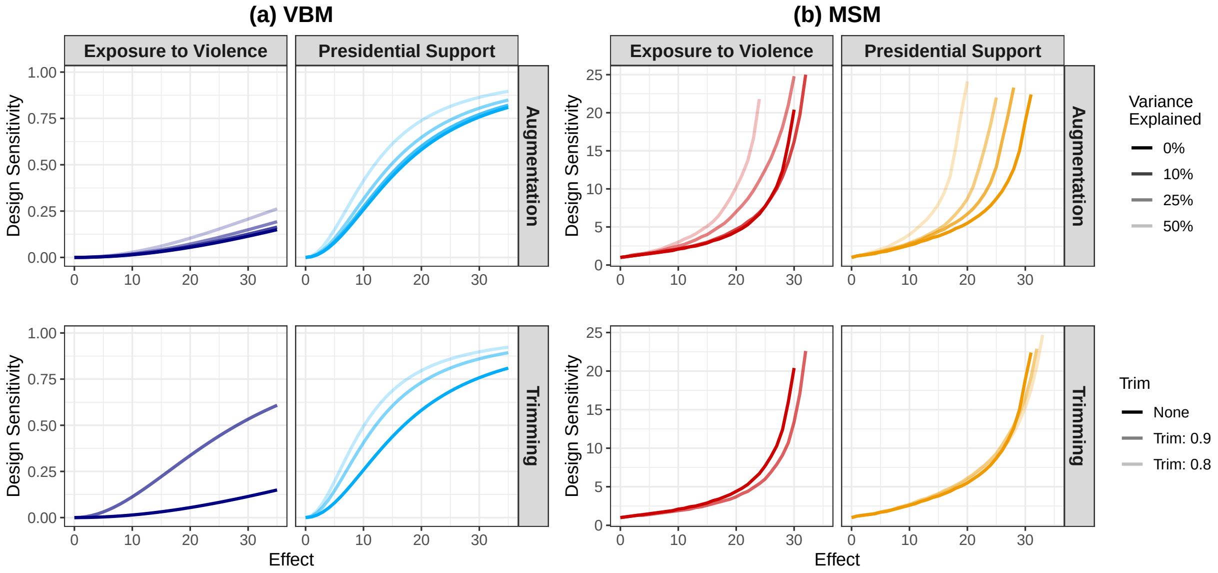


Figure 3.2: Design sensitivities under augmentation and trimming for (a) the variance-based model and (b) the marginal sensitivity model.

propensity scores greater than these values). This improves the design sensitivity uniformly across all effect sizes. Trimming a small number of extreme weights results in large improvements in the design sensitivity, even for a relatively low effect size. For example, for the hypothesis of exposure to violence, trimming weights that correspond to propensity scores greater than 0.9 would result in excluding 3 observations, out of 1,123 total observations (i.e., 0.26% of control units). For an effect size of 10, this would correspond to an increase in the design sensitivity from $\tilde{R}^2 = 0.01$ to $\tilde{R}^2 = 0.11$. This implies that assuming an effect size of 10, without trimming, we could only identify a true effect if the imbalance from an omitted confounder explained less than 1% of the variation in the ideal weights. However, after trimming, we would be able to identify a true effect even accounting for a possible confounder that is up to 10 times more imbalanced.

Importantly, neither design choice (i.e., augmenting and trimming) appear to hurt design sensitivity for the variance-based sensitivity models in this setting. However, from estimating the design sensitivities, it is clear that there are substantial gains to robustness from trimming a few observations from the study. In contrast, while fitting a predictive outcome model can help improve design sensitivity, these improvements are more marginal.

Illustration on the Marginal Sensitivity Model

We now illustrate design sensitivity on the marginal sensitivity model. We see a large improvement in the design sensitivity for the marginal sensitivity model from augmentation across all effect sizes. This is likely because the marginal sensitivity model is more susceptible to extreme values in the outcome. By augmenting the weighted estimator, we are able to reduce extreme values in the outcome. Even an outcome model that can explain 10% of the variation in the control outcomes can result in a substantial improvement in the design sensitivity of the marginal sensitivity model. (See Figure 3.2-(b) for illustration.)

In contrast to the variance-based sensitivity models, we see that for for the marginal sensitivity models, at small effect sizes, trimming does not affect the design sensitivity. However, for large effect sizes (i.e., $\tau > 25$), trimming actually results in a slight reduction in the design sensitivity. As such, in settings where researchers are concerned about robustness to a worst-case error, design sensitivity would suggest that researchers should not perform trimming, and fitting a predictive outcome model would be most helpful at improving robustness.

Remark on the Choice of Sensitivity Model and Interpretation

While design sensitivity is useful for deciding which weighting approaches we expect to be most robust given a particular mode of sensitivity analysis, we do not view it as particularly helpful for choosing among different sensitivity models. Design sensitivities for the marginal and variance-based models are parameterized very differently and are not directly comparable. As such, throughout the chapter and analysis, we have restricted attention to maximizing the design sensitivity for a *fixed* set of sensitivity models. We refer readers to Huang and Pimentel (2022) for more discussion about comparing the performance of different sensitivity models, as well as Rosenbaum (2015) for discussion on comparisons of different sensitivity models below the design sensitivity threshold. In general, we recommend that practitioners decide *a priori* which sensitivity model best captures the type of unobserved confounder that concerns them most from a substantive perspective and pursue design sensitivity under that model.

The interpretation of design sensitivity magnitude depends on the underlying sensitivity model. As such, determining whether a design sensitivity is large or small requires researchers to reason about whether the amount of confounding represented by $\tilde{\Gamma}$ is plausible. We recommend the use of existing tools for sensitivity analysis such as formal benchmarking (Huang and Pimentel, 2022; Huang, 2022) and amplification (Soriano et al., 2021) to help with interpretation.

3.6 Conclusion

In this chapter, we introduced *design sensitivity* for weighted estimators. This asymptotic measure of robustness allows researchers to consider how certain design choices in their observational studies can affect sensitivity to omitted confounders at the design stage. Design sensitivity can be estimated for a general set of sensitivity models that meet a relatively weak set of regularity conditions. We derive the design sensitivities for two commonly used sensitivity models: the variance-based sensitivity model, which constrains an average error from omitting a confounder, and the marginal sensitivity model, which constrains a worst-case error from omitting a confounder. We show that trimming and augmentation—two common design choices that researchers make in practice—can influence design sensitivity. Thus, beyond the standard discussions of variance reduction for trimming and double robustness for augmentation, these decisions can also impact robustness to omitted confounders. We illustrate our framework in a study of the 2016 Colombian peace agreement, for which trimming drastically improves design sensitivity under the variance-based sensitivity model, and augmentation improves design sensitivity under the marginal sensitivity model.

Several lines of future work follow naturally. While we provide explicit calculations for two commonly used sensitivity models, the framework introduced applies more generally. An interesting avenue of future work could compare how design sensitivities across a wider array of different sensitivity models respond to design choices. Future work could also examine design sensitivity with respect to other design choices, including the choice between ATT, ATE, and quantile effect estimands (Greifer and Stuart, 2021).

The idea of "sharp" sensitivity analysis (Dorn and Guo, 2021), or sensitivity analysis that asymptotically recovers exactly the set of true effect estimates without any conservatism, is somewhat related to design sensitivity. Both approaches aim to make the estimated intervals $[\hat{L}_{\nu(\Gamma,w)}, \hat{U}_{\nu(\Gamma,w)}]$ as small as possible in large samples. However, Dorn and Guo emphasize changing the sensitivity analysis method itself by incorporating additional constraints, in contrast to our focus on changing other aspects of the design while holding the sensitivity analysis fixed. As such, design sensitivity is not well suited to guiding user choice between sharp and non-sharp sensitivity analysis methods. However, Theorems 3.1-3.2 should apply to sharp sensitivity analysis such that power and design sensitivity formulas could be derived. Based on findings in Dorn and Guo (2021), we anticipate that design sensitivity may behave differently for their sharp sensitivity analysis approach than for methods like the marginal sensitivity analysis of Zhao et al. (2019) which has been proven not to be sharp. For example, we expect that the width of sharp limiting sets may not be affected by the variance of the study outcomes, in which case augmentation might no longer provide any benefit to design sensitivity. We leave such questions to be taken up by future authors.

Finally, design sensitivity is defined in the context of the asymptotic limiting distributions. This formulation allows us to disentangle uncertainty from sampling error from uncertainty from omitted confounding. However, it may not provide clear guidance in small-sample settings or in settings where several designs exhibit similar design sensitivities. In the context of matching, Rosenbaum (2015) computes Bahadur efficiencies of sensitivity analyses as a way to compare robustness of different design specifications with more granularity than design sensitivity can provide. Future work should explore the potential of this approach for weighting estimators.

Bibliography

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132.

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. (2021). The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*.

Ben-Michael, E., Feller, A., and Rothstein, J. (2018). The augmented synthetic control method. *arXiv preprint arXiv:1811.04170*.

Ben-Michael, E., Feller, A., and Rothstein, J. (2020). Variation in impacts of letters of recommendation on college admissions decisions: Approximate balancing weights for treatment effect heterogeneity in observational studies.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, pages 1196–1217.

Björnberg, K. A., Vahter, M., Petersson-Grawe, K., Glynn, A., Cnattingius, S., Darnerud, P., Atuma, S., Aune, M., Becker, W., and Berglund, M. (2003). Methyl mercury and inorganic mercury in swedish pregnant women and in cord blood: influence of fish consumption. *Environmental Health Perspectives*, 111(4):637–641.

Brown, L. D., Casella, G., and Gene Hwang, J. (1995). Optimal confidence sets, bioequivalence, and the limacon of pascal. *Journal of the American Statistical Association*, 90(431):880–889.

Chattopadhyay, A., Christopher H. Hase, and Zubizarreta, J. R. (2020). Balancing Versus Modeling Approaches to Weighting in Practice. *Statistics in Medicine*, 39(24):3227–3254.

Chattopadhyay, A. and Zubizarreta, J. R. (2021). On the implied weights of linear regression for causal inference. *arXiv preprint arXiv:2104.06581*.

Cinelli, C. and Hazlett, C. (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B*, 82(1):39–67.

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Deville, J. C., Särndal, C. E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368.

Dorn, J. and Guo, K. (2021). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *arXiv preprint arXiv:2102.04543*.

Firpo, S. and Possebom, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2).

Fogarty, C. B. (2020). Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, 115(531):1518–1530.

Franks, A., D'Amour, A., and Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, pages 1–33.

Greifer, N. and Stuart, E. A. (2021). Choosing the estimand when matching or weighting in observational studies. *arXiv preprint arXiv:2106.10577*.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hartman, E. and Huang, M. (2022). Sensitivity analysis for survey weights. *arXiv preprint arXiv:2206.07119*.

Hazlett, C. and Parente, F. (2023). From" is it unconfounded?" to" how much confounding would it take?": Applying the sensitivity-based approach to assess causes of support for peace in colombia. *Journal of Politics (forthcoming)*.

Heller, R., Rosenbaum, P. R., and Small, D. S. (2009). Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*, 104(487):1090–1101.

Hirshberg, D. A., Maleki, A., and Zubizarreta, J. (2019). Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*.

Hong, G., Yang, F., and Qin, X. (2020). Did you conduct a sensitivity analysis? a new weighting-based approach for evaluations of the average treatment effect for the treated. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Howard, S. R. and Pimentel, S. D. (2021). The uniform general signed rank test and its design sensitivity. *Biometrika*, 108(2):381–396.

Hsu, J. Y., Small, D. S., and Rosenbaum, P. R. (2013). Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108(501):135–148.

Huang, M. (2022). Sensitivity analysis in the generalization of experimental results. *arXiv preprint arXiv:2202.03408*.

Huang, M. and Pimentel, S. D. (2022). Variance-based sensitivity analysis for weighting estimators result in more informative bounds. *arXiv preprint arXiv:2208.01691*.

Ishikawa, K. and He, N. (2023). Kernel conditional moment constraints for confounding robust inference. *arXiv preprint arXiv:2302.13348*.

Jin, Y., Ren, Z., and Zhou, Z. (2022). Sensitivity analysis under the f-sensitivity models: Definition, estimation and inference. *arXiv preprint arXiv:2203.04373*.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Keele, L., Ben-Michael, E., Feller, A., Kelz, R., and Miratrix, L. (2020). Hospital Quality Risk Standardization via Approximate Balancing Weights.

Klaassen, C. A. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620.

Mahaffey, K. R., Clickner, R. P., and Bodurow, C. C. (2004). Blood organic mercury and dietary mercury intake: National health and nutrition examination survey, 1999 and 2000. *Environmental health perspectives*, 112(5):562–570.

Neyman, J. (1990 [1923]). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Pimentel, S. D. and Kelz, R. R. (2020). Optimal tradeoffs in matched designs comparing us-trained and internationally trained surgeons. *Journal of the American Statistical Association*, 115(532):1675–1688.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, 110(510):515–527.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika*, 91(1):153–164.

Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician*, 59(2):147–152.

Rosenbaum, P. R. (2007). Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics*, 63(4):1164–1171.

Rosenbaum, P. R. (2010). *Design of observational studies*, volume 10. Springer.

Rosenbaum, P. R. (2011). What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? In *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, pages 87–114. Springer.

Rosenbaum, P. R. (2015). Bahadur efficiency of sensitivity analyses in observational studies. *Journal of the American Statistical Association*, 110(509):205–217.

Rosenbaum, P. R. et al. (2010). *Design of observational studies*, volume 10. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218.

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

Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405.

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

Rubin, D. B. (1980a). Discussion of 'Randomization analysis of experimental data: The Fisher randomization test comment' by Basu. *Journal of the American Statistical Association*, 75(371):591–593.

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

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36.

Soriano, D., Ben-Michael, E., Bickel, P. J., Feller, A., and Pimentel, S. D. (2021). Interpretable sensitivity analysis for balancing weights. *arXiv preprint arXiv:2102.13218*.

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

Tan, Z. (2007). Comment: Understanding or, ps and dr. *Statistical Science*, 22(4):560–568.

Tan, Z. (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158.

Tudball, M., Hughes, R., Tilling, K., Bowden, J., and Zhao, Q. (2019). Sample-constrained partial identification with application to selection bias. *arXiv preprint arXiv:1906.10159*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274.

Veitch, V. and Zaveri, A. (2020). Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems*, 33:10999–11009.

Wang, Y. and Zubizarreta, J. R. (2019). Minimal approximately balancing weights: asymptotic properties and practical considerations. *arXiv preprint arXiv:1705.00998*.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493.

