Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding

Anonymous Author(s)
Affiliation
Address
email

Abstract

It is a truth universally acknowledged that an observed association without known mechanism must be in want of a causal estimate. Causal estimates from observational data will be biased in the presence of 'unobserved confounding'. However, we might hope that the influence of unobserved confounders is weak relative to a 'large' estimated effect. The purpose of this paper is to develop *Austen plots*, a sensitivity analysis tool to aid such judgments by making it easier to reason about potential bias induced by unobserved confounding. We formalize confounding strength in terms of how strongly the unobserved confounding influences treatment assignment and outcome. For a target level of bias, an Austen plot shows the minimum values of treatment and outcome influence required to induce that level of bias. Austen plots generalize the classic sensitivity analysis approach of Imbens [Imb03]. Critically, Austen plots allow *any* approach for modeling the observed data. We illustrate the tool by assessing biases for several real causal inference problems, using a variety of machine learning approaches for the initial data analysis. Code, demo data, and a tutorial are available at [removed].

The high costs of randomized controlled trials coupled with the relative availability of (large scale) observational data motivate attempts to infer causal relationships from observational data. For example, we may wish to use a database of electronic health records to estimate the effect of a treatment. Causal inference from observational data must account for possible *confounders* that influence both treatment assignment and the outcome; e.g., wealth may be a common cause influencing whether a patient takes an expensive drug and whether they recover. Often, causal inference is based on the assumption of 'no unobserved confounding'; i.e., the assumption that the observed covariates include all common causes of the treatment assignment and outcome. This assumption is fundamentally untestable from observed data, but its violation can induce bias in the estimation of the treatment effect—the unobserved confounding may completely or in part explain the observed association. Our aim in this paper is to develop a sensitivity analysis tool to aid in reasoning about potential bias induced by unobserved confounding.

Intuitively, if we estimate a large positive effect then we might expect the real effect is also positive, even in the presence of mild unobserved confounding. For example, consider the association between smoking and lung cancer. One could argue that this association arises from a genetic mutation that predisposes carriers to both an increased desire to smoke and to a greater risk of lung cancer. However, the association between smoking and lung cancer is large—is it plausible that some unknown genetic association could have a strong enough influence to explain the association? Answering such questions requires a domain expert to make a judgment about whether plausible confounding is "mild" relative to the "large" effect. In particular, the domain expert must translate judgments about the strength of the unobserved confounding into judgments about the bias induced in the estimate of the effect. Accordingly, we must formalize what is meant by strength of unobserved confounding, and to show how to translate judgments about confounding strength into judgments about bias.

A prototypical example, due to Imbens [Imb03] (building on [RR83]), illustrates the broad approach. The observed data consists of a treatment T, an outcome Y, and covariates X that may causally affect the treatment and outcome. Imbens [Imb03] then posits an additional unobserved binary confounder U for each patient, and supposes that the observed data and unobserved confounder were generated according to:

$$\begin{aligned} U_i &\overset{\text{iid}}{\sim} \text{Bern}(1/2) \\ T_i \mid X_i, U_i &\overset{ind}{\sim} \text{Bern}(\text{sig}(\gamma X_i + \alpha U_i)) \\ Y_i \mid X_i, T_i, U_i &\overset{ind}{\sim} \text{Norm}(\tau T_i + \beta X_i + \delta U_i, \sigma^2). \end{aligned}$$

where sig is the sigmoid function. If we had observed U_i, we could estimate $(\hat{\tau}, \hat{\gamma}, \hat{\beta}, \hat{\alpha}, \hat{\delta}, \hat{\sigma}^2)$ from the data and report $\hat{\tau}$ as the estimate of the average treatment effect. Since U_i is not observed, it is not possible to identify the parameters from the data. Instead, we make (subjective) judgments about plausible values of α—how strongly U_i affects the treatment assignment—and δ—how strongly U_i affects the outcome. Contingent on plausible $\alpha = \alpha^*$ and $\delta = \delta^*$, the other parameters can be estimated. This yields an estimate of the treatment effect $\hat{\tau}(\alpha^*, \delta^*)$ under the presumed values of the sensitivity parameters.

