Journal of the Royal Statistical Society Series A: Statistics in Society, 2023, **186**, 707–721
https://doi.org/10.1093/jrsssa/qnad032
Advance access publication 22 March 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Interpretable sensitivity analysis for balancing weights

Dan Soriano[1], Eli Ben-Michael[2], Peter J. Bickel[1], Avi Feller[3] and Samuel D. Pimentel[1]

[1]Department of Statistics, University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720, USA
[2]Department of Statistics & Data Science and the Heinz College of Information Systems and Public Policy, Carnegie Mellon University, 4800 Forbes Ave, Pittsburgh, PA 15213, USA
[3]Goldman School of Public Policy & Department of Statistics, University of California, Berkeley, 2607 Hearst Avenue, Berkeley, CA 94720, USA

Address for correspondence: Dan Soriano, Department of Statistics, University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720, USA. Email: dan_soriano@berkeley.edu

Abstract

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 paper, 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.

Keywords: balancing weights, causal inference, confounding, observational study, sensitivity analysis, weighting

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).

Received: August 24, 2021. **Revised:** February 2, 2023. **Accepted:** February 17, 2023

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 paper, 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 Background, notation, and review

2.1 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), $i \in \{1, \ldots, n\}$, 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, 1990 [1923]; Rubin, 1974). We assume stable treatment and no interference between units (Rubin, 1980), so the observed outcome is $Y_i = (1 - Z_i)Y_i(0) + Z_i Y_i(1)$. An estimand of interest is the *Population Average Treatment Effect* (PATE),

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

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 5 and Online Supplementary Material, Section SM-3 in the supplementary material.

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 & Rubin, 1983b).

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

Assumption 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 1 and 2, we can non-parametrically identify μ_1, solely with the outcomes from units receiving treatment,

$$\mu_1 = \mathbb{E}\left[\frac{ZY}{\pi(X)}\right]. \tag{2}$$

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. (2021) 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 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 3 (Marginal sensitivity model). For $\Lambda \geq 1$, the true propensity score satisfies

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

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 1 holds. If, for example, $\Lambda = 2$, Assumption 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 $h(x, y) = \log \mathrm{OR}(\pi(x), \pi(x, y))$,

$$\mathcal{H}(\Lambda) = \{h : \mathcal{X} \times \mathbb{R} \to \mathbb{R} : \|h\|_\infty \leq \log \Lambda\}, \tag{3}$$

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

$$\pi^{(h)}(x, y) = \left[1 + \left(\frac{1}{\pi(x)} - 1\right) \mathrm{e}^{h(x,y)}\right]^{-1}. \tag{4}$$

[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.

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

$$\mu_1^{(h)} = \mathbb{E}\left[\frac{Z}{\pi^{(h)}(X, Y(1))}\right]^{-1} \mathbb{E}\left[\frac{ZY}{\pi^{(h)}(X, Y(1))}\right]. \tag{5}$$

Under the marginal sensitivity model in Assumption 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 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 weight estimators that achieve approximate (rather than exact) balance on covariates, such as the stable balancing weights of Zubizarreta (2015). In Section 4, we then introduce an amplification that allows us to better interpret and calibrate marginal sensitivity analyses.

2.2 Weighting estimators under strong ignorability

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

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{Z_i \hat{\gamma}(X_i)}{\sum_{i=1}^{n} Z_i \hat{\gamma}(X_i)} Y_i. \tag{6}$$

Under strong ignorability (Assumptions 1 and 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). 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 & Schafer, 2007).

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

Most balancing weight 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* (SBWs) that find weights $\hat{\gamma}(X)$ that solve

$$\begin{aligned} &\min_{\gamma(X)\in\mathbb{R}^{n_1}} \int Z\gamma(X)^2 \, \mathrm{d}P_n \\ &\quad \text{subject to } \left\| \int Z\gamma(X)\phi(X) - \phi(X) \, \mathrm{d}P_n \right\|_\infty \leq \lambda \quad \gamma(X) \geq 0, \end{aligned} \tag{7}$$

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

[2] Other possibilities include soft balance penalties rather than hard constraints (e.g., Ben-Michael et al., 2023; Keele et al., 2023) and non-parametric measures of balance (e.g., Hirshberg et al., 2019).