Zhang, Y. and Zhao, Q. (2022). Bounds and semiparametric inference in l_∞ and l_2-sensitivity analysis for observational studies. *arXiv preprint arXiv:2211.04697*.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*, 47(2):965–993.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Zhao, Q., Small, D. S., and Rosenbaum, P. R. (2018). Cross-screening in observational studies that test many hypotheses. *Journal of the American Statistical Association*, 113(523):1070–1084.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.

Appendix A

Supplementary materials for Chapter 2

A.1 Proofs

Proof of Theorem 2.1

Proof. We prove that, after centering, the difference between the mean computed from estimating and evaluating the inverse probability function γ on bootstrap data and the mean computed from using the true function γ and evaluating on actual data is of order $n^{-1/2}$.

For simplicity, we consider estimating the population mean from an independent and identically distributed random sample with missing outcome data. For unit i, let Y_i be the outcome, X_i be a vector of observed covariates, and Z_i be a response indicator, where $Z_i = 1$ if we observe unit i's outcome and $Z_i = 0$ otherwise. In addition, let $\gamma_P(X) = 1/\pi_P(X)$ be the *population weight* associated with the unit with covariate X. We consider using estimator $\hat{\mu}^{(h)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}^{(h)}(X_i, Y_i)Z_iY_i$ to estimate $\mu^{(h)} = \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X,Y]] = \mathbb{E}[\mathbb{E}[\frac{ZY}{\pi_P^{(h)}(X,Y)}|X,Y]] = \mathbb{E}[\frac{ZY}{\pi_P^{(h)}(X,Y)}] = \mathbb{E}[\gamma_P^{(h)}(X,Y)ZY]$ (by the law of iterated expectations) from observed data $O_i = (X_i, Z_i, Y_iZ_i)_{i=1}^{n}$ drawn from joint distribution $P(\cdot)$. Theorem 2.1 applies for any known deviation from ignorability represented by the log odds ratio $h(x,y) = \log \mathrm{OR}(\pi(x), \pi(x,y))$. Without loss of generality, we use $h(x,y) = 0$ and suppress the dependency of $\hat{\mu}^{(h)}$ and $\mu^{(h)}$ on $h(x,y)$ for notational simplicity.

We sample split to make the proof and arguments simpler and more transparent (see Klaassen, 1987). The proof can equivalently be done without sample splitting, but we sample split to avoid the associated complexities. We split the data into two equally sized samples, $i = 1, \ldots, m$ and $i = m+1, \ldots, n$. For both samples, we take an iid bootstrap sample of size m from the respective empirical distribution to obtain data $O_i^* = (X_i^*, Z_i^*, Y_i^*Z_i^*)_{i=1}^{m}$ and $O_i^* = (X_i^*, Z_i^*, Y_i^*Z_i^*)_{i=m+1}^{n}$. Let $\hat{\gamma}^*$ denote an estimate of γ using bootstrap data. We estimate $\hat{\gamma}^*(X)$ in one bootstrap sample and evaluate in the other bootstrap sample. We then switch roles and take a weighted average of the two estimates proportional to $\sum_{i=1}^{m} Z_i^*$ in both bootstrap samples to obtain an efficient estimate. This sample splitting approach with reversing roles and averaging yields the same estimate as without sample splitting to order $o(n^{-1/2})$. We demonstrate this through simulation (see Appendix A.2). We examine the case where we evaluate on the bootstrap sample from the second half of the data and estimate $\hat{\gamma}^*(X)$ from the bootstrap sample from the first half.

We make the following mild assumptions on how $\hat{\gamma}$ is constructed:

Assumption A.1. Consider function $\tilde{\gamma} : \mathscr{X}^m \times \{0,1\}^m \to \mathbb{R}^+$. As an example, consider the function corresponding to the stable balancing weights optimization problem (2.7). Let $\hat{\gamma}_n^*(x) = \tilde{\gamma}_{P_m^*}(X_1^*, \ldots, X_m^*, Z_1^*, \ldots, Z_m^*, x)$ and $\hat{\gamma}_n(x) = \tilde{\gamma}_{P_m}(X_1, \ldots, X_m, Z_1, \ldots, Z_m, x)$, where P_m^* and P_m are the empirical distributions for the bootstrap sample from the first half of the data and the actual first half of the data, respectively, be such that:

1. $\tilde{\gamma}$ is uniformly bounded in m and x.
2. $\mathbb{E}_1\left[\left(\sup_x \left|\hat{\gamma}_n^*(x) - \hat{\gamma}_n(x)\right|\right)^2\right] = o_p(1)$.
3. $\sup_x \left|\hat{\gamma}_n(x) - \gamma_P(x)\right| = o_p(1)$.[1]
4. $\mathbb{E}\left[\hat{\gamma}_n(X)ZY\right] - \mathbb{E}\left[\gamma_P(X)ZY\right] = o(n^{-1/2})$.
5. Y has a finite second moment: $\mathbb{E}[Y^2] \leq M$, where M is a constant.

Assumption A.1.4 assumes that the bias of $\hat{\mu}$ for estimating μ is of order $o(n^{-1/2})$. The assumptions for Theorem 3 in Wang and Zubizarreta (2019) and Theorem 2 in Hirshberg et al. (2019) are possible conditions under which our set of assumptions hold. These are representative of typical assumptions in this setting where the estimator is assumed to be a function of d covariates with assumptions on the number of components of an orthogonal expansion. There are various alternative assumptions, all of which boil down to requiring that $\gamma_P(\cdot)$ can be characterized by a low-dimensional structure.

In conjunction with Assumptions A.1.2 and A.1.3, Assumption A.1.4 can be implausible with high-dimensional covariates or when the covariate distribution can be specified only by a high-dimensional parametric model which requires estimation. This caution is independent of the method used to estimate $\gamma_P(X)$. We provide an example to illustrate the issues that can arise in high-dimensional settings. Suppose X has dimension p and that $\hat{\gamma}_n(X)$ uses Nadaraya-Watson type kernel density estimation for $\gamma_P(X)$. Further, assume that $\gamma_P(\cdot)$ has bounded partial derivatives of order $\leq s$. Then, it is well known that if $\hat{\gamma}_n(X)$ has bandwidths $h_1 = \cdots = h_p = h$, then $\mathbb{E}\left[\hat{\gamma}_n(X)|X\right] = \gamma_P(X) + O(h^s)$ and $\mathbb{E}\left[\left|\hat{\gamma}_n(X) - \mathbb{E}\left[\hat{\gamma}_n(X)|X\right]\right|^2\right] = \Omega((nh^p)^{-1})$. In order to have $nh^p \to \infty$ and $nh^{2s} \to 0$, we must have:

[1] Wang and Zubizarreta (2019)'s Theorem 2 proves that Assumption A.1.3 holds for weights estimated by SBW (2.7).

1. $h \to 0$ slower than $n^{-\frac{1}{p}}$ and
2. $h \to 0$ faster than $n^{-\frac{1}{2s}}$.

This is possible only if $s \geq p/2$. In fact, more sophisticated heuristics yield replacement of $\frac{1}{2s}$ by $\frac{1}{4s}$. Intuitively, if p is large, this assumption is unrealistic in any case. It implies that $\gamma_P(\cdot)$ has a Taylor expansion to order s with $\Omega(p^s)$ bounded coefficients, which means $p^{\frac{p}{2}}$ for $s \geq p/2$. For $p = 100$, this yields 100^{50}! This example illustrates that there is reason to be skeptical of the plausibility of Assumption A.1.4 in high-dimensional settings. Additional research into propensity score estimation with high-dimensional covariates would seem important.

These assumptions together imply that $\hat{\gamma}_n^*$ is consistently uniform for γ. Assumption A.1 verifies

$$\begin{aligned}
&\mathbb{E}_1\left[\left(\sup_x\left|\hat{\gamma}_n^*(x)-\gamma_P(x)\right|Y_{m+1}Z_{m+1}\right)^2\right]\\
=&\mathbb{E}_1\left[\left(\sup_x\left|\hat{\gamma}_n^*(x)-\gamma_P(x)\right|\right)^2\right]\mathbb{E}_1\left[Y_{m+1}^2Z_{m+1}^2\right]\\
=&o(1),
\end{aligned}$$

where $\mathbb{E}_1$ denotes the conditional expectation given the first sample. Note, the conditions in Assumption A.1 are stronger than needed and could be relaxed.

We proceed conditional on the first sample $O_i = (X_i, Z_i, Y_iZ_i)_{i=1}^m$ and the first bootstrap sample $O_i^* = (X_i^*, Z_i^*, Y_i^*Z_i^*)_{i=1}^m$. Therefore, $\hat{\gamma}_n^*$ is a completely known function. Let $\mathbb{E}^*$ denote the conditional expectation of the second bootstrap sample given the actual second sample.

Since

$$\mathbb{E}^*\left[\frac{1}{m}\sum_{i=m+1}^{2m}\hat{\gamma}_n^*(X_i^*)Z_i^*Y_i^*\right] = \frac{1}{m}\sum_{i=m+1}^{2m}\hat{\gamma}_n^*(X_i)Z_iY_i,$$

then, by Theorem 2.1 from Bickel and Freedman (1981),

$$\frac{1}{m}\sum_{i=m+1}^{2m}\hat{\gamma}_n^*(X_i^*)Z_i^*Y_i^* - \frac{1}{m}\sum_{i=m+1}^{2m}\hat{\gamma}_n^*(X_i)Z_iY_i \tag{A.1}$$

$$\text{and} \quad \frac{1}{m}\sum_{i=m+1}^{2m}\left(\hat{\gamma}_n^*(X_i)Z_iY_i - \mathbb{E}_1\left[\hat{\gamma}_n^*(X_{m+1})Z_{m+1}Y_{m+1}\right]\right) \tag{A.2}$$

have the same limiting distribution. Since (A.1) and (A.2) have the same limiting distribution, instead of showing

$$\begin{aligned}&\frac{1}{m}\sum_{i=m+1}^{2m}\hat{\gamma}_n^*(X_i^*)Z_i^*Y_i^* - \mathbb{E}^*\Big[\frac{1}{m}\sum_{i=m+1}^{2m}\hat{\gamma}_n^*(X_i^*)Z_i^*Y_i^*\Big]\\ =&\frac{1}{m}\sum_{i=m+1}^{2m}\gamma_P(X_i)Z_iY_i - \mathbb{E}_1\Big[\gamma_P(X_{m+1})Z_{m+1}Y_{m+1}\Big] + o_p(n^{-1/2})\end{aligned} \tag{A.3}$$

to show that the bootstrap can be validly applied, it suffices to show that the difference between the mean with the true γ and the mean with $\hat{\gamma}_n^*$ estimated on the bootstrap data is of order $n^{-1/2}$. Therefore, we show

$$\begin{aligned}&\frac{1}{m}\sum_{i=m+1}^{2m}\hat{\gamma}_n^*(X_i)Z_iY_i - \mathbb{E}_1\Big[\hat{\gamma}_n^*(X_{m+1})Z_{m+1}Y_{m+1}\Big]\\ =&\frac{1}{m}\sum_{i=m+1}^{2m}\gamma_P(X_i)Z_iY_i - \mathbb{E}_1\Big[\gamma_P(X_{m+1})Z_{m+1}Y_{m+1}\Big] + o_p(n^{-1/2}).\end{aligned} \tag{A.4}$$

We have now reduced the problem to showing that the true function γ can be replaced with $\hat{\gamma}_n^*$. In order to show this, we use properties of $\hat{\gamma}_n^*$ from Assumption A.1. First, we let

$$\Delta(X_i, Y_i, Z_i) = (\hat{\gamma}_n^*(X_i) - \gamma_P(X_i))Z_iY_i - \mathbb{E}_1\Big[(\hat{\gamma}_n^*(X_{m+1}) - \gamma_P(X_{m+1}))Z_{m+1}Y_{m+1}\Big].$$

Note that the difference between the terms on the left and right hand sides of (A.4) is equal to $\frac{1}{m}\sum\limits_{i=m+1}^{2m}\Delta(X_i, Y_i, Z_i)$. Additionally, note that $\mathbb{E}_1\Big[\Delta(X_i, Y_i, Z_i)\Big] = 0$. Therefore,

$$\begin{aligned}&\mathbb{E}_1\Big[\Big(\frac{1}{m}\sum_{i=m+1}^{2m}\Delta(X_i, Y_i, Z_i)\Big)^2\Big]\\ =&\frac{1}{m}\mathbb{E}_1\Big[\Delta(X_{m+1}, Y_{m+1}, Z_{m+1})^2\Big].\end{aligned}$$

Since $m = \Omega(n)$, by Assumption A.1,

$$\begin{aligned}
&\mathbb{E}_1\left[\Delta(X_{m+1}, Y_{m+1}, Z_{m+1})^2\right] \\
=&\mathbb{E}_1\left[([\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1})]\, Z_{m+1}Y_{m+1})^2\right] \\
&- 2\mathbb{E}_1\Big\{\left[\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1})\right] Z_{m+1}Y_{m+1}\mathbb{E}_1\Big[\left[\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1})\right] Z_{m+1}Y_{m+1}\Big]\Big\} \\
&+ \mathbb{E}_1\Big\{\mathbb{E}_1\Big[\left[\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1})\right] Z_{m+1}Y_{m+1}\Big]^2\Big\} \\
=&\mathbb{E}_1\Big[\Big(\left[\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1})\right] Z_{m+1}Y_{m+1}\Big)^2\Big] - \mathbb{E}_1\Big[\left[\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1})\right] Z_{m+1}Y_{m+1}\Big]^2 \\
\le&\mathbb{E}_1\Big[\Big(\left[\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1})\right] Z_{m+1}Y_{m+1}\Big)^2\Big] \\
\le&M \cdot \mathbb{E}_1\left[(\hat{\gamma}^*_n(X_{m+1}) - \gamma_P(X_{m+1}))^2\right] \\
=&o_p(1).
\end{aligned}$$

Therefore, (A.4) follows. □

A.2 Simulation for sample splitting

We conduct simulations to demonstrate the validity of the sample splitting technique that we use to prove Theorem 2.1 in Appendix A.1. We show that the bootstrap distributions for the balancing weights estimates of μ_0 with and without sample splitting are quite similar.