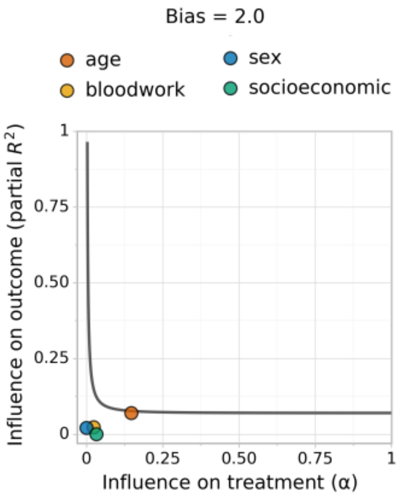

Figure 1: Austen plot showing how strong an unobserved confounder would need to be to induce a bias of 2 in an observational study of the effect of combination blood pressure medications on diastolic blood pressure [Dor+16]. We chose this bias to equal the nominal average treatment effect estimated from the data. We model the outcome with Bayesian Additive Regression Trees and the treatment assignment with logistic regression—Austen plots accommodate any choice of models. The curve shows all values treatment and outcome influence that would induce a bias of 2. The colored dots show the influence strength of (groups of) observed covariates, given all other covariates. For example, an unobserved confounder with as much influence as the patient's age might induce a bias of about 2.

The approach just outlined has a major drawback: it relies on a parametric model for the full data generating process. The assumed model is equivalent to assuming that, had U been observed, it would have been appropriate to use logistic regression to model treatment assignment, and linear regression to model the outcome. This assumption also implies a simple, parametric model for the relationships governing the observed data. This restriction is out of step with modern practice, where we use flexible machine-learning methods to model these relationships. For example, the assumption forbids the use of neural networks or random forests, though such methods are often state-of-the-art for causal effect estimation.

Austen plots The purpose of this paper is to introduce *Austen plots*, an adaptation of Imbens' approach that fully decouples sensitivity analysis and modeling of the observed data. An example Austen plot is shown in Figure 1. The high-level idea is to posit a generative model that uses a simple, interpretable parametric form for the influence of the unobserved confounder, but that *puts no constraints on the model for the observed data*. We then use the parametric part of the model to formalize "confounding strength" and to compute the induced bias as a function of the confounding.

We further adapt two innovations pioneered by Imbens [Imb03]. First, we find a parameterization of the model so that the sensitivity parameters, measuring strength of confounding, are on a standardized, unitless scale. This allows us to compare the strength of hypothetical unobserved confounding to the strength of observed covariates, measured from data. Second, we plot the curve of all values of the sensitivity parameter that would yield given level of bias. This moves the analyst judgment from "what are plausible values of the sensitivity parameters?" to "are sensitivity parameters this extreme plausible?"

Figure 1, an Austen plot for an observational study of the effect of combination medications on diastolic blood pressure, illustrates the idea. A bias of 2 would suffice to undermine the qualitative conclusion that the blood-pressure treatment is effective. Examining the plot, an unobserved confounder as strong as age could induce this amount of confounding, but no other (group of) observed confounders has so much influence. Accordingly, if a domain expert thinks an unobserved confounder as strong as age is unlikely then they may conclude that the treatment is likely effective. Or, if such a confounder is plausible, they may conclude that the study fails to establish efficacy.

The purpose of this paper is adapting Imbens' sensitivity analysis approach to allow for arbitrary models for observed data. The contributions are: 1. Positing a generative model that is both easily interpretable and where the required bias calculations are tractable. 2. Deriving a reparameterization that standardizes the scale of influence strength, and showing how to estimate the influence strength of observed covariates for reference. And, 3. illustrative examples showing that Austen plots preserve the key elements of Imbens' approach and are informative about sensitivity to unobserved confounding in real-world data.