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 (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

$$\min_{\beta \in \mathbb{R}^d} \underbrace{\int Z[\beta \cdot \phi(X)]_+^2 - \beta \cdot \phi(X)\, dP_n}_{\text{balancing loss}} + \underbrace{\lambda \|\beta\|_1}_{\text{regularization}}, \tag{8}$$

where $[x]_+ = \max\{0, x\}$. The weights are recovered from the dual solution as $\hat{\gamma}(X_i) = [\hat{\beta} \cdot \phi(X_i)]_+$. 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 weight estimators.

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

$$\hat{\mu}_1^{(h)} = \left(\sum_{Z_i=1} \hat{\gamma}^{(h)}(X_i, Y_i(1)) \right)^{-1} \sum_{Z_i=1} \hat{\gamma}^{(h)}(X_i, Y_i(1)) Y_i, \tag{9}$$

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 weight optimization problem (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

$$\left[L^{(h)}, U^{(h)} \right] = \left[Q_{\frac{\alpha}{2}}\left(\hat{\mu}_{1,b}^{*(h)} \right), Q_{1-\frac{\alpha}{2}}\left(\hat{\mu}_{1,b}^{*(h)} \right) \right]. \tag{10}$$

$Q_\alpha(\hat{\mu}_{1,b}^{*(h)})$ 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 (9) using bootstrap sample $b \in \{1, \ldots, B\}$.

Note, * in $\hat{\mu}_{1,b}^{*(b)}$ 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 $[L^{(b)}, U^{(b)}]$ is an asymptotically valid confidence interval for $\mu_1^{(b)}$ with at least $(1-\alpha)$-coverage under high-level assumptions outlined in the Online supplementary material on how well the balancing weights estimate the propensity scores.

Theorem 1 Under Online Supplementary Material, Assumption 1, for every $b \in H(\Lambda)$,

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

and

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

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 Online Supplementary Material, Assumption 1 and the marginal sensitivity model (3).

Since each of the confidence intervals $[L^{(b)}, U^{(b)}]$ 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 3, where

$$L^{\text{union}} = \inf_{b\in\mathcal{H}(\Lambda)} L^{(b)}, \quad U^{\text{union}} = \sup_{b\in\mathcal{H}(\Lambda)} U^{(b)}. \tag{11}$$

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

$$[L, U] = \left[Q_{\frac{\alpha}{2}}\left(\inf_{b\in\mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(b)} \right), Q_{1-\frac{\alpha}{2}}\left(\sup_{b\in\mathcal{H}(\Lambda)} \hat{\mu}_{1,b}^{*(b)} \right) \right]. \tag{12}$$

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 that this can be overly conservative; see Sections 5.2 and 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:

$$\begin{aligned} \min / \max_{r\in\mathbb{R}^{n_1}} \quad & \hat{\mu}_1^{(b)} = \frac{\sum_{i=1}^{n} Z_i\big(1 + r_i\big[\hat{\gamma}(X_i) - 1\big]\big) Y_i}{\sum_{i=1}^{n} Z_i\big(1 + r_i\big[\hat{\gamma}(X_i) - 1\big]\big)} \\ \text{subject to} \quad & r_i \in [\Lambda^{-1}, \Lambda], \ \text{ for all } i \in \{1, \ldots, n\}, \end{aligned} \tag{13}$$

where $r_i = \mathrm{OR}\{\pi(X_i), \pi(X_i, Y_i(1))\}$ 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 $b = 1, \ldots, B$, re-estimate the weights and compute the extrema $\inf_{\mathrm{h} \in \mathcal{H}(\Lambda)} \hat{\mu}^{*(b)}_{1,b}$ and $\sup_{\mathrm{h} \in \mathcal{H}(\Lambda)} \hat{\mu}^{*(b)}_{1,b}$ under the collection of sensitivity models $\mathcal{H}(\Lambda)$ by solving (13).