The setup of the simulations is as follows. We draw 10,000 iid samples where covariates X_1 and X_2 are drawn from standard normal distributions, treatment indicator Z_i is a bernoulli random variable with probability $= 0.5 + 0.07X_{1i} + 0.07X_{2i} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 0.03^2)$, and $Y_i = 0.2Z_i + 0.5X_{1i} + 0.5X_{2i} + \delta_i$, where $\delta_i \sim \mathcal{N}(0, 0.2^2)$. We run 1,000 simulations and estimate μ_0 with and without sample splitting using weights obtained by entropy balancing with exact balance from Hainmueller (2012). We observe in Figure A.1 that the bootstrap distributions of the estimates with and without sample splitting are comparable.

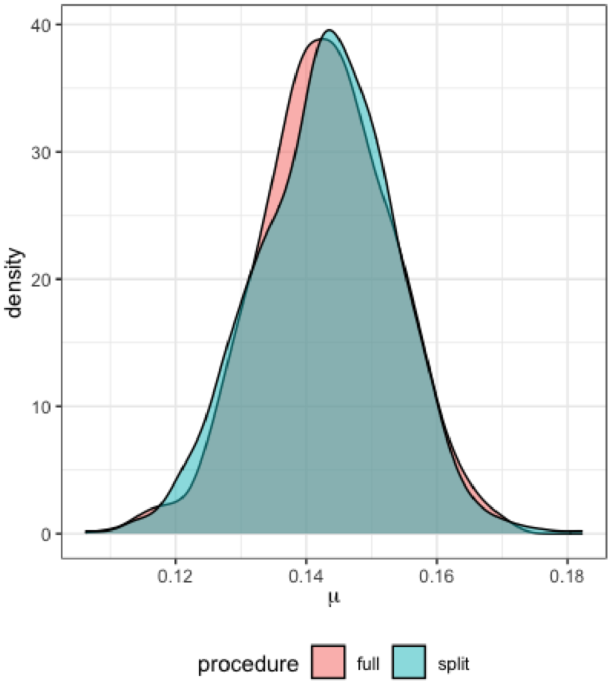


Figure A.1: Bootstrap distributions of estimates of μ_0 with the full data and with sample splitting

A.3 Average treatment effect on the treated

In many settings, researchers are interested in estimating the *Population Average Treatment Effect on the Treated* (PATT):

$$\tau_T = \mathbb{E}[Y(1) - Y(0)|Z = 1] = \mu_{11} - \mu_{01}, \tag{A.5}$$

where $\mu_{11} = \mathbb{E}[Y(1)|Z = 1]$ and $\mu_{01} = \mathbb{E}[Y(0)|Z = 1]$. Since μ_{11} is identifiable from observed data, we primarily focus on estimating μ_{01}.

Our procedure for performing sensitivity analysis outlined in Section 2.3 largely still holds. The primary details that differ for the PATT are as follows. First, for a particular $h \in H(\Lambda)$, we can write the shifted estimand as

$$\mu_{01}^{(h)} = \mathbb{E}\left[(1-Z)\frac{\pi^{(h)}(X, Y(0))}{1-\pi^{(h)}(X, Y(0))}\right]^{-1} \mathbb{E}\left[(1-Z)\frac{\pi^{(h)}(X, Y(0))}{1-\pi^{(h)}(X, Y(0))}Y\right]. \tag{A.6}$$

The corresponding shifted estimator for $\mu_{01}^{(h)}$ is

$$\hat{\mu}_{01}^{(h)} = \left(\sum_{Z_i=0} e^{-h(X_i, Y_i(0))}\hat{\gamma}(X_i)\right)^{-1} \sum_{Z_i=0} e^{-h(X_i, Y_i(0))}\hat{\gamma}(X_i)Y_i. \tag{A.7}$$

We make the following modifications to our amplification described in Section 2.4 for the ATT. Where $U \in \mathbb{R}$ represents a latent unmeasured confounding variable, standardized to have mean zero and variance 1, we consider a working model for the conditional expectation of the control potential outcome:

$$\mathbb{E}[Y(0) \mid X = x, U = u, Z = 1] = f(x) + \beta_{u0} \cdot u. \tag{A.8}$$

Then, we define the bias to be the difference between the true expected value of control potential outcomes for treated units μ_{01} and the IPW estimand. We decompose the bias into (i) the strength of the unmeasured confounder U in predicting $Y(0)$ for treated units beyond the observed covariates, β_{u0} and (ii) the imbalance in U, δ_{u0}:

$$\mathbb{E}[Y(0) \mid Z = 1] - \mathbb{E}\left[\frac{1-Z}{\mathbb{P}(Z=1)}\frac{\pi(X)}{1-\pi(X)}Y\right] = \beta_{u0} \cdot \underbrace{\left(\mathbb{E}\left[U \mid Z = 1\right] - \mathbb{E}\left[\frac{1-Z}{\mathbb{P}(Z=1)}\frac{\pi(X)}{1-\pi(X)}U\right]\right)}_{\delta_{u0}}.$$

Next, we derive upper and lower bounds for this product by using the partial identification of μ_{01} under the marginal sensitivity model:

$$\inf_{h \in \mathcal{H}(\Lambda^*)} \mu_{01}^{(h)} - \mathbb{E}\left[\frac{1-Z}{\mathbb{P}(Z=1)}\frac{\pi(X)}{1-\pi(X)}Y\right] \leq \beta_{u0} \cdot \delta_{u0} \leq \sup_{h \in \mathcal{H}(\Lambda^*)} \mu_{01}^{(h)} - \mathbb{E}\left[\frac{1-Z}{\mathbb{P}(Z=1)}\frac{\pi(X)}{1-\pi(X)}Y\right].$$

Finally, we construct finite-sample versions of these population bounds by bounding the bias as the maximum of the absolute values of the highest and lowest possible differences in the estimated values,

$$|\beta_{u0} \cdot \delta_{u0}| \leq \max \left\{ \left| \inf_{h \in \mathcal{H}(\Lambda^*)} \hat{\mu}_{01}^{(h)} - \hat{\mu}_{01} \right|, \left| \sup_{h \in \mathcal{H}(\Lambda^*)} \hat{\mu}_{01}^{(h)} - \hat{\mu}_{01} \right| \right\}.$$

Appendix B

Supplementary materials for Chapter 3

B.1 Proofs

Theorem 3.1 (Power of a Sensitivity Analysis)

For a general class of sensitivity models $\nu(\Gamma, w)$, define $\tau_{\nu(\Gamma,w)}$ as the minimum value in the set of possible point estimates (i.e., $\tau_{\nu(\Gamma,w)} := \inf_{\tilde{w} \in \nu(\Gamma,w)} \tau(\tilde{w})$). Define $\xi_{\nu(\Gamma,w)} := \tau_W - \tau_{\nu(\Gamma,w)}$. Finally, define $k_\alpha := 1 - \Phi(\alpha)$. Then, the power of a sensitivity analysis is defined as:

$$\Pr\left(\frac{\sqrt{n}(\hat{\tau}_W - \xi_{\nu(\Gamma,w)})}{\sigma_{\nu(\Gamma,w)}} \geq k_\alpha\right) = \Pr\left(\frac{\sqrt{n}\cdot(\hat{\tau}_W - \tau_W)}{\sigma_W} \geq \frac{k_\alpha \cdot \sigma_{\nu(\Gamma,w)} + \sqrt{n}\cdot(\xi_{\nu(\Gamma,w)} - \tau_W)}{\sigma_W}\right)$$

$$\simeq 1 - \Phi\left(\frac{k_\alpha \cdot \sigma_{\nu(\Gamma,w)} + \sqrt{n}\cdot(\xi_{\nu(\Gamma,w)} - \tau_W)}{\sigma_W}\right),$$

Proof.

$$\Pr\left(\frac{\sqrt{n}(\hat{\tau}_W - \xi_{\nu(\Gamma,w)})}{\sigma_{\nu(\Gamma,w)}} \geq k_\alpha\right) = \Pr\left(\frac{\sqrt{n}(\hat{\tau}_W - \xi_{\nu(\Gamma,w)})}{\sigma_W} \geq k_\alpha \cdot \frac{\sigma_{\nu(\Gamma,w)}}{\sigma_W}\right)$$

Adding and subtracting $\sqrt{n}\tau_W/\sigma_W$ to both sides results in the following:

$$= \Pr\left(\frac{\sqrt{n}\cdot(\hat{\tau}_W - \tau_W)}{\sigma_W} \geq \frac{k_\alpha \cdot \sigma_{\nu(\Gamma,w)} + \sqrt{n}\cdot(\xi_{\nu(\Gamma,w)} - \tau_W)}{\sigma_W}\right)$$

Noting that $\hat{\tau}_W \xrightarrow{d} N(\tau_W, \sigma_W^2)$ concludes the proof:

$$\simeq 1 - \Phi\left(\frac{k_\alpha \cdot \sigma_{\nu(\Gamma,w)} + \sqrt{n}\cdot(\xi_{\nu(\Gamma,w)} - \tau_W)}{\sigma_W}\right)$$

□

Theorem 3.3 (Design Sensitivity for the Variance-Based Sensitivity Model)

Proof. As an overview of the proof, we will first show that $\sigma_{\nu_{vbm}(R^2,w)} < \infty$. Then, we will invoke the results from Theorem 3.1 to solve for the R^2 parameter for which $\xi_{\nu_{vbm}(R^2,w)}$ is equal to τ_W. In order for $\sigma_{\nu_{vbm}(R^2,w)} < \infty$, the endpoints of the range of potential point estimates under the variance-based sensitivity model must have finite variance. This is a secondary result, proven in Huang and Pimentel (2022), Theorem 3.2. We assume researchers are using inverse propensity score weights (for strategies on generalization to other estimation approaches see Soriano et al. (2021)). The result follows from using a Z-estimation framework, and showing that the vector of parameters $\hat{\theta}$ converge in distribution to a Normal distribution with finite variance. The original proof is done with respect to a general missingness indicator. We will provide the set-up for completeness here, written with respect to the ATT setting.

To begin, we define μ_w as the expectation of the weights:

$$\mu_w = \mathbb{E}((1-Z)w) \equiv \mathbb{E}((1-Z)\cdot(1+\exp(-\beta X))),$$

where the second equivalence arises from the assumption that we are using inverse propensity score weights. Then, we define μ as the average, re-weighted outcome across the control units:

$$\mu = \frac{\mathbb{E}((1-Z)Y(1+\exp(-\beta^\top X)))}{\mu_w}.$$

Define $\mu_w^2 = \mathbb{E}((1-Z)w^2)$ and $\sigma_Y^2 = \mathbb{E}((1-Z)Y^2)$ as the second moment of the weights and the outcomes, respectively. Then, we define the vector $\theta = (\mu, \mu_w, \beta, \mu_w^2, \mu_Y, \mu_Y^2)^\top \in \Theta$. Define the function $Q : 0,1 \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d+5}$, where for $t = (1-z, x^\top, y) \in \{0,1\} \times \mathbb{R}^d \times \mathbb{R}$:

$$Q(t \mid \theta) = \begin{pmatrix} Q_1(t|\theta) \\ Q_2(t|\theta) \\ Q_3(t|\theta) \\ Q_4(t|\theta) \\ Q_5(t|\theta) \\ Q_6(t|\theta) \end{pmatrix} := \begin{pmatrix} \left((1-z) - \frac{\exp(\beta^\top x)}{1+\exp(\beta^\top x)}\right) x \\ \mu_w - (1-z)\left(1+\exp(-\beta^\top x)\right) \\ \mu_w \mu - (1-z)y\left(1+\exp(-\beta^\top x)\right) \\ \mu_w^2 - (1-z)\left(1+\exp(-\beta^\top x)\right)^2 \\ \mu_y - (1-z)y \\ \mu_y^2 - (1-z)y^2 \end{pmatrix} \tag{B.1}$$

Finally, we define $\Phi(\theta)$ as:

$$\Phi(\theta) = \int Q(t|\theta) d\mathbb{P}(t),$$

where $T = (1-Z, X^\top, (1-Z)Y)^\top \sim \mathbb{P}$, where $\mathbb{P}$ represents the true distribution generating the data. It is simple to see that $\Phi(\theta^*) = 0$, when θ^* is equal to the true parameter values.

Then, the Z-estimates $\hat{\theta}$ satisfy the following estimating equations :

$$\Phi_n(\hat{\theta}) : = \frac{1}{n}\sum_{i=1}^{n} Q(T_i|\hat{\theta}) \tag{B.2}$$

$$= \begin{pmatrix} \left(\frac{1}{n}\sum_{i=1}^{n}(1-Z_i) - \frac{\exp(\hat{\beta}^\top X_i)}{1+\exp(\hat{\beta}^\top \mathbf{X}_i)}\right) X_i \\ \hat{\mu}_w - \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\left(1+\exp(-\hat{\beta}^\top X_i)\right) \\ \hat{\mu}_w\mu - \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)Y_i\left(1+\exp(-\hat{\beta}^\top X_i)\right) \\ \hat{\mu}_w^2 - \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\left(1+\exp(-\hat{\beta}^\top X_i)\right)^2 \\ \hat{\mu}_y - \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)Y_i \\ \hat{\mu}_y^2 - \frac{1}{n}\sum_{i=1}^{n}((1-Z_i)Y_i^2) \end{pmatrix} = 0 \tag{B.3}$$

We define $\Sigma := \mathbb{E}(Q(t \mid \theta)Q(t \mid \theta)^\top)$. We invoke the following regularity conditions:

Assumption B.1 (Regularity Conditions)**.** Assume that the parameter space Θ is compact, and that θ is in the interior of Θ. Furthermore, (Y, X) satisfies the following:

1. $\mathbb{E}(Y^4) < \infty$
2. $\det\left(\mathbb{E}\left(\frac{\exp(\beta^\top X)}{(1+\exp(\beta^\top X))^2} XX^\top\right)\right) > 0$
3. $\forall$ compact subsets $S \subset \mathbb{R}^d$, $\mathbb{E}(\sup_{\beta\in S}\exp(\beta^\top X)) < \infty$

Under these assumptions, we can apply results from Huang and Pimentel (2022, Theorem 3.2) to show that the parameters $\hat{\theta}$ converge in distribution to $N(\theta, \dot{\Phi}_0^{-1}\Sigma\dot{\Phi}_0)$. In particular, the first and second regularity conditions are necessary to show that Σ is finite, and the last regularity condition is necessary for the convergence in distribution. Applying the Delta method, it follows that $\sigma_{\nu_{vbm}(R^2,w)} < \infty$. These are the same regularity conditions necessary for standard weighted estimators to converge in distribution to a Normal distribution. We can think of the standard weighted estimator as a special case, in which $R^2 = 0$, and thus, we must only consider the first three elements in Q and θ.