The key advantages of Austen plots as a sensitivity analysis method are[1] 1. Plausibility judgments are made on directly interpretable quantities, the total confounding influence on Y and T. Additionally, the Austen plot model does not rely on the detailed nature of the unobserved confounding—there may be one or many unobserved confounders, with any sort of distribution—all that matters is the total confounding influence. 2. The unobserved strength of confounding can be directly compared to the strength of observed covariates. 3. The method is entirely post-hoc. That is, the analyst does not need to consider any aspect of the sensitivity analysis when modeling the observed data. In particular, producing Austen plots requires *only predictions* from the data models. We provide software and a tutorial for producing the plots.[2]

Notation For concreteness, we focus on the estimation of the average effect of a binary treatment. The data are generated independently and identically $(Y_i, T_i, X_i, U_i) \overset{\text{iid}}{\sim} P$, where U_i is not observed and P is some unknown probability distribution. The average treatment affect (ATE) is

$$\texttt{ATE} = \mathbb{E}[Y \mid \mathrm{do}(T=1)] - \mathbb{E}[Y \mid \mathrm{do}(T=0)].$$

The use of Pearl's do notation indicates that the effect of interest is causal. The results that follow can also be simply adapted to the average treatment effect on the treated, see appendix A.

The traditional approach to causal estimation assumes that the observed covariates X contain all common causes of Y and T. If this 'no unobserved confounding' assumption holds, then the ATE is equal to parameter, τ, of the observed data distribution, where

$$\tau = \mathbb{E}[\mathbb{E}[Y \mid X, T=1] - \mathbb{E}[Y \mid X, T=0]]. \tag{0.1}$$

The parameter τ can be estimated from a finite data sample. The general approach proceeds in two steps. First, we produce estimates $\hat{g}$ and $\hat{Q}$ for the propensity score g and the conditional expected outcome Q, where

Definition 1. The *propensity score* g is $g(x) = \mathrm{P}(T=1 \mid X=x)$ and the *conditional expected outcome* Q is $Q(t,x) = \mathbb{E}[Y \mid T=t, X=x]$.

In modern practice, Q and g are often estimated by fitting flexible machine learning models. The second step is to plug the estimated $\hat{Q}$ and $\hat{g}$ in to some downstream estimator $\hat{\tau}$. For example, following 0.1, the estimator

$$\hat{\tau}^Q = \frac{1}{n}\sum_i \hat{Q}(1, x_i) - \hat{Q}(0, x_i),$$

is a natural choice. Other estimators incorporate $\hat{g}$.

We are interested in the case of possible unobserved confounding. That is, where U causally affects Y and T. If there is unobserved confounding then the parameter τ is not equal to the ATE, so $\hat{\tau}$ is a biased estimate. Inference about the ATE then divides into two tasks. First, the statistical task: estimating τ as accurately as possible from the observed data. And, second, the causal (domain-specific) problem of assessing $\texttt{bias} = \texttt{ATE} - \tau$. We emphasize that our focus here is bias due to causal misidentification, not the statistical bias of the estimator. Our aim is to reason about the bias induced by unobserved confounding—the second task—in a way that imposes no constraints on the modeling choices for $\hat{Q}$, $\hat{g}$ and $\hat{\tau}$ used in the initial analysis.

[1] See section 4 for a more detailed comparison with related work.

[2] Supplementary material.

1 Sensitivity Model

Our sensitivity analysis should impose no constraints on how the *observed* data is modeled. However, sensitivity analysis demands some assumption on the relationship between the observed data and the *unobserved* confounder. It is convenient to formalize such assumptions by specifying a probabilistic model for how the data is generated. The strength of confounding is then formalized in terms of the parameters of the model (the sensitivity parameters). Then, the bias induced by the confounding can be derived from the assumed model. Our task is to posit a generative model that both yields a useful and easily interpretable sensitivity analysis, and that avoids imposing any assumptions about the observed data.