Step 3 Obtain valid confidence intervals for sensitivity analysis

$$L = Q_{\frac{\alpha}{2}}\left(\inf_{b \in \mathcal{H}(\Lambda)} \hat{\mu}^{*(b)}_{1,b}\right), \quad U = Q_{1-\frac{\alpha}{2}}\left(\sup_{b \in \mathcal{H}(\Lambda)} \hat{\mu}^{*(b)}_{1,b}\right). \tag{14}$$

Replacing $\hat{\gamma}(X_i)$ in Equation (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) and Pimentel and Kelz (2020).

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 & Silber, 2009). Under the marginal sensitivity model in Assumption 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, 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

[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 (13) using the full observed data for a particular value of Λ.

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

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,

$$\mathbb{E}[Y(1) \mid X = x,\, U = u] = f(x) + \beta_u \cdot u. \tag{15}$$

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 (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,

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

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

$$|\beta_u \cdot \delta_u| \leq \max\left\{\left|\inf_{h \in \mathcal{H}(\Lambda^*)} \hat{\mu}_1^{(h)} - \hat{\mu}_1\right|, \left|\sup_{h \in \mathcal{H}(\Lambda^*)} \hat{\mu}_1^{(h)} - \hat{\mu}_1\right|\right\}. \tag{16}$$

Recall that $\hat{\mu}_1$ (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 5, we illustrate our sensitivity analysis procedure and how our amplification can produce more interpretable results.

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 Online Supplementary Material, Section SM-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 values 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 a Monte Carlo error.

We fix the bias equal to the maximum absolute value of the upper and lower bounds on the bias in Equation (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 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.

5.1 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 (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 1 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.

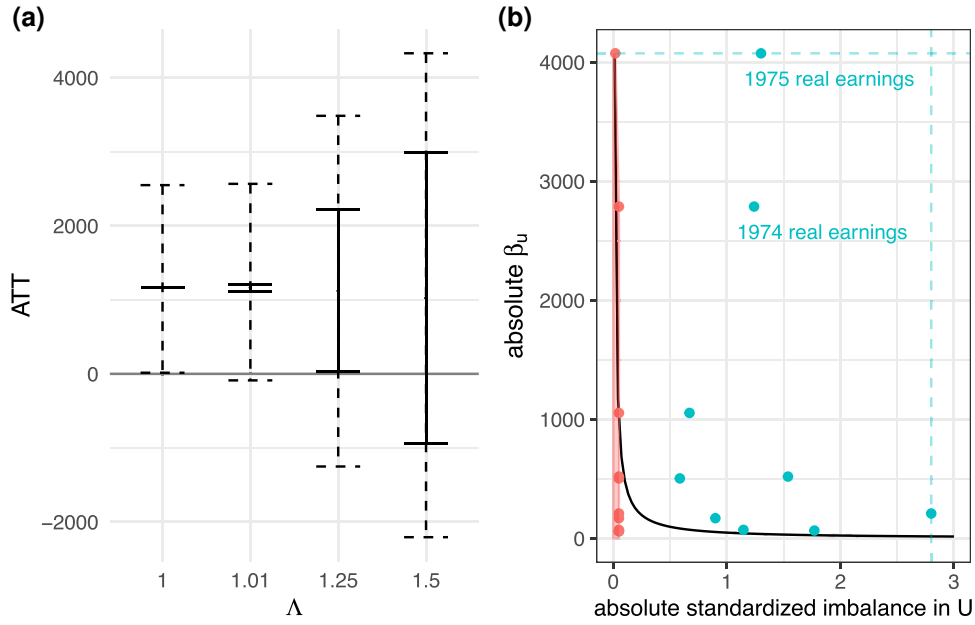

Figure 1. Sensitivity analysis results with the LaLonde data. (a) Point estimate and confidence intervals. Solid intervals are point estimate intervals and dotted intervals are 95% confidence intervals. (b) Contour plot illustrating amplification of the sensitivity analysis with comparison to observed variables. 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 (the blue ● points farther from the y-axis) and after weighting (the red ● points closer to the y-axis). 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.