Because $\sigma_{\nu_{vbm}(R^2,w)} < \infty$, the results for the theorem follow almost immediately from Theorem 3.1. Recall that the design sensitivity is defined as the minimum parameter value for a set of sensitivity models ν for which $\xi_{\nu(\Gamma,w)} > \tau_W$ (i.e., when $\sqrt{n}(\xi_{\nu(\Gamma,w)} - \tau_W) = 0$). To solve for the design sensitivity for the variance-based sensitivity model, we begin by noting that the error term $\xi_{\nu_{vbm}(R^2,w)}$ is equal to the maximum bias for a set of sensitivity models. Following Huang and Pimentel (2022), the maximum asymptotic bias that can occur is given by:

$$\begin{aligned} &\xi_{\nu_{vbm}(R^2,w)} \\ &\quad := \max_{\tilde{w}\in\nu_{vbm}(R^2)} \text{Bias}(\tau_W \mid \tilde{w}) \\ &\quad = \sqrt{1-\text{cor}(w, Y \mid Z=0)^2} \cdot \sqrt{\frac{R^2}{1-R^2} \cdot \text{var}(w \mid Z=0) \cdot \text{var}(Y \mid Z=0)}. \end{aligned}$$

To solve for $\tilde{R}^2$, we set $\xi_{\nu_{vbm}(R^2,w)}$ equal to τ_W:

$$\begin{aligned}
\sqrt{1-\text{cor}(w,Y \mid Z=0)^2}\cdot\sqrt{\frac{\tilde{R}^2}{1-\tilde{R}^2}\cdot\text{var}(w \mid Z=0)\cdot\text{var}(Y \mid Z=0)} &= \tau_W\\
\frac{\tilde{R}^2}{1-\tilde{R}^2} = \frac{1}{1-\text{cor}(w,Y \mid Z=0)^2} &\cdot \frac{\tau_W^2}{\text{var}(w \mid Z=0) \mid \text{var}(Y \mid Z=0)}\\
\tilde{R}^2 = \frac{a^2}{1+a^2} \text{ where } a^2 = \frac{1}{1-\text{cor}(w,Y \mid Z=0)^2} &\cdot \frac{\tau_W^2}{\text{var}(w \mid Z=0)\cdot\text{var}(Y \mid Z=0)}
\end{aligned}$$

□

Theorem 3.4 (Design Sensitivity for the Marginal Sensitivity Model)

Proof. For a fixed value of Λ, we reject the null hypothesis that $\tau = 0$ if

$$\begin{aligned}
&\lim_{n\to\infty}\min_{\tilde{w}\in\nu_{msm}(\Lambda,w)}\hat{\tau}_{\tilde{w}} > 0\\
\Longleftrightarrow\ &\lim_{n\to\infty}\left[\frac{1}{\sum_{i=1}^n Z_i}\sum_{i=1}^n Y_i Z_i - \max_{\tilde{w}\in\nu_{msm}(\Lambda,w)}\frac{\sum_{i=1}^n \tilde{w}_i Y_i(1-Z_i)}{\sum_{i=1}^n \tilde{w}_i(1-Z_i)}\right] > 0\\
\Longleftrightarrow\ &\lim_{n\to\infty}\frac{1}{\sum_{i=1}^n Z_i}\sum_{i=1}^n Y_i Z_i > \lim_{n\to\infty}\max_{\tilde{w}\in\nu_{msm}(\Lambda,w)}\frac{\sum_{i=1}^n \tilde{w}_i Y_i(1-Z_i)}{\sum_{i=1}^n \tilde{w}_i(1-Z_i)}.
\end{aligned}$$

Therefore, we can compute the design sensitivity $\tilde{\Lambda}$ by finding Λ such that

$$\lim_{n\to\infty}\frac{1}{\sum_{i=1}^n Z_i}\sum_{i=1}^n Y_i Z_i = \lim_{n\to\infty}\max_{\tilde{w}\in\nu_{msm}(\Lambda,w)}\frac{\sum_{i=1}^n \tilde{w}_i Y_i(1-Z_i)}{\sum_{i=1}^n \tilde{w}_i(1-Z_i)}. \tag{B.4}$$

The term on the left hand side of the estimating equation (B.4) is the observed data sample mean of $Y(1)$ and is equal to $\mathbb{E}\left[Y(1) \mid Z=1\right]$ by the law of large numbers. We focus on showing that the right hand side limit exists and computing its value.

For notational simplicity, let $\hat{\mu}_0(\tilde{w}) := \frac{\sum_{i=1}^n \tilde{w}_i Y_i(1-Z_i)}{\sum_{i=1}^n \tilde{w}_i(1-Z_i)}$. Without loss of generality, let the first m units be control units such that $Z_1 = \cdots = Z_m = 0, Z_{m+1} = \cdots = Z_n = 1$, where $1 \leq m < n$. Additionally, let Y be ordered from largest to smallest such that $Y_1 \geq Y_2 \geq \cdots \geq Y_m$ and let $Y_i = 0$ for $i \notin \{1,\ldots,m\}$. Then, by Proposition 2 from Zhao et al. (2019),

$$\max_{\tilde{w}\in\nu_{msm}(\Lambda,w)} \hat{\mu}_0(\tilde{w}) = \max_{a\in\{0,\dots,m\}} \frac{\sum\limits_{i=\min\{a,1\}}^{a} \Lambda w_i Y_i + \sum\limits_{i=\min\{a+1,m+1\}}^{\max\{a+1,m\}} \frac{1}{\Lambda} w_i Y_i}{\sum\limits_{i=\min\{a,1\}}^{a} \Lambda w_i + \sum\limits_{i=\min\{a+1,m+1\}}^{\max\{a+1,m\}} \frac{1}{\Lambda} w_i}$$

$$= \max_{c\in\mathbb{R}} \frac{\sum\limits_{i=1}^{m} \Lambda \mathbb{1}\{Y_i \geq c\} w_i Y_i + \sum\limits_{i=1}^{m} \frac{1}{\Lambda}\mathbb{1}\{Y_i < c\} w_i Y_i}{\sum\limits_{i=1}^{m} \Lambda \mathbb{1}\{Y_i \geq c\} w_i + \sum\limits_{i=1}^{m} \frac{1}{\Lambda}\mathbb{1}\{Y_i < c\} w_i}.$$

The following lemma allows us to state the limit of $\max\limits_{\tilde{w}\in\nu_{msm}(\Lambda,w)} \hat{\mu}_0(\tilde{w})$.

Lemma B.1 (Limit of $\max\limits_{\tilde{w}\in\nu_{msm}(\Lambda,w)} \hat{\mu}_0(\tilde{w})$)**.**
Under Assumption 3.2 and $\mathbb{E}\left(Y_i^2\right), \mathbb{E}\left(w_i^2\right) < \infty$,

$$\max_{\tilde{w}\in\nu_{msm}(\Lambda,w)} \hat{\mu}_0(\tilde{w}) = \max_{c\in\mathbb{R}} \frac{\sum\limits_{i=1}^{m} \Lambda \mathbb{1}\{Y_i \geq c\} w_i Y_i + \sum\limits_{i=1}^{m} \frac{1}{\Lambda}\mathbb{1}\{Y_i < c\} w_i Y_i}{\sum\limits_{i=1}^{m} \Lambda \mathbb{1}\{Y_i \geq c\} w_i + \sum\limits_{i=1}^{m} \frac{1}{\Lambda}\mathbb{1}\{Y_i < c\} w_i} \tag{B.5}$$

$$\xrightarrow{p} \max_{c\in\mathbb{R}} \frac{\Lambda\mathbb{E}\left[\mathbb{1}\{Y(0) \geq c\} wY(0)|Z=0\right] + \frac{1}{\Lambda}\mathbb{E}\left[\mathbb{1}\{Y(0) < c\} wY(0)|Z=0\right]}{\Lambda\mathbb{E}\left[\mathbb{1}\{Y(0) \geq c\} w|Z=0\right] + \frac{1}{\Lambda}\mathbb{E}\left[\mathbb{1}\{Y(0) < c\} w|Z=0\right]} \tag{B.6}$$

Proof. We break $\max\limits_{\tilde{w}\in\nu_{msm}(\Lambda,w)} \hat{\mu}_0(\tilde{w})$ into four functions and show that each function converges uniformly to its corresponding expectation. As a result, $\max\limits_{\tilde{w}\in\nu_{msm}(\Lambda,w)} \hat{\mu}_0(\tilde{w})$ converges to its expectation. First, let

$$\max_{\tilde{w}\in\nu_{msm}(\Lambda,w)} \hat{\mu}_0(\tilde{w}) = \max_{c\in\mathbb{R}} \frac{\Lambda\overline{g}_1(c) + \frac{1}{\Lambda}\overline{g}_2(c)}{\Lambda\overline{g}_3(c) + \frac{1}{\Lambda}\overline{g}_4(c)},$$

where $\overline{g}_t(c) = \frac{1}{m}\sum\limits_{i=1}^{m} g_t(Y_i, w_i; c)$ for $t \in \{1,2,3,4\}$ and

1. $g_1(Y_i, w_i; c) = \mathbb{1}\{Y_i \geq c\} w_i Y_i$
2. $g_2(Y_i, w_i; c) = \mathbb{1}\{Y_i < c\} w_i Y_i$
3. $g_3(Y_i, w_i; c) = \mathbb{1}\{Y_i \geq c\} w_i$
4. $g_4(Y_i, w_i; c) = \mathbb{1}\{Y_i < c\} w_i$.

We show that the class of functions $\mathcal{F} = \{g_1(y, w; c) : c \in \mathbb{R} \cup \{-\infty, \infty\}\}$, each element of which maps (Y_i, w_i) to the real line, is Glivenko-Cantelli and therefore $\overline{g}_1(c)$ converges uniformly to its expectation, $\mathbb{E}\left[\mathbb{1}\left\{Y(0) \geq c\right\} wY(0)|Z = 0\right]$. A similar result can be shown for $\overline{g}_2(c)$, $\overline{g}_3(c)$, and $\overline{g}_4(c)$. To clarify our exposition, we focus initially on the case in which the distribution of Y is continuous.

Let P be the probability distribution from which $(Y_1, w_1), ..., (Y_m, w_m)$ is a random sample and let F be the cdf of Y. Choose any $\epsilon > 0$. By our assumptions and the Cauchy-Schwarz theorem, $\mathbb{E}|Y_i w_i| < \infty$; therefore, there exist constants M_ϵ^- and M_ϵ^+ sufficiently large such that $\mathbb{E}\left[|Y_i w_i| \mathbb{1}\{Y_i < -M_\epsilon^-\}\right] < \epsilon$ and $\mathbb{E}\left[|Y_i w_i| \mathbb{1}\{Y_i > M_\epsilon^+\}\right] < \epsilon$. Define $p_\epsilon^- = F(-M_\epsilon^-)$ and $p_\epsilon^+ = F(M_\epsilon^+)$, and let $\Delta_\epsilon^- = F(0) - p_\epsilon^-$ and $\Delta_\epsilon^+ = p_\epsilon^+ - F(0)$. Finally, choose any $k \in \mathbb{N}$.

Define functions $f_j^- = \mathbb{1}\left\{Y_i \geq F^{-1}\left(p_\epsilon^- + \frac{j\Delta_\epsilon^-}{k}\right)\right\} w_i Y_i$ and $f_j^+ = \mathbb{1}\left\{Y_i \geq F^{-1}\left(p_\epsilon^+ - \frac{j\Delta_\epsilon^+}{k}\right)\right\} w_i Y_i$ for $j \in \{0, \dots, k\}$. If Y_i has only nonnegative (nonpositive) support, the quantities $M_\epsilon^-, p_\epsilon^-, \Delta_\epsilon^-, f_i^-$ $(M_\epsilon^+, p_\epsilon^+, \Delta_\epsilon^+, f_i^+)$ are unnecessary. In addition, let $\underline{f} = g_1(Y_i, w_i; -\infty) = Y_i w_i$, $\overline{f} = g_1(Y_i, w_i; \infty) = 0$, and note that $f_k^- = f_k^+ = \mathbb{1}\{Y_i \geq 0\} w_i Y_i$. For any two real-valued functions $\ell(Y_i, w_i), u(Y_i, w_i)$ such that $\ell(Y_i, w_i), u(Y_i, w_i)$ for all (Y_i, w_i), define the bracket $[\ell, u] = \{f \in \mathcal{F} : \ell(Y_i, w_i) \leq f(Y_i, w_i) \leq u(Y_i, w_i)\}$ as the set of all functions contained between them. Then the brackets

$$\begin{aligned}
&\left[\underline{f}, f_0^-\right], \left[f_0^-, f_1^-\right], \left[f_1^-, f_2^-\right], \dots, \left[f_{k-1}^-, f_k^-\right] \text{ and} \\
&\left[\overline{f}, f_0^+\right], \left[f_0^+, f_1^+\right], \left[f_1^+, f_2^+\right], \dots, \left[f_{k-1}^+, f_k^+\right]
\end{aligned}$$

form a coverage of the function class $\mathcal{F}$ since every function in $\mathcal{F}$ belongs to at least one bracket.