To begin, consider the functional form of the sensitivity model used by Imbens [Imb03].

$$\operatorname{logit} \mathrm{P}(T = 1 \mid x, u) = h(x) + \alpha u \tag{1.1}$$

$$\mathbb{E}[Y \mid t, x, u] = l(t, x) + \delta u, \tag{1.2}$$

for some functions h and l. That is, the propensity score is logit-linear in the unobserved confounder, and the conditional expected outcome is linear.

By rearranging (1.1) to solve for u and plugging in to (1.2), we see that it's equivalent to assume $\mathbb{E}[Y \mid t, x, u] = \tilde{l}(t, x) + \tilde{\delta} \operatorname{logit} \mathrm{P}(T = 1 \mid x, u)$. That is, the unobserved confounder u only influences the outcome through the propensity score. Accordingly, by positing a distribution on $\mathrm{P}(T = 1 \mid x, u)$ directly, we can circumvent the need to explicitly articulate U (and h).

Definition 2. Let $\tilde{g}(x, u) = \mathrm{P}(T = 1 \mid x, u)$ denote the propensity score given observed covariates x and the unobserved confounder u.

The insight is that we can posit a sensitivity model by defining a distribution on $\tilde{g}$ directly. The logit-linear model does not directly lead to a tractable sensitivity analysis. Instead, we choose:

$$\tilde{g}(X, U) \mid X \sim \operatorname{Beta}(g(X)({}^1\!/\!_\alpha - 1), (1 - g(X))({}^1\!/\!_\alpha - 1)).$$

The sensitivity parameter α plays the same role as in Imbens' model: it controls the influence of the unobserved confounder U on treatment assignment. When α is close to 0 then $\tilde{g}(X, U) \mid X$ is tightly concentrated around $g(X)$, and the unobserved confounder has little influence. That is, U minimally affects our belief about who is likely to receive treatment. Conversely, when α is close to 1 then $\tilde{g}$ concentrates near 0 and 1; i.e., knowing U would let us accurately predict treatment assignment. Indeed, it can be shown that α is the change in our belief about how likely a unit was to have gotten the treatment, given that they were actually observed to be treated (or not):

$$\alpha = \mathbb{E}[\tilde{g}(X, U) \mid T = 1] - \mathbb{E}[\tilde{g}(X, U) \mid T = 0]. \tag{1.3}$$

With the $\tilde{g}$ model in hand, we define our sensitivity model:

Assumption 1 (Sensitivity Model)**.**

$$\begin{aligned} \tilde{g}(X, U) \mid X &\sim \operatorname{Beta}(g(X)({}^1\!/\!_\alpha - 1), (1 - g(X))({}^1\!/\!_\alpha - 1)) \\ T \mid X, U &\sim \operatorname{Bern}(\tilde{g}(X, U)) \\ \mathbb{E}[Y \mid T, X, U] &= Q(T, X) + \delta\big(\operatorname{logit} \tilde{g}(X, U) - \mathbb{E}[\operatorname{logit} \tilde{g}(X, U) \mid X, T]\big). \end{aligned}$$

This model has been constructed to satisfy the requirement that the propensity score and conditional expected outcome are the g and Q actually present in the observed data:

$$\begin{aligned} \mathrm{P}(T = 1 \mid X) &= \mathbb{E}[\mathbb{E}[T \mid X, U] \mid X] = \mathbb{E}[\tilde{g}(X, U) \mid X] = g(X) \\ \mathbb{E}[Y \mid T, X] &= \mathbb{E}[\mathbb{E}[Y \mid T, X, U] \mid T, X] = Q(T, X). \end{aligned}$$

The sensitivity parameters are α, controlling the dependence between the unobserved confounder the treatment assignment, and δ, controlling the relationship with the outcome. In effect, by making an assumption about the propensity score directly, we have sidestepped the need to explicitly articulate the parts of the observed/unobserved relationship that are not actually relevant for the treatment effect estimation.