Figure 1 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 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 these strongly substantiate 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 least the red shaded region appear misleadingly small.

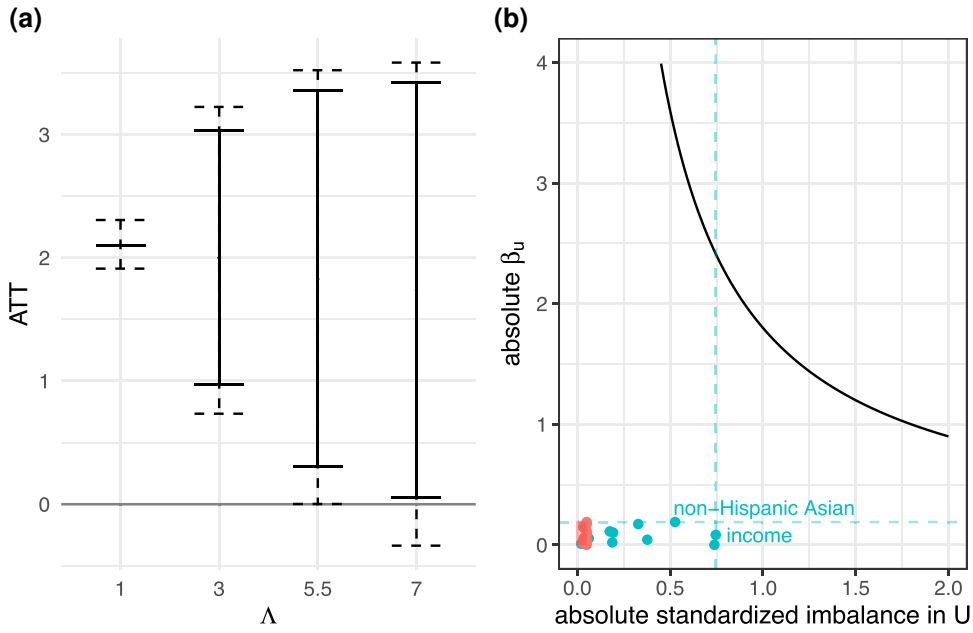

Figure 2. Sensitivity analysis results with the fish diet data. (a) Point estimate and confidence intervals. Solid intervals are point estimate intervals and dotted intervals are 95% confidence intervals. (b) Contour plot illustrating amplification of the sensitivity analysis, with comparison to observed variables. 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 (the blue ● points farther from the y-axis) and after weighting (the red ● points closer to the y-axis). 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.

5.2 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 (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. 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) 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 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 1 and the fish consumption data results in Figure 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 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 (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 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 and precision compared to existing competitors.

6 Discussion

Balancing weight estimation is a popular approach for estimating treatment effects by weighting units to balance covariates. In this paper, 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 5.2, Dorn and Guo (2021) show that the intervals obtained from solving the linear programming problem (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 weight 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 &