Now that we have a set of $2(k+1)$ brackets $[\ell, u]$ that cover $\mathcal{F}$, we show that they are ϵ-brackets in the sense that $P(u - \ell) < \epsilon$ for bracket where $Pf = \int f dP$. Let $C = E(w_i^2) < \infty$.

$$P|f_0^- - \underline{f}| = \mathbb{E}\left|\mathbb{1}\left\{Y_i < M_\epsilon^-\right\} w_i Y_i\right| < \epsilon \qquad \text{and} \tag{B.7}$$

$$P|f_0^+ - \overline{f}| = \mathbb{E}\left|\mathbb{1}\left\{Y_i > M_\epsilon^+\right\} w_i Y_i\right| < \epsilon \tag{B.8}$$

by our initial choices of M_ϵ^- and M_ϵ^+. For $j = 1, \dots, k$,

$$\begin{aligned}
P|f_j^- - f_{j-1}^-| &\leq \mathbb{E}\left[\mathbb{1}\left\{Y_i \in \left[F^{-1}\left(p_\epsilon^- + \frac{(j-1)\Delta_\epsilon^-}{k}\right), F^{-1}\left(p_\epsilon^- + \frac{j\Delta_\epsilon^-}{k}\right)\right)\right\} w_i M_\epsilon^-\right] \\
&\leq C M_\epsilon^- \cdot \Pr\left(Y_i \in \left[F^{-1}\left(p_\epsilon^- + \frac{(j-1)\Delta_\epsilon^-}{k}\right), F^{-1}\left(p_\epsilon^- + \frac{j\Delta_\epsilon^-}{k}\right)\right)\right) \\
&= \frac{C M_\epsilon^- \Delta_\epsilon^-}{k} \leq \frac{C M_\epsilon^-}{k}.
\end{aligned} \tag{B.9}$$

The second line follow from the Cauchy-Schwarz inequality. Similarly,

$$\begin{aligned}P|f_j^+ - f_{j-1}^+| &\leq \mathbb{E}\left[\mathbb{1}\left\{Y_i \in \left[F^{-1}\left(p_\epsilon^+ - \frac{j\Delta_\epsilon^+}{k}\right), F^{-1}\left(p_\epsilon^+ - \frac{(j-1)\Delta_\epsilon^+}{k}\right)\right)\right\} w_i M_\epsilon^+\right] \\ &\leq CM_\epsilon^+ \cdot \Pr\left(Y_i \in \left[F^{-1}\left(p_\epsilon^+ - \frac{j\Delta_\epsilon^+}{k}\right), F^{-1}\left(p_\epsilon^+ - \frac{(j-1)\Delta_\epsilon^+}{k}\right)\right)\right) \\ &= \frac{CM_\epsilon^+\Delta_\epsilon^+}{k} \leq \frac{CM_\epsilon^+}{k}. \end{aligned} \tag{B.10}$$

Since k was chosen arbitrarily, we can select a value large enough such that $\frac{CM_\epsilon^-}{k}, \frac{CM_\epsilon^+}{k} < \epsilon$. Therefore, by Theorem 19.4 from Van der Vaart (2000), since the bracketing numbers are finite for every $\epsilon > 0$, the class of functions $\mathcal{F}$ is P-Glivenko-Cantelli. Since $\mathcal{F}$ is Glivenko-Cantelli, by Theorem 19.1 from Van der Vaart (2000), uniform convergence holds.

If the distribution of Y is not continuous, the above argument works until statements (B.9) and (B.10), which may not hold because the probability that Y_i lies in a small region may still be large if a point probability mass is contained within it. We can modify the argument to account for such point masses as follows. Consider the set of points $\mathcal{Y}$ for which Y_i has a point probability mass greater than or equal to $\epsilon/2$; this set must be finite in cardinality. Increase M_ϵ^+ and M_ϵ^- so that $-M_\epsilon^- < y < M_\epsilon^+$ for all $y \in \mathcal{Y}$. Choose any $y_0 \in \mathcal{Y}$, and suppose without loss of generality that $y_0 < 0$. We can split any bracket $[f_j^-, f_{j+1}^-]$ that contains $g(Y_i, w_i; y_0)$ into the following two brackets:

$$\left[f_j^-,\ \mathbb{1}\left\{Y_i \geq y_0\right\} w_i Y_i\right] \qquad \text{and} \qquad \left[\mathbb{1}\left\{Y_i > y_0\right\} w_i Y_i,\ f_{j+1}^-\right]$$

Since the largest remaining probability point masses are all smaller than $\epsilon/2$, it is now possible to choose k sufficiently large that each bracket $[\ell, u]$ satisfies $P|u - \ell| < \epsilon$.

If Lemma B.1 is not true, then assume that there is some value ϵ for which we can always find some n such that (B.5) and (B.6) are different by at least ϵ. Since each of the functions $\overline{g}_1(c)$, $\overline{g}_2(c)$, $\overline{g}_3(c)$, and $\overline{g}_4(c)$ differ from their expectations by at most ϵ_1, ϵ_2, ϵ_3, and ϵ_4, respectively, we can construct a new ϵ which upper bounds the difference between (B.5) and (B.6). If there is not uniform convergence of (B.5) and (B.6), then the two terms have to be different by at least ϵ. Therefore, Lemma B.1 follows by contradiction. □

Noting that (B.6) can equivalently be written with indicator functions in terms of the conditional CDF of $Y(0)$ given $Z = 0$, Theorem 3.4 follows from Lemma B.1. □

Corollaries to Theorem 3.4 for Trimmed and Augmented Weighting Estimators

Corollary B.1 (Design Sensitivity for the Marginal Sensitivity Model for Trimming)**.** Define $G_{\theta,m}(Y)$ as the following function:

$$G_{\theta,m}(Y) = \begin{cases} 1 & \text{if } Y \geq F^{-1}_{Y|w<m,Z=0}(1-\theta) \\ 0 & \text{if } Y < F^{-1}_{Y|w<m,Z=0}(1-\theta) \end{cases},$$

where $F_{y|x}$ represents the population c.d.f. of y given x under the favorable situation and m represents the trimming cutoff. Let $\tilde{\Lambda}$ be any solution to the following estimating equation:

$$\begin{aligned}&\mathbb{E}[wY(0) \mid w < m, Z = 0] + \tau_{trim} = \\ &\sup_{\theta \in [0,1]} \frac{\Lambda \mathbb{E}\left[wY(0) \cdot G_{\theta,m}(Y(0)) \mid w < m, Z = 0\right] + \frac{1}{\Lambda}\mathbb{E}\left[wY(0) \cdot (1 - G_{\theta,m}(Y(0))) \mid w < m, Z = 0\right]}{\Lambda \mathbb{E}\left[w \cdot G_{\theta,m}(Y(0)) \mid w < m, Z = 0\right] + \frac{1}{\Lambda}\mathbb{E}\left[w \cdot (1 - G_{\theta,m}(Y(0))) \mid w < m, Z = 0\right]},\end{aligned}$$

where $\tau_{trim} = \mathbb{E}\left[Y(1) - Y(0) \mid Z = 1, w < m\right]$. Then $\widetilde{\Lambda}$ is the design sensitivity.

Proof. The proof of Corollary B.1 is equivalent to the proof of Theorem 3.4 after removing units with $w_i \geq m$. □

Corollary B.2 (Design Sensitivity for the Marginal Sensitivity Model for Augmentation)**.** Define $\tilde{\Lambda}$ as any solution to the following estimating equation (where $F_{y|x}$ represents the population cdf of y given x under the favorable situation):

$$\begin{aligned}&\mathbb{E}[we \mid Z = 0] + \tau \\ &= \sup_{\theta \in [0,1]} \frac{\Lambda \mathbb{E}\left[we\mathbb{1}\left\{e \geq F^{-1}_{e|Z=0}(1-\theta)\right\} \mid Z = 0\right] + \frac{1}{\Lambda}\mathbb{E}\left[we\mathbb{1}\left\{e < F^{-1}_{e|Z=0}(1-\theta)\right\} \mid Z = 0\right]}{\Lambda \mathbb{E}\left[w\mathbb{1}\left\{e \geq F^{-1}_{e|Z=0}(1-\theta)\right\} \mid Z = 0\right] + \frac{1}{\Lambda}\mathbb{E}\left[w\mathbb{1}\left\{e < F^{-1}_{e|Z=0}(1-\theta)\right\} \mid Z = 0\right]},\end{aligned}$$

where $e := Y - g(X)$ are the residuals from an arbitrary outcome model g. Then $\widetilde{\Lambda}$ is the design sensitivity.

Proof. The proof of Corollary B.2 is equivalent to the proof of Theorem 3.4 after replacing the outcomes with residuals. □

Theorem 3.5 (Impact of Augmentation on Design Sensitivity)

Define $e := Y - g(X)$ as the residual from an arbitrary outcome model g used to augment a weighted estimate. Then, for the variance-based sensitivity model, the design sensitivity from an augmented weighted estimators will be greater than the design sensitivity for a standard weighted estimator if the following holds:

$$\text{var}(e \mid Z = 0) \leq \frac{1 - \text{cor}(w, Y \mid Z = 0)^2}{1 - \text{cor}(w, e \mid Z = 0)^2} \cdot \text{var}(Y \mid Z = 0)$$

Proof. To begin, we will first derive the design sensitivity for the variance-based sensitivity model for augmented weighted estimators. If we treat the model $g(X)$ as fixed, then it is simple to show that under the same regularity assumptions as the ones invoked in Theorem 3.3 (i.e., Assumption B.1), $\sigma^{aug}_{\nu(R^2,w)} < \infty$. In particular, we can apply the same proof, but substitute the residuals for the outcomes. The regularity conditions effectively state that the fourth moment of the residuals must be finite.

Following Huang (2022) (Theorem 5.1), note that the maximum asymptotic bias that can occur for an augmented weighted estimator is:

$$\begin{aligned}
\xi^{aug}_{\nu_{vbm}(R^2,w)} &:= \max_{\tilde{w} \in \nu_{vbm}(R^2,w)} \text{Bias}(\tau_W^{aug} \mid \tilde{w}) \\
&= \sqrt{1 - \text{cor}(w, e \mid Z = 0)^2} \cdot \sqrt{\frac{R^2}{1-R^2} \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(e \mid Z = 0)}.
\end{aligned}$$

Then, because $\sigma^{aug}_{\nu(R^2,w)} < \infty$, following Theorem 3.3, the design sensitivity can be algebraically solved for:

$$\tilde{R}^2_{aug} = \frac{b^2}{1+b^2} \text{ where } b^2 = \frac{1}{1 - \text{cor}(w, e \mid Z = 0)^2} \cdot \frac{\tau^2_{aug}}{\text{var}(w \mid Z = 0) \cdot \text{var}(e \mid Z = 0)}$$

To compare $\tilde{R}^2_{aug}$ and $\tilde{R}^2$, we can re-write $\tilde{R}^2_{aug}$ as follows:

$$\begin{aligned}
\tilde{R}^2_{aug} &= \frac{b^2}{1+b^2} \\
&= \frac{\frac{1}{1-\text{cor}(w,e|Z=0)^2} \cdot \frac{\tau^2_{aug}}{\text{var}(w|Z=0)\cdot\text{var}(e|Z=0)}}{1 + \frac{1}{1-\text{cor}(w,e|Z=0)^2} \cdot \frac{\tau^2_{aug}}{\text{var}(w|Z=0)\cdot\text{var}(e|Z=0)}} \\
&= \frac{\tau^2_{aug}}{(1 - \text{cor}(w, e \mid Z = 0))^2 \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(e \mid Z = 0) + \tau^2_{aug}}
\end{aligned}$$

Then:

$$\begin{aligned}
\frac{\tilde{R}^2_{aug}}{\tilde{R}^2} &= \frac{(1 - \text{cor}(w, Y \mid Z = 0)^2) \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(Y \mid Z = 0) + \tau^2_W}{(1 - \text{cor}(w, e \mid Z = 0)^2) \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(e \mid Z = 0) + \tau^2_{aug}} \cdot \frac{\tau^2_{aug}}{\tau^2_W} \\
&= \frac{\tau^2_{aug}}{\tau^2_W} \cdot \frac{(1 - \text{cor}(w, Y \mid Z = 0)^2) \cdot \text{var}(w) \cdot \text{var}(Y) + \tau^2_W}{(1 - \text{cor}(w, e \mid Z = 0)^2) \cdot \text{var}(w) \cdot \text{var}(e) + \tau^2_{aug}}
\end{aligned}$$

Because we are in the favorable setting, in which there is no omitted confounding, the weighted estimator will recover the estimand (i.e., the ATT) consistently (i.e., $\hat{\tau}_W \xrightarrow{p} \tau_W \equiv \tau$). Similarly, because the augmented weighted estimator is doubly robust, augmenting will also recover the estimand consistently (i.e., $\hat{\tau}_{aug} \xrightarrow{p} \tau_{aug} \equiv \tau$), regardless of the outcome model. Thus, $\tau_W = \tau_{aug}$. Then, the above is greater than 1 if the following criteria holds:

$$\begin{aligned}
(1 - \text{cor}(w, e)^2) \cdot \text{var}(e \mid Z = 0) &\leq (1 - \text{cor}(w, Y \mid Z = 0)^2) \cdot \text{var}(Y \mid Z = 0) \\
\text{var}(e \mid Z = 0) &\leq \frac{1 - \text{cor}(w, Y \mid Z = 0)^2}{1 - \text{cor}(w, e \mid Z = 0)^2} \cdot \text{var}(Y \mid Z = 0)
\end{aligned}$$

□

Theorem 3.6 (Impact of Trimming Weights on Design Sensitivity)

Define some cutoff m such that weights above the cutoff are trimmed. Furthermore, assume the trimmed weights are centered at mean 1 and the projection of the trimmed, ideal weights are centered on the trimmed, estimated weights. Then, for the variance-based sensitivity model, if the following holds:

$$\underbrace{\frac{\text{var}(w \mid w < m, Z = 0)}{\text{var}(w \mid Z = 0)}}_{\text{(1)Variance reduction in } w} \leq \underbrace{\frac{1 - \text{cor}(w, Y \mid Z = 0)^2}{1 - \text{cor}(w, Y \mid w < m, Z = 0)^2}}_{\text{(2) Change in relationship between } w \text{ and } Y} \cdot \underbrace{\frac{\text{var}(Y \mid Z = 0)}{\text{var}(Y \mid w < m, Z = 0)}}_{\text{(3)Variance reduction in } Y},$$

the design sensitivity from a trimmed estimator will be greater than the standard weighted estimator.

Proof. Like Theorem 3.5, we will begin by deriving the design sensitivity for weighted estimators with trimming, under the variance-based sensitivity model. Furthermore, we have assumed that the trimmed weights are centered at mean 1: $\mathbb{E}(w \mid w < m) = 1$. This assumption trivially holds as long as researchers normalize the trimmed weights to be mean 1.