Bias We now turn to calculating the bias induced by unobserved confounding. By assumption, X and U together suffice to render the average treatment effect identifiable as:

$$\mathtt{ATE} = \mathbb{E}[\mathbb{E}[Y \mid T=1, X, U] - \mathbb{E}[Y \mid T=0, X, U]].$$

Plugging in our sensitivity model yields,

$$\mathtt{ATE} = \mathbb{E}[Q(1,X) - Q(0,X)] + \delta(\mathbb{E}[\operatorname{logit} \tilde{g}(X,U) \mid X, T=1] - \mathbb{E}[\operatorname{logit} \tilde{g}(X,U) \mid X, T=0]).$$

The first term is the observed-data estimate τ, so

$$\mathtt{bias} = \delta(\mathbb{E}[\operatorname{logit} \tilde{g}(X,U) \mid X, T=1] - \mathbb{E}[\operatorname{logit} \tilde{g}(X,U) \mid X, T=0]).$$

Then, by invoking Beta-Bernoulli conjugacy and standard Beta identities, we arrive at,

Theorem 3. *Under our sensitivity model, Assumption 1, an unobserved confounder with influence α and δ induces bias in the estimated treatment effect equal to*

$$\begin{aligned}\mathtt{bias} = \delta\mathbb{E}\big[&\psi\big(g(X)(1/\alpha - 1) + 1\big) - \psi\big((1-g(X))(1/\alpha - 1)\big) \\ &- \psi\big(g(X)(1/\alpha - 1)\big) + \psi\big((1-g(X))(1/\alpha - 1) + 1\big)\big],\end{aligned}$$

where ψ is the digamma function

Reparameterization The model in the previous section provides a formalization of confounding strength and tells us how much bias is induced by a given strength of confounding. This lets us translate judgments about confounding strength to judgments about bias. However, δ may be difficult to interpret. Following Imbens [Imb03], we will reexpress the outcome-confounder strength in terms of the partial coefficient of determination:

$$R^2_{Y,\mathrm{par}}(\alpha, \delta) = \frac{\mathbb{E}(Y - Q(T,X))^2 - \mathbb{E}(Y - \mathbb{E}[Y \mid T, X, U])^2}{\mathbb{E}(Y - Q(T,X))^2}.$$

This parameterization has two advantages over δ. First, $R^2_{Y,\mathrm{par}}$ has a familiar interpretation—the proportion of previously unexplained variation in Y that is explained by the unobserved covariate U. Second, $R^2_{Y,\mathrm{par}}$ has a fixed, unitless scale—enabling easy comparisons with reference values.

The key to computing the reparameterization is the following result (proof in appendix):

Theorem 4. *Under our sensitivity model, Assumption 1, the outcome influence is*

$$R^2_{Y,\mathrm{par}}(\alpha, \delta) = \delta^2 \sum_{t=0}^{1} \frac{\mathbb{E}\big[\psi_1\big(g(X)^t(1-g(X))^{1-t}(1/\alpha - 1) + 1[T=t]\big)\big]}{\mathbb{E}[(Y - Q(T,X))^2]},$$

where ψ_1 is the trigamma function.

We do not reparameterize the strength of confounding on treatment assignment because, by design, α is already interpretable and on a fixed, unitless scale.

Estimating bias In combination, Theorems 3 and 4 yield an expression for the bias in terms of α and $R^2_{Y,\mathrm{par}}$. In practice, we can estimate the bias induced by confounding by fitting models for $\hat{Q}$ and $\hat{g}$ and replacing the expectations by means over the data. To avoid problems associated with overfitting, we recommend a data splitting approach. Namely, split the data into k folds and, for each fold, estimate $Q(t_i, x_i)$ and $g(x_i)$ by fitting the $\hat{Q}$ and $\hat{g}$ models on the other $k-1$ folds.