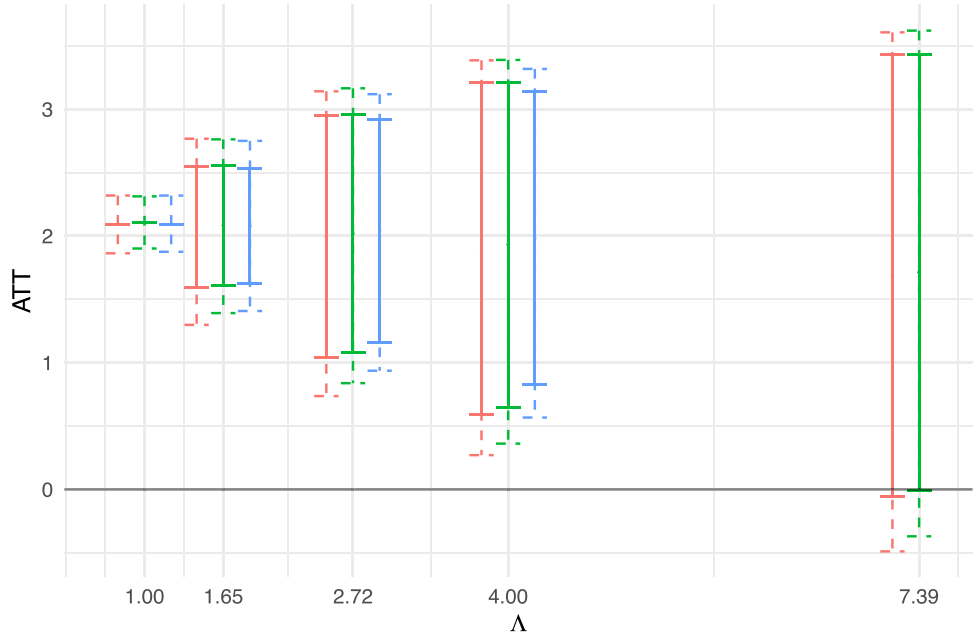

Figure 3. Comparison of confidence interval width after sensitivity analysis for three approaches in the fish consumption example. We compare the intervals constructed using stabilized weights followed by our proposed sensitivity analysis (green intervals in the middle for each value of Λ) against those obtained by fitting IPW weights and conducting sensitivity analysis as described in Zhao et al. (2019) (red intervals on the left), and against those obtained by IPW and the approach of Dorn and Guo (2021) (blue intervals on the right), 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.

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.

Acknowledgments

All data and code used in this paper are available upon request. We would like to thank Kevin Guo and Skip Hirshberg for useful discussion and comments.

Conflict of interest: None declared.

Funding

This research was supported in part by the Hellman Family Fund at UC Berkeley, the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010, the Office of Naval Research (ONR) through grant N00014-17-1-2176, and the Two Sigma PhD fellowship. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, nor the Office of Naval Research.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series A* online.

References

Abadie A., & Gardeazabal J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, *93*(1), 113–132. https://doi.org/10.1257/000282803321455188

Athey S., Imbens G. W., & 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. https://doi.org/10.1111/rssb.12268

Ben-Michael E., Feller A., Hirshberg D. A., & Zubizarreta J. R. (2021). 'The balancing act in causal inference', arXiv, arXiv:2110.14831, preprint: not peer reviewed.

Ben-Michael E., Feller A., & Rothstein J. (2018). 'The augmented synthetic control method', arXiv, arXiv:1811.04170, preprint: not peer reviewed.

Ben-Michael E., Feller A., & Rothstein J. (2023). Varying impacts of letters of recommendation on college admissions. *Annals of Applied Statistics*.

Björnberg K. A., Vahter M., Petersson-Grawe K., Glynn A., Cnattingius S., Darnerud P., Atuma S., Aune M., Becker W., & 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. https://doi.org/10.1289/ehp.111-1241457

Brown L. D., Casella G., & Gene Hwang J. (1995). Optimal confidence sets, bioequivalence, and the limacon of Pascal. *Journal of the American Statistical Association*, *90*(431), 880–889. https://doi.org/10.2307/2291322

Chattopadhyay A., Hase C. H., & Zubizarreta J. R. (2020). Balancing versus modeling approaches to weighting in practice. *Statistics in Medicine*, *39*(24), 3227–3254. https://doi.org/10.1002/sim.8659

Cinelli C., & Hazlett C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Cornfield J., Haenszel W., Hammond E. C., Lilienfeld A. M., Shimkin M. B., & 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. https://doi.org/10.1093/jnci/22.1.173

Dehejia R. H., & Wahba S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858

Deville J. C., & Särndal C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382. https://doi.org/10.1080/01621459.1992.10475217

Deville J. C., Särndal C. E., & Sautory O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*(423), 1013–1020. https://doi.org/10.1080/01621459.1993.10476369

Dorn J., & Guo K. (2021). 'Sharp sensitivity analysis for inverse propensity weighting via quantile balancing', arXiv, arXiv:2102.04543, preprint: not peer reviewed.