We will begin by deriving the maximum asymptotic bias for a trimmed weighted estimator. To begin, define the cutoff for trimming to be some threshold m, such that any observations $w \geq m$ are trimmed. Then, the estimand of interest is thus the average treatment effect, across the treated, subset to units that associated with weights $w < m$:

$$\begin{aligned} \tau^{trim} &:= \mathbb{E}(Y(1) - Y(0) \mid Z = 1, X \in \mathcal{A}) \\ &\equiv \mathbb{E}(Y(1) - Y(0) \mid Z = 1, w < m) \end{aligned}$$

Notably, the additional condition of $w < m$ is an observable condition, given the observed covariates X (as the estimated weights are a function of X). The bias for a trimmed weighted estimator is:

$$\begin{aligned} &\left|\text{Bias}(\hat{\tau}_W^{trim})\right| = \left|\mathbb{E}\left(\hat{\tau}_W^{trim}\right) - \tau^{trim}\right| \\ &\text{By conditional ignorability:} \\ &= \left|\mathbb{E}\left(\sum_{i=1}^n (1 - Z)w \cdot \mathbb{1}\{w < m\}Y\right) - \mathbb{E}\left(\sum_{i=1}^n w^*(1 - Z) \cdot \mathbb{1}\{w < m\}Y\right)\right| \\ &= \left|\mathbb{E}(wY \mid Z = 0, w < m) - \mathbb{E}(w^*Y \mid Z = 0, w < m)\right| \\ &= \left|\mathbb{E}((w - w^*) \cdot Y \mid Z = 0, w < m)\right| \end{aligned}$$

By construction, $\mathbb{E}(w \mid Z = 0, w < m) = \mathbb{E}(w^* \mid Z = 0, w < m)$:

$$\begin{aligned}
&= |\mathbb{E}((w - w^*) \cdot Y \mid Z = 0, w < m) - \mathbb{E}(w - w^* \mid Z = 0, w < m) \cdot \mathbb{E}(Y \mid Z = 0, w < m)| \\
&= |\text{cov}(w - w^*, Y \mid Z = 0, w < m)| \\
&= |\text{cor}(w - w^*, Y \mid Z = 0, w < m)| \cdot \sqrt{\text{var}(w - w^* \mid Z = 0, w < m) \cdot \text{var}(Y \mid Z = 0, w < m)} \\
&= |\text{cor}(w - w^*, Y \mid Z = 0, w < m)| \cdot \sqrt{\text{var}(w \mid Z = 0, w < m) \cdot \frac{R^2}{1 - R^2} \text{var}(Y \mid Z = 0, w < m)}.
\end{aligned}$$

The last line follows from the fact the projection of the trimmed, ideal weights in the observed covariate space of X are centered on the trimmed, estimated weights. For intuition, first define an indicator $V := \mathbb{1}\{w > m\}$. Then, the trimmed, ideal weights can be written as $w^* \cdot V$. Similarly, the trimmed, estimated weights can be written as $w \cdot V$. Then, if the projection of the ideal weights in X are centered on w (a condition that trivially is met when using inverse propensity score weights), it follows immediately that $\mathbb{E}(w^* \cdot V \mid X) = w \cdot V$. As a result, within the space of $w < m$, the residual error (i.e., $w^* - w$) is orthogonal to the estimated weights w.

To bound the correlation term, we apply the recursive formula of partial correlation:

$$\begin{aligned}
-\sqrt{1 - \text{cor}(w, Y \mid Z = 0, w < m)^2} &\leq \text{cor}(w - w^*, Y \mid Z = 0, w < m) \\
&\leq \sqrt{1 - \text{cor}(w, Y \mid Z = 0, w < m)^2}
\end{aligned}$$

Then, the maximum asymptotic bias for a trimmed weighted estimator is:

$$\begin{aligned}
\xi^{trim}_{\sigma(R^2, w)} := &\max_{\tilde{w} \in \sigma(R^2)} \text{Bias}(\hat{\tau}^{trim}_W) \\
= &\sqrt{1 - \text{cor}(w, Y \mid Z = 0, w < m)^2} \cdot \\
&\sqrt{\frac{R^2}{1 - R^2} \cdot \text{var}(w \mid Z = 0, w < m) \cdot \text{var}(Y \mid Z = 0, w < m)}. \qquad \text{(B.11)}
\end{aligned}$$

Solving for the R^2 such that $\xi^{trim}_{\sigma(R^2, w)} = \tau_W$,

$$\tilde{R}^2_{trim} = \frac{c^2}{1 + c^2},$$

where

$$c^2 := \frac{1}{1 - \text{cor}(w, Y \mid Z = 0, w < m)^2} \cdot \frac{\tau_W^{2trim}}{\text{var}(w \mid w < m, Z = 0) \cdot \text{var}(Y \mid w < m, Z = 0)}.$$

Then,

$$\begin{aligned}
\frac{\tilde{R}^2_{trim}}{\tilde{R}^2} &= \frac{(1 - \text{cor}(w, Y \mid Z = 0)^2) \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(Y \mid Z = 0) + \tau_W^2}{(1 - \text{cor}(w, Y \mid Z = 0, w < m)^2) \cdot \text{var}(w \mid w < m, Z = 0) \cdot \text{var}(Y \mid w < m, Z = 0) + \tau_W^{2trim}} \cdot \frac{\tau_W^{2trim}}{\tau_W^2} \\
&= \frac{\tau_W^{2trim}}{\tau_W^2} \frac{(1 - \text{cor}(w, Y \mid Z = 0)^2) \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(Y \mid Z = 0) + \tau_W^2}{(1 - \text{cor}(w, Y \mid Z = 0, w < m)^2) \cdot \text{var}(w \mid w < m, Z = 0) \cdot \text{var}(Y \mid w < m, Z = 0) + \tau_W^{2trim}}
\end{aligned}$$

Let $\tau_W^{trim} := \tau_W + c$:

$$= \frac{(\tau_W + c)^2}{\tau_W^2} \frac{(1 - \text{cor}(w, Y \mid Z = 0)^2) \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(Y \mid Z = 0) + \tau_W^2}{(1 - \text{cor}(w, Y \mid Z = 0, w < m)^2)\text{var}(w \mid Z = 0, w < m)\text{var}(Y \mid Z = 0, w < m) + (\tau_W + c)^2}$$

In order for there to be an improvement in design sensitivity from trimming, the following must hold:

$$\begin{aligned}\frac{(\tau_W + c)^2}{\tau_W^2} \cdot \big((1 - \text{cor}(w, Y \mid Z = 0)^2) \cdot \text{var}(w \mid Z = 0) \cdot \text{var}(Y \mid Z = 0) + \tau_W^2\big) \geq \\ (1 - \text{cor}(w, Y \mid Z = 0, w < m)^2) \cdot \text{var}(w \mid Z = 0, w < m) \cdot \text{var}(Y \mid w < m, Z = 0) \\ +(\tau_W + c)^2\end{aligned}$$

Re-arranging:

$$\frac{\text{var}(w \mid Z = 0)}{\text{var}(w \mid Z = 0, w < m)} \geq \frac{1 - \text{cor}(w, Y \mid Z = 0, w < m)}{1 - \text{cor}(w, Y \mid Z = 0)^2} \cdot \left(\frac{\tau_W}{\tau_W + c}\right)^2 \cdot \frac{\text{var}(Y \mid Z = 0, w < m)}{\text{var}(Y \mid Z = 0)}$$

Because we are assuming a constant treatment effect, $c = 0$, which allows us to arrive at Equation (3.9):

$$\frac{\text{var}(w \mid Z = 0)}{\text{var}(w \mid Z = 0, w < m)} \geq \frac{1 - \text{cor}(w, Y \mid Z = 0, w < m)}{1 - \text{cor}(w, Y \mid Z = 0)^2} \cdot \frac{\text{var}(Y \mid Z = 0, w < m)}{\text{var}(Y \mid Z = 0)}$$

The results of Theorem 3.6 may be easily extended in settings when researchers are interested in a smoothed trimmed estimator. Due to the non-smoothness of trimming, traditional trimming methods ignore the uncertainty in the design stage from estimating weights and conduct inference excluding units with extreme estimated weights. Yang and Ding (2018) develop a smooth trimming estimator that weights all units continuously, assigning extremely small weights to units with large weights instead of removing them, and is asymptotically linear. Therefore, the bootstrap can be used to construct confidence intervals. Design sensitivity considers the large sample limits of the weights, so design stage uncertainty is not present. Furthermore, in asymptotic settings, the smoothed and non-smoothed trimmed estimators are equivalent.

However, in settings when researchers are interested in calculating the power of a sensitivity analysis, in addition to design sensitivity, it can be helpful to consider the smoothed trimmed estimator. In particular, the (standard) trimmed estimator is non-smooth, and as a result, will not be amenable to a bootstrap-style procedure to estimating power (Yang and Ding, 2018). Following Yang and Ding (2018), we define a smoothed trimmed estimator, denoted as $\hat{\tau}_W^{smooth}$, which approximates the trimmed estimator arbitrarily well using a tuning parameter $\epsilon > 0$. Applying Theorem 3.1, the power of a sensitivity analysis for $\hat{\tau}_W^{smooth}$ is as follows:

$$\begin{aligned}
&\Pr\left(\frac{\sqrt{n}(\hat{\tau}_W^{smooth} - \xi_{\nu(\Gamma,w)}^{smooth}}{\sigma_{\nu(\Gamma,w)}^{smooth}} \geq k_\alpha\right) \\
&\quad = \Pr\left(\frac{\sqrt{n}\cdot(\hat{\tau}_W^{smooth} - \tau^{trim})}{\sigma_W^{smooth}} \geq \frac{k_\alpha \cdot \sigma_{\nu(\Gamma,w)}^{smooth} + \sqrt{n}\cdot(\xi_{\nu(\Gamma,w)}^{smooth} - \tau^{trim})}{\sigma_W^{smooth}}\right) \\
&\quad = \Pr\left(\frac{\sqrt{n}\cdot(\hat{\tau}_W^{smooth} - \tau^{trim})}{\sigma_W^{smooth}} \geq \frac{k_\alpha \cdot \sigma_{\nu(\Gamma,w)}^{smooth} + \sqrt{n}\cdot(\xi_{\nu(\Gamma,w)}^{trim} + \delta - \tau^{trim})}{\sigma_W^{smooth}}\right) \qquad \text{(B.12)} \\
&\quad \to 1 - \Phi\left(\frac{k_\alpha \cdot \sigma_{\nu(\Gamma,w)}^{smooth} + \sqrt{n}\cdot(\xi_{\nu(\Gamma,w)}^{trim} + \delta - \tau^{trim})}{\sigma_W^{smooth}}\right),
\end{aligned}$$

where δ is a function of ϵ. The expression from Equation (B.12), which introduces the δ constant follows directly from the fact that we may express the bias for the smoothed trimming estimator as a function of the bias of a standard (non-smooth) trimmed estimator and a function of ϵ:

$$\begin{aligned}
\text{Bias}(\hat{\tau}_W^{smooth}) &= \mathbb{E}(\hat{\tau}_W^{smooth}) - \tau^{trim} \\
&= \underbrace{\mathbb{E}(\hat{\tau}_W^{smooth}) - \mathbb{E}(\hat{\tau}_W^{trim})}_{(*)} + \underbrace{\mathbb{E}(\hat{\tau}_W^{trim}) - \tau^{trim}}_{\equiv \text{Bias}(\hat{\tau}_W^{trim})}
\end{aligned}$$

The first term (denoted by $(*)$) is equal to δ, which can be made arbitrarily large or small by tuning $\epsilon > 0$:

$$\begin{aligned}
&\mathbb{E}(\hat{\tau}_W^{smooth}) - \mathbb{E}(\hat{\tau}_W^{trim}) \\
=&\mathbb{E}\left(\frac{1}{\sum_{i=1}^n (1-Z)wV'}\sum_{i=1}^n (1-Z)w\cdot V'Y\right) - \\
&\mathbb{E}\left(\frac{1}{\sum_{i=1}^n (1-Z)w\mathbb{1}\{w<m\}}\sum_{i=1}^n (1-Z)w\cdot \mathbb{1}\{w<m\}Y\right) \\
=&\mathbb{E}(w(V' - \mathbb{1}\{w<m\})Y \mid Z=0) \\
=&\underbrace{\mathbb{E}(w(V' - \mathbb{1}\{w<m\})Y \mid Z=0, w<m)P(w<m \mid Z=0)}_{=0} + \\
&\mathbb{E}(w(V' - \mathbb{1}\{w<m\})Y \mid Z=0, w\geq m)P(w\geq m \mid Z=0) \\
:=&\varepsilon \cdot \mathbb{E}(wY \mid Z=0, w\geq m)P(w\geq m \mid Z=0) \\
\equiv&\delta
\end{aligned}$$

As such, the bias bound of the smoothed trimmed estimator can be written as the bias bound in Equation (B.11) and an arbitrary constant δ, which is a function of ϵ:

$$\begin{aligned}
\xi^{smooth}_{\sigma(R^2,w)} &= \max_{\tilde{w}\in\nu(\Gamma,w)} \text{Bias}(\hat{\tau}^{smooth}_W) \\
&= \max_{\tilde{w}\in\nu(\Gamma,w)} \text{Bias}(\hat{\tau}^{trim}_W) + \delta \\
&= \xi^{trim}_{\nu(\Gamma,w)} + \delta
\end{aligned}$$

Therefore, we have shown that as $n \to \infty$, for an arbitrarily small $\epsilon > 0$, the design sensitivity will be within a δ-neighborhood of the value $\tilde{\Gamma}$, for which $\xi^{trim}_{\nu(\Gamma,w)} = \tau_W$.

□

B.2 Using a Planning Sample to Estimate Design Sensitivity

Calibrating Design Sensitivity to Outcome Data with a Planning Sample

To estimate design sensitivity, researchers must posit an outcome model. One way to calibrate their priors to the existing outcome data is to utilize a planning sample. This is done by holding out part of the sample to use as a 'planning sample' (akin to a pilot sample in experimental studies). The remainder of the sample is then used for the analysis. We will refer to the holdout sample as the 'analysis sample.' We will assume in this section that researchers are randomly sampling observations from a fixed dataset to construct a planning sample. However, in cases when researchers have access to auxiliary outcome data (i.e., historical datasets), they may utilize these external datasets as the planning sample, and treat the full observational study as the analysis sample.

We will outline two approaches that researchers may use to estimate design sensitivity using a planning sample. The first approach proposes drawing a planning sample, and simply estimating the design sensitivity across the planning sample. Table B.1 outlines in more detail.

The approach outlined in Table B.1 allows researchers to calibrate the quantities needed to calculate design sensitivity using a planning sample. However, in settings where the outcome distribution may be heavy tailed, the planning sample may be unable to capture the full complexity present in the outcomes, which can result in an over-estimation of design sensitivity. In particular, this is of concern to the marginal sensitivity models, in which robustness to unmeasured confounding and the design sensitivity are often characterized by outliers (Huang and Pimentel, 2022). One alternative way to leverage a planning sample, but additionally account for more complex variation across the full dataset is by first fitting an outcome model across the planning sample units, and use this model to simulate outcomes across the units in the analysis sample. Table B.2 summarizes the procedure.