2 Calibration using observed data

The analyst must make judgments about the influence a hypothetical unobserved confounder might have on treatment assignment and outcome. To calibrate such judgments, we'd like to have a reference point for how much the observed covariates influence the treatment assignment and outcome. In the sensitivity model, the degree of influence is measured by $R^2_{Y,\mathrm{par}}$ and α. We want to measure the degree of influence of an observed covariate Z given the other observed covariates $X \backslash Z$.

For the outcome, this can be measured as:

$$R^2_{Y,X\backslash Z} := \frac{\mathbb{E}(Y - \mathbb{E}[Y \mid T, X\backslash Z])^2 - \mathbb{E}(Y - Q(T,X))^2}{\mathbb{E}(Y - \mathbb{E}[Y \mid T, X\backslash Z])^2}.$$

In practice, estimate the quantity by fitting a new regression model $\hat{Q}_Z$ that predicts Y from T and $X\backslash Z$. Then we compute

$$\hat{R}^2_{Y,X\backslash Z} = \frac{\frac{1}{n}\sum_i (y_i - \hat{Q}_Z(t_i, x_i\backslash z_i))^2 - \frac{1}{n}\sum_i (y_i - \hat{Q}(t_i, x_i))^2}{\frac{1}{n}\sum_i (y_i - \hat{Q}_Z(t_i, x_i\backslash z_i))^2}.$$

It is less clear how to produce the analogous estimate for the influence on treatment assignment. To facilitate the estimation, we reexpress α in a more convenient form (proof in appendix):

Theorem 5. *Under our sensitivity model, Assumption 1,*

$$\alpha = 1 - \frac{\mathbb{E}[\tilde{g}(X,U)(1-\tilde{g}(X,U))]}{\mathbb{E}[g(X)(1-g(X))]}.$$

Then, we can measure influence of observed covariate Z on treatment assignment given $X\backslash Z$ in an analogous fashion to the outcome. We define $g_{X\backslash Z}(X\backslash Z) = \mathrm{P}(T = 1 \mid X\backslash Z)$, then fit a model for $g_{X\backslash Z}$ by predicting T from $X\backslash Z$, and estimate

$$\hat{\alpha}_{X\backslash Z} = 1 - \frac{\frac{1}{n}\sum_i \hat{g}(x_i)(1-\hat{g}(x_i))}{\frac{1}{n}\sum_i \hat{g}_{X\backslash Z}(x_i\backslash z_i)(1-\hat{g}_{X\backslash Z}(x_i\backslash z_i))}.$$

Grouping covariates The estimated values $\hat{\alpha}_{X\backslash Z}$ and $\hat{R}^2_{Y,X\backslash Z}$ measure the influence of Z conditioned on all the other confounders. In some cases, this can be misleading. For example, if some piece of information is important but there are multiple covariates providing redundant measurements, then the estimated influence of each covariate will be small. To avoid this, we suggest grouping together related or strongly dependent covariates and computing the influence of the entire group in aggregate. For example, grouping income, location, and race as 'socioeconomic variables'.