Firpo S., & Possebom V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, *6*(2), 20160026. https://doi.org/10.1515/jci-2016-0026

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. https://doi.org/10.1080/01621459.2019.1632072

Franks A., D'Amour A., & Feller A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, *115*(532), 1–38. https://doi.org/10.1080/01621459.2019.1604369

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. https://doi.org/10.1093/pan/mpr025

Hirshberg D. A., Maleki A., & Zubizarreta J. (2019). 'Minimax linear estimation of the retargeted mean', arXiv, arXiv:1901.10296, preprint: not peer reviewed.

Hong G., Yang F., & Qin X. (2021). 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)*, *184*(1), 227–254. https://doi.org/10.1111/rssa.12621

Huang M. (2022). 'Sensitivity analysis in the generalization of experimental results', arXiv, arXiv:2202.03408, preprint: not peer reviewed.

Huang M., & Pimentel S. D. (2022). 'Variance-based sensitivity analysis for weighting estimators result in more informative bounds', arXiv, arXiv:2208.01691, preprint: not peer reviewed.

Kang J. D., & Schafer J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539. https://doi.org/10.1214/07-STS227

Keele L., Ben-Michael E., Feller A., Kelz R., & Miratrix L. (2023). Hospital quality risk standardization via approximate balancing weights. *Annals of Applied Statistics*.

LaLonde R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, *76*(4), 604–620.

Mahaffey K. R., Clickner R. P., & 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. https://doi.org/10.1289/ehp.6587

Neyman J. (1990 [1923]). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472. https://doi.org/10.1214/ss/1177012031

Pimentel S. D., & 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. https://doi.org/10.1080/01621459.2020.1720693

Pimentel S. D., Kelz R. R., Silber J. H., & 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. https://doi.org/10.1080/01621459.2014.997879

Rosenbaum P. R. (2002). *Observational studies*. Springer.

Rosenbaum P. R., & 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. https://doi.org/10.1111/j.2517-6161.1983.tb01242.x

Rosenbaum P. R., & Rubin D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum P. R., & Silber J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, *104*(488), 1398–1405. https://doi.org/10.1198/jasa.2009.tm08470

Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593. https://doi.org/10.2307/2287653

Tan Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, *101*(476), 1619–1637. https://doi.org/10.1198/016214506000000023

Tan Z. (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, *107*(1), 137–158. https://doi.org/10.1093/biomet/asz059

Tudball M., Hughes R., Tilling K., Bowden J., & Zhao Q. (2019). 'Sample-constrained partial identification with application to selection bias', arXiv, arXiv:1906.10159, preprint: not peer reviewed.

VanderWeele T. J., & Ding P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, *167*(4), 268–274. https://doi.org/10.7326/M16-2607

Veitch V., & 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., & Zubizarreta J. R. (2019). 'Minimal approximately balancing weights: Asymptotic properties and practical considerations', arXiv, arXiv:1705.00998, preprint: not peer reviewed.

Zhao Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*, *47*(2), 965–993. https://doi.org/10.1214/18-AOS1698

Zhao Q., Small D. S., & 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)*, *81*(4), 735–761. https://doi.org/10.1111/rssb.12327

Zhao Q., Small D. S., & Rosenbaum P. R. (2018). Cross-screening in observational studies that test many hypotheses. *Journal of the American Statistical Association*, *113*(523), 1070–1084. https://doi.org/10.1080/01621459.2017.1407770

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. https://doi.org/10.1080/01621459.2015.1023805