Table B.1: **Estimating Design Sensitivity using a Planning Sample**

Step 1. Fix an effect size τ.

Step 2. Estimate the weights $\hat{w}_i$ using the full sample.

Step 3. Across the units in the control group, generate a planning sample by randomly sampling n_{plan} observations. Denote the set of indices that correspond to the planning sample as $\mathcal{P}$.

Step 4. If using the variance-based sensitivity model:

a. Calculate the sample variance of the outcomes (i.e., $\widehat{\text{var}}(Y_i \mid i \in \mathcal{P})$) and the sample correlation between the outcomes and the estimated weights (i.e., $\widehat{\text{cor}}(\hat{w}_i, Y_i \mid i \in \mathcal{P})$).

b. Calculate the sample variance of the estimated weights across the full sample.

If using the marginal sensitivity model:

a. Calculate the weighted average of the control outcomes in the planning sample:

$$\hat{\mu}_0^{plan} \leftarrow \frac{\sum_{i:i\in\mathcal{P}} \hat{w}_i Y_i}{\sum_{i:i\in\mathcal{P}} \hat{w}_i}.$$

b. Generate the average treatment outcome, given a fixed τ:

$$\hat{\mu}_1 \leftarrow \tau + \frac{\sum_{i:i\in\mathcal{P}} \hat{w}_i Y_i}{\sum_{i:i\in\mathcal{P}} \hat{w}_i}.$$

Step 5. Using the components generated in Step 4, estimate the design sensitivities under the variance-based sensitivity model using Theorem 3.3 and the marginal sensitivity model using Theorem 3.4.

The procedure outlined in Table B.2 allows researchers to flexibly calibrate design sensitivity using a planning sample. The fitted outcome model can be of arbitrary specification, and researchers can leverage flexible, black box machine learning models to estimate $Y_i(0)$. To simulate the noise ϵ_i^* in Step 5-(b), we currently assume the errors are normally distributed. However, researchers may relax this assumption and posit any, arbitrary distribution for the residuals, using the residuals across the planning sample to calibrate the necessary parameters for the distribution.

Table B.2: **Estimating Design Sensitivity using a Planning Sample and Simulated Outcomes**

Step 1. Fix an effect size τ.

Step 2. Estimate the weights $\hat{w}_i$ using the full sample.

Step 3. Across the units in the control group, generate a planning sample by randomly sampling n_{plan} observations. Denote the set of indices that correspond to the planning sample as $\mathcal{P}$.

Step 4. Across the planning sample $\mathcal{P}$, fit an outcome model $\hat{g}_{\mathcal{P}}(X_i)$ for the control units. Use the residuals from the fitted outcome model to estimate the variance in the residuals (i.e., unexplained variation in the outcomes, denoted as $\hat{\sigma}^2_{e,\text{plan}}$).

Step 5. Simulate new data for the units in the analysis sample (i.e., $i \notin \mathcal{P}$) by parametrically sampling residuals from $\hat{g}_{\mathcal{P}}(X_i)$:

a. Randomly sample units across the analysis sample, with replacement. We refer to this as the *bootstrap sample*, $\mathcal{B}$.

b. For all units in $\mathcal{B}$, estimate the outcome $Y_i^*(0)$:

$$Y^*(0) \leftarrow \hat{g}_{\mathcal{P}}(X_i) + \epsilon^*,$$

where $\epsilon_i^* \sim N(0, \hat{\sigma}^2_{e,\text{plan}})$.

Step 6. Apply Steps 4 and 5 from Table B.1, but using $\mathcal{B}$ instead of the planning sample $\mathcal{P}$.

To illustrate the proposed procedure, we turn to the empirical application. For simplicity, we will focus on the setting in which researchers are interested in the impact of presidential support on support for the FARC peace deal. We draw 100 different planning samples from the data, and use Table B.2 to estimate the design sensitivity for a variety of different effect sizes across both the variance-based and marginal sensitivity models. Figure B.1 provides a visualization for the distribution of estimated design sensitivities across the different planning samples. The estimated design sensitivities using the planning sample are mostly centered around the oracle design sensitivities, calibrated using the full dataset.

Remark on Sample Boundedness. From Figure B.1, we see that as the effect size increases, the spread of estimated design sensitivities under the marginal sensitivity model from

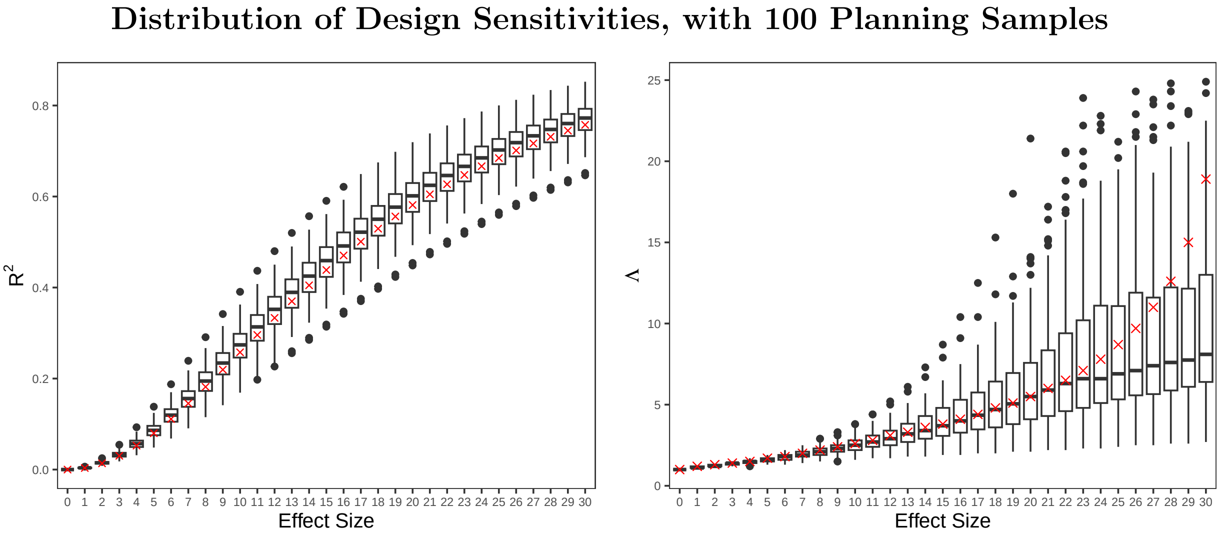


Figure B.1: Distribution of estimated design sensitivity measures across 100 different planning samples. Design sensitivities were estimated using the procedure outlined in Table B.2. The red × points denote the oracle design sensitivities, calibrated using the full dataset.

the planning samples also increases. This is likely because the marginal sensitivity model is susceptible to sample boundedness. More specifically, Huang and Pimentel (2022) highlighted that because of the inherent stabilization within the model, the marginal sensitivity model can only recover a worst-case bias bound defined by the range of observed control outcomes. As a result, as the effect size increases towards the sample bounds, $\Lambda \to \infty$. Because the range of observed control outcomes depends on the observed data, variation in the drawn planning sample can drive variation in the estimated design sensitivity for the marginal sensitivity model. The variance-based sensitivity model is not susceptible to sample boundedness; as a result, the estimated design sensitivities across different planning samples remain relatively stable, even as the effect size increases.

Both approaches proposed in the following subsection provide researchers with a way to estimate design sensitivity using a planning sample. We have also illustrated that the design sensitivities estimated from a planning sample are similar to the design sensitivities estimated using the full dataset. However, that what is arguably most important in practice is that the *relative* estimates of design sensitivity from different design choices, such as using trimmed weights or an augmented weighted estimator, stably inform the optimal design choice. See Section B.2 for more details.

Estimating Design Sensitivity for Augmentation, without Prior Specification of Outcome Model

We consider two settings for which researchers may estimate design sensitivity for augmentation. First, we consider the setting in which researchers are assessing whether or not they should fit an outcome model to begin with to perform augmentation. To estimate the impact of augmentation on design sensitivity without a prior specification of an outcome model, researchers can instead construct a *proxy outcome model* that explains a fixed r^2 of the variation in the outcomes. They can then estimate the design sensitivity. If they find that an extremely high r^2 value is needed—i.e., they must explain a large percentage of the variation in the outcomes—for improvements in design sensitivity, this can be infeasible to do in practice, and choose to not augment their estimation. Table B.3 summarizes the procedure.

To make this more concrete, we can consider the empirical application. Assume first that researchers are interested in optimizing for robustness with respect to an average error (i.e., the variance-based sensitivity model). From Figure 3.2, we see that even if researchers were to estimate an outcome model that could explain 50% of the variation, the design sensitivity for the variance-based sensitivity model would only improve slightly. In contrast, if researchers are worried about potential, worst-case confounding (i.e., the marginal sensitivity model), we see that if they were to augment the weighted estimator with an outcome model that could explain even only 25% of the variation in the outcomes, the design sensitivity would improve substantially, especially in cases where the treatment effect might be relatively small.

In the second setting, researchers already have estimated an outcome model of interest *a priori*. Then, design sensitivity can be estimated in the same manner as in the standard case. However, instead of calibrating to the outcome distribution, researchers must calibrate design sensitivity to the underlying residuals.

It is worth noting that design sensitivity is usually not hurt from augmenting with an outcome model. However, employing a proxy outcome model first can help researchers determine if it is worth fitting a complex outcome model, and also better assess practical trade-offs, like whether to gather more covariate data that could feasibly help explain variation in the outcomes.

Using a Planning Sample to Improve Power in the Colombian Peace Agreement Study

Heller et al. (2009) propose randomly splitting the data in an observational study into a planning sample and an analysis sample to inform design decisions. We consider using a similar strategy for weighted observational studies to select between traditional inverse propensity score (IPW), trimmed, and augmented weighting estimators and illustrate that sample splitting can improve power for the FARC example. Table B.4 shows our best estimates of the design sensitivities for the FARC example with the presidential support treatment from following the steps outlined in Table B.1 with $\tau = 22$ and using the full sample for planning.

Table B.3: **Estimating Design Sensitivity for Augmentation using a Proxy Outcome Model**

Step 1. Set some r^2, which represents the variation explained in the outcome model.

Step 2. Draw a planning sample.

Step 3. Estimate a model $\hat{m}$ across the planning sample to generate outcomes.

Step 4. Generate an outcome across the rest of the sample not in the planning sample. Add noise to each prediction to guarantee that the outcomes are the same variance as the outcomes in the planning sample. Denote this as $\tilde{Y}$.

Step 5. Construct a proxy outcome model g^* that can explain r^2 of the outcomes.

1. Generate a scaled version of the outcomes:
$$Z^* := (Y^* - \mathbb{E}(Y^*))/sd(Y^*)$$
2. Create a proxy covariate X^* that is correlated with the generated outcomes:
$$X^* := \sqrt{r^2} \cdot Z^* + \sqrt{1 - r^2} \cdot v,$$
where v is a standard normal random variable.
3. Estimate a linear model with the proxy covariate on the outcomes
$$Y^* \sim X^*$$
4. From the linear model fit in the previous step, use the fitted values $\hat{Y}^*$ and the residuals e^*.

Step 6. Estimate the design sensitivity.

For the variance-based sensitivity model, the estimated design sensitivities are similar for the standard and augmented weighting estimators, while trimming leads to a significant improvement in design sensitivity. For the marginal sensitivity model, augmentation and trimming perform similarly in terms of design sensitivity, with both outperforming weighting alone.

In practice, we cannot use the full data to estimate the design sensitivities, as this would violate the design principle. Instead, we may use a planning sample to help calibrate our estimates of design sensitivity to the observed data. However, using a planning sample comes at a cost; in particular, if we hold out part of the data to use as a planning sample, we cannot

Table B.4: Design sensitivity estimated using full FARC data

	VBM			MSM		
Treatment	$\tilde{R}^2$	$\tilde{R}^2_{aug}$	$\tilde{R}^2_{trim}$	$\tilde{\Lambda}$	$\tilde{\Lambda}_{aug}$	$\tilde{\Lambda}_{trim}$
Presidential Support	0.63	0.63	0.76	6.46	7.09	7.13

use these observations in our analysis (i.e., we restrict ourselves to a smaller sample size).

To estimate power, we randomly split the data into planning and analysis samples and use the planning sample to estimate design sensitivities for each method and both sensitivity models according to the steps outlined in Table B.1. We then conduct sensitivity analyses for the variance-based and marginal sensitivity models under particular values of R^2 and Λ, respectively, using the selected estimation strategies and record whether or not the null hypothesis of no treatment effect is rejected at a 5% significance level. We repeat this process for 1,000 random splits of the data, estimating power as the proportion of random splits for which we reject.

We estimate the power for each combination of estimator and sensitivity model when using 10% of the FARC data with the presidential support treatment for the planning sample and the remaining 90% for the analysis sample. The results are available in Table B.5. For the marginal sensitivity model with $\Lambda = 4$ and the variance-based sensitivity model with $R^2 = 0.25$, the sensitivity analysis for each estimator rejects the null hypothesis of no effect. Additionally, implementing the method selected using the planning sample achieves near or equal to 100% power in both scenarios. Conversely, the power is near zero for each estimator with $\Lambda = 6$, rendering the choice of estimator moot. The repeated sample splits with $\Lambda = 5$ for the marginal sensitivity model and $R^2 = 0.35$ and 0.67 for the variance-based model highlight the potential gains in power from using a planning sample to make design decisions. For the former sensitivity model, the augmented weighting estimator greatly outperforms the two alternative estimators. Implementing the method selected by the sample splitting approach yields 68% power, far higher than would be achieved by using the IPW or trimmed weighting estimator. For $R^2 = 0.35$ and $R^2 = 0.67$, using a planning sample leads us to select the trimmed estimator for each sample split, maximizing power. We repeat the same exercise using 20% of the data for the planning sample.

Estimated power for analysis sample using FARC data

	Reject?			10% of Data for Planning				20% of Data for Planning			
	Full Sample			Analysis Sample			Chosen	Analysis Sample			Chosen
	IPW	Trim	Aug.	IPW	Trim	Aug.	Method	IPW	Trim	Aug.	Method
Marginal Sensitivity Model											
$\Lambda = 4$	1	1	1	0.97	1.00	1.00	0.99	0.84	1.00	1.00	0.93
$\Lambda = 5$	0	0	1	0.10	0.28	0.86	0.68	0.18	0.40	0.73	0.51
$\Lambda = 6$	0	0	0	0.01	0.00	0.02	0.02	0.03	0.02	0.06	0.04
Variance-based Sensitivity Model											
$R^2 = .25$	1	1	1	0.81	1.00	0.85	1.00	0.67	1.00	0.73	1.00
$R^2 = .35$	0	1	0	0.18	1.00	0.27	1.00	0.22	1.00	0.28	1.00
$R^2 = .67$	0	0	0	0.00	0.79	0.00	0.79	0.00	0.96	0.00	0.96

Table B.5: For values under full sample, 1 represents rejection of the null hypothesis of no effect at a 95% significance level for the corresponding estimator and sensitivity parameter value using the full FARC data, while 0 represents failure to reject. Under analysis sample, we display the proportion of rejections across repeated splits of the data as estimated power. Chosen method is the estimated power using the method selected using the planning sample.

B.3 Detailed Simulation Results

Simulation Parameters

The simulation setup is under a *favorable situation* defined as:

1. The study is free of unmeasured bias. In the case of the marginal sensitivity model, the marginal sensitivity model is satisfied with $\Lambda = 1$. In the case of the variance-based sensitivity model, $R^2 = 0$. Equivalently, for both sensitivity models $w_i^* = w_i$.

2. The null hypothesis of no treatment effect is false, and a specific alternative is true. In our case, the alternative we consider is that the data is generated by a stochastic model with a treatment effect as follows:

$$P(Z_i = 1 \mid X_i) \propto \frac{\exp(\beta_\pi X_i)}{1 + \exp(\beta_\pi X_i)}, \qquad Y_i = \beta_y X_i + \tau Z_i + u_i, \tag{B.13}$$

where $X_i \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$, and $u_i \overset{iid}{\sim} N(0, \sigma_y^2)$.

We vary the different parameters, $\{\beta_\pi, \beta_y, \tau, \sigma_y\}$. We vary τ to control the effect size, β_π for the variance in the weights, and β_y and σ_y to alter the variance in the outcomes. The correlation between the weights and the outcome depends on β_π, β_y, and σ_y. The base parameters are set as follows: $\tau = 1$, $\mu_x = 0$, $\sigma_x = 1$, $\beta_y = 1$, $\sigma_y = 1$, $\beta_\pi = 1$. For now, we will assume a constant treatment effect τ for all units i. However, we relax this assumption in Section B.3 and allow for heterogeneous treatment effects.

Drivers of Design Sensitivity

To simulate the drivers of design sensitivity for each sensitivity model, we modify $\{\beta_\pi, \beta_y, \tau, \sigma_y\}$ such that one element of the data generating process changes, while holding the others constant. We are essentially estimating the derivative of the design sensitivity, with respect to the effect size, the variance in the estimated weights, the variance in the outcomes, and the correlation between the estimated weights and the outcome. Simulation results are presented in both Table B.6 and Figure 3.1.

Treatment Effect Heterogeneity

To allow for treatment effect heterogeneity, we modify the data generating process under the favorable situation (B.13) so that it allows for the individual treatment effect for unit i to depend on its covariate value X_i:

$$Y_i = \beta_y X_i + \tau_i Z_i + u_i, \qquad \tau_i = \tau_0 + \beta_\tau X_i. \tag{B.14}$$

In this setup, β_τ controls the degree of treatment effect heterogeneity. As β_τ increases in magnitude, the individual treatment effects depend more on the covariate values, while

$\beta_\tau = 0$ recovers the constant effects case. When β_τ and β_π are the same sign, the weights and individual treatment effects are positively correlated; otherwise, they are negatively correlated.

As discussed in Section 3.4, the impact of trimming on design sensitivity can depend on the correlation between the weights and the treatment effects. For positive correlation, trimming units with large weights also removes units with larger treatment effects, reducing the ATT and thus the design sensitivity. The reverse is true when the weights and effect sizes are negatively correlated.

The simulations presented in Table B.7 examine the impact of trimming on design sensitivity for varying levels of treatment effect heterogeneity. In the constant effects case with $\beta_\tau = 0$, trimming improves design sensitivity for both the variance-based and marginal sensitivity models. As β_τ decreases, trimming increases the treatment effect and thus increases the design sensitivity compared to trimming with constant effects. Conversely, the treatment effect and design sensitivity decrease as β_τ increases, eventually causing trimming to hurt design sensitivity compared to not trimming at all. Higher levels of effect heterogeneity are required for trimming to hurt design sensitivity for the variance-based sensitivity model than the marginal sensitivity model.

Assessing Augmentation under Model Misspecification

Theorem 3.5 establishes the conditions under which outcome model augmentation improves design sensitivity compared to a standard weighted estimator for the variance-based sensitivity model. We now evaluate the impact of model specification on design sensitivity through simulation. The results are displayed in Table B.8. In line with the data generating process in the favorable situation (B.13), the outcome Y is modelled as a linear function of X for the correctly specified case. We also consider several model misspecifications, with each misspecified model replacing X with W. We consider a noise model with $W_i \overset{iid}{\sim} N(0,1)$, $W_i = X_i^3$ for misspecification 1, $W_i = \exp\{X_i/2\}$ for misspecification 2, and $W_i = \log(X_i^4)$ for misspecification 3.

The correctly specified model yields the largest improvements in design sensitivity compared to weighting alone for both sensitivity models. The improvement in design sensitivity stems from the reduction in the variance of the outcome from augmentation, which can be seen by comparing the variance of Y to the variance of the residual e, where the residual e plays the role of a pseudo outcome for the augmented weighted estimator. The design sensitivities are unchanged for the noise model; however, it is not advisable to implement this model in practice since it could lead to less precise estimates in a finite-sample. While performing augmentation with the first two misspecified models does not help design sensitivity as much as with the correctly specified model, both models result in higher design sensitivities than the standard weighted estimator, highlighting that even misspecified outcome models could lead to improvements. On the other hand, the third misspecified model yields lower design sensitivity values for the marginal sensitivity model and only minor improvements for the variance-based sensitivity model.

Table B.6: Drivers of design sensitivity

Simulation Parameters								
τ	β_y	β_π	σ_y	$\text{var}(Y \mid Z=0)$	$\text{var}(w \mid Z=0)$	$\text{cor}(w, Y \mid Z=0)$	$\tilde{\Lambda}$	$\tilde{R}^2$
Effect Size								
0.25	1	1	1	1.83	1.28	0.54	1.27	0.04
0.50	1	1	1	1.82	1.30	0.54	1.59	0.13
0.75	1	1	1	1.83	1.35	0.53	2.01	0.24
1.00	1	1	1	1.82	1.29	0.54	2.55	0.37
1.25	1	1	1	1.83	1.26	0.54	3.27	0.49
1.50	1	1	1	1.83	1.30	0.54	4.16	0.57
1.75	1	1	1	1.83	1.29	0.54	5.35	0.64
2.00	1	1	1	1.83	1.35	0.53	7.02	0.70
Variance in Outcomes								
1.00	0.5	1	0.5	0.46	1.31	0.54	6.89	0.70
1.00	1.0	1	1.0	1.83	1.30	0.54	2.53	0.37
1.00	1.5	1	1.5	4.11	1.31	0.53	1.86	0.20
1.00	2.0	1	2.0	7.33	1.29	0.54	1.58	0.13
1.00	2.5	1	2.5	11.42	1.29	0.54	1.45	0.09
Variance in Weights								
1.00	0.80	0.50	1.11	1.84	0.27	0.54	2.55	0.74
1.00	0.88	0.75	1.07	1.83	0.66	0.54	2.56	0.54
1.00	1.00	1.00	1.00	1.83	1.32	0.53	2.53	0.37
1.00	1.17	1.25	0.88	1.82	2.30	0.54	2.54	0.25
1.00	1.40	1.50	0.65	1.83	4.30	0.53	2.50	0.15
1.00	1.64	1.75	0.14	1.83	6.52	0.54	2.60	0.11
Correlation between Weights and Outcomes								
1.00	-1.4	1	0.46	1.83	1.29	-0.75	2.56	0.49
1.00	-1.2	1	0.80	1.83	1.31	-0.64	2.56	0.41
1.00	-1.0	1	1.00	1.82	1.32	-0.53	2.55	0.37
1.00	-0.8	1	1.14	1.83	1.26	-0.43	2.55	0.34
1.00	-0.6	1	1.24	1.83	1.27	-0.32	2.55	0.32
1.00	-0.4	1	1.30	1.83	1.27	-0.21	2.56	0.31
1.00	-0.2	1	1.34	1.83	1.31	-0.11	2.56	0.30
1.00	0.0	1	1.35	1.83	1.29	0.00	2.56	0.30
1.00	0.2	1	1.34	1.83	1.32	0.11	2.56	0.29
1.00	0.4	1	1.30	1.83	1.27	0.22	2.56	0.31
1.00	0.6	1	1.24	1.83	1.30	0.32	2.55	0.32
1.00	0.8	1	1.14	1.83	1.29	0.43	2.56	0.34
1.00	1.0	1	1.00	1.83	1.31	0.53	2.52	0.36
1.00	1.2	1	0.80	1.82	1.28	0.64	2.57	0.42
1.00	1.4	1	0.46	1.83	1.31	0.75	2.53	0.49

Impact of trimming on design sensitivity for varying treatment effect heterogeneity

β_τ	Trimmed ATT	$\tilde{\Lambda}$	$\tilde{\Lambda}_{trim}$	Change	$\tilde{R}^2$	$\tilde{R}^2_{trim}$	Change
-1.5	2.64	6.10	15.34	9.24	0.42	0.73	0.31
-1.0	2.49	6.55	12.86	6.31	0.44	0.71	0.27
-0.5	2.36	6.38	10.80	4.42	0.44	0.69	0.24
0.0	2.23	6.22	9.19	2.97	0.43	0.66	0.23
0.5	2.10	6.55	8.01	1.46	0.44	0.63	0.19
1.0	1.96	6.24	6.82	0.58	0.44	0.60	0.16
1.5	1.83	6.27	5.89	-0.38	0.43	0.57	0.14
2.0	1.70	6.06	5.08	-0.98	0.43	0.53	0.10
3.0	1.43	6.47	3.81	-2.66	0.44	0.44	-0.00
4.0	1.17	6.16	2.97	-3.19	0.44	0.35	-0.09

Table B.7: We vary the amount of treatment effect heterogeneity and assess the impact of trimming on design sensitivity. For the simulation, we set the average treatment effect to be equal to 2.23, $\text{var}(Y \mid Z = 0) = 2.48$, $\text{var}(Y \mid Z = 0, w < m) = 2.4$, $\text{cor}(w, Y \mid Z = 0) = 0.5$, $\text{cor}(w, Y \mid Z = 0, w < m) = 0.6$.

Impact of augmentation on design sensitivity under outcome model misspecification

	Standard IPW		Augmented IPW							
σ_y^2	$\text{cor}(w, Y \mid Z = 0)$	$\text{var}(Y \mid Z = 0)$	$\text{cor}(e, Y \mid Z = 0)$	$\text{var}(e \mid Z = 0)$	$\tilde{\Lambda}$	$\tilde{\Lambda}_{aug}$	Change	$\tilde{R}^2$	$\tilde{R}^2_{aug}$	Change
Outcome Model Type: Correct										
0.50	0.70	1.08	-0.00	0.25	6.60	87.64	81.04	0.76	0.87	0.12
1.00	0.54	1.84	0.00	1.01	4.15	7.18	3.03	0.57	0.63	0.06
1.50	0.42	3.08	0.00	2.25	2.97	3.59	0.62	0.41	0.44	0.03
2.00	0.33	4.84	0.00	4.01	2.37	2.58	0.21	0.28	0.30	0.01
3.00	0.23	9.81	-0.00	8.97	1.83	1.88	0.05	0.16	0.16	0.00
Outcome Model Type: Noise										
0.50	0.70	1.08	0.70	1.08	6.50	6.50	0.00	0.76	0.76	0.00
1.00	0.54	1.83	0.54	1.83	4.18	4.18	0.00	0.58	0.58	0.00
1.50	0.41	3.08	0.41	3.08	2.94	2.94	0.00	0.40	0.40	-0.00
2.00	0.33	4.84	0.33	4.84	2.36	2.36	0.00	0.28	0.28	-0.00
3.00	0.23	9.79	0.23	9.79	1.84	1.84	0.00	0.16	0.16	0.00
Outcome Model Type: Misspecification 1										
0.50	0.69	1.08	0.46	0.61	6.46	22.90	16.44	0.75	0.78	0.03
1.00	0.54	1.83	0.31	1.37	4.17	5.62	1.45	0.57	0.59	0.01
1.50	0.41	3.09	0.23	2.62	2.95	3.28	0.33	0.40	0.41	0.01
2.00	0.33	4.82	0.17	4.35	2.36	2.49	0.13	0.28	0.29	0.01
3.00	0.23	9.82	0.12	9.34	1.83	1.86	0.03	0.16	0.16	0.00
Outcome Model Type: Misspecification 2										
0.50	0.69	1.08	-0.15	0.34	6.43	45.31	38.88	0.75	0.84	0.09
1.00	0.54	1.83	-0.08	1.08	4.12	6.27	2.15	0.57	0.62	0.05
1.50	0.41	3.10	-0.06	2.35	2.98	3.41	0.43	0.40	0.43	0.02
2.00	0.33	4.84	-0.04	4.09	2.36	2.53	0.17	0.29	0.30	0.01
3.00	0.23	9.86	-0.03	9.12	1.84	1.87	0.03	0.16	0.16	0.00
Outcome Model Type: Misspecification 3										
0.50	0.70	1.08	0.74	0.96	6.62	4.67	-1.95	0.76	0.80	0.04
1.00	0.53	1.83	0.55	1.71	4.15	3.50	-0.65	0.56	0.59	0.02
1.50	0.42	3.08	0.42	2.96	2.96	2.74	-0.22	0.40	0.41	0.01
2.00	0.33	4.83	0.34	4.71	2.37	2.28	-0.09	0.28	0.29	0.01
3.00	0.23	9.80	0.23	9.68	1.84	1.82	-0.02	0.16	0.16	0.00

Table B.8: For the correctly specified outcome model, we model the control potential outcome $Y(0)$ as a linear function of X. For other outcome models, we replace X with W, where $W_i \overset{iid}{\sim} N(0,1)$ for the noise model, $W_i = X_i^3$ for misspecification 1, $W_i = \exp\{X_i/2\}$ for misspecification 2, and $W_i = \log\left(X_i^4\right)$ for misspecification 3.