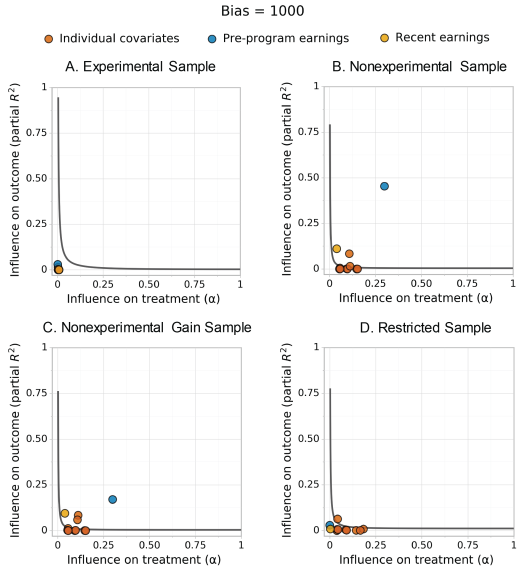

Figure 2: Austen plots preserve the qualitative conclusions of Imbens' analysis without imposing any restriction on the modeling of the observed data. In each plot, the black solid line indicates the partial R^2 and α values that would induce a bias of at least \$1000. Each plot also includes estimates for the strength of confounding for each of the nine covariates (red circles) as well as recent lag in earnings (RE75 and pos75, yellow circles), and the all preprogram earnings (RE74, pos74, RE75, pos75, green circles).

3 Examples

We now examine several examples of Austen plots for sensitivity analysis, showing: (1) We preserve the qualitative usefulness of Imbens' approach, without any modeling restrictions. (2) Austen plots are informative about bias due to unobserved confounding in real observational studies. (3) The bias estimates tend to be conservative.[3]

Imbens' analysis To demonstrate the use of Austen plots, we replicate Imbens [Imb03] example and produce

[3] Code and data in supplementary material.

sensitivity plots for variations on the LaLonde job training data [LaL86]. We use exactly the same data splitting and adjustment sets as Imbens [Imb03]. We find that the conclusions about the effects of unobserved confounding are substantively the same as Imbens [Imb03]. That is, we arrive at sensible sensitivity conclusions while liberating ourselves from the need for parametric assumptions on the observed data. We report bias for the average treatment effect on the treated.

The original purpose of the LaLonde job training data was to analyze the effect of a job training program on the annual earnings of a participant. The data consists of both an experimental (randomly assigned) part, and an observational sample from the Panel Study of Income Dynamics (PSID). We test on (1) the experimental sample, (2) the experimental treated with observational controls, (3) the same as 2, except with outcome defined as change in earnings since 1974, and (4) the same as 2, except individuals with high earnings pretreatment (>$5000) are dropped. We adjust for: married, age, education, race, and earnings in 1974 and 1975. There are large differences in these background covariates between the experimental sample and the PSID controls—this is a main challenge for the LaLonde setup.

Deviating from Imbens, we fit random forests for $\hat{Q}$ and $\hat{g}$. This demonstrates the sensitivity analysis in the case where the observed data model does not have a simple parametric specification.

Austen plots for these analyses are displayed in Figure 2. Following Imbens, we choose a bias of $1000 (for context, the effect estimate from the RCT is about $1800). The experimental sample (panel A) is robust to unobserved confounding: inducing a bias of $1000 would require an unobserved confounder with a much stronger effect than any of the measured covariates or earning variables. By contrast, the non-experimental samples (panels B and C) are much more sensitive to unobserved confounding. Several of the covariates, if unobserved, would suffice to bias the estimate by $1000. Note that the sensitivity curves are the same for both B and C, since the outcome is just a linear transformation. Finally, the restricted sample (panel D) is both significantly more robust to bias than the full non-experimental samples, and the influence of the observed covariates is much reduced. Imposing the restriction mitigates the treatment-control population mismatch.

Practical relevance Figure 3 shows Austen plots for two effects estimated from observational data. The first study is based on data from the Infant Health and Development Program (IHDP) [BG+92], an experiment targeted at low-birth-weight, premature infants that provided child care and home visits. We look at a study measuring the effect of the level of participation in IHDP child development centers in the two years preceding an IQ test on the outcome of the IQ test [Hil11, §6.1]. Level of participation is not randomly assigned, so Hill [Hil11] estimates the effect by using Bayesian Additive Regression Trees (BART) [Chi+10] to control for a range of covariates.

The second plot corresponds to the estimate of the effect of combination blood pressure medications on diastolic blood pressure described in [Dor+16]. The data is derived from an American survey that includes a variety of socioeconomic and health factors. We again use BART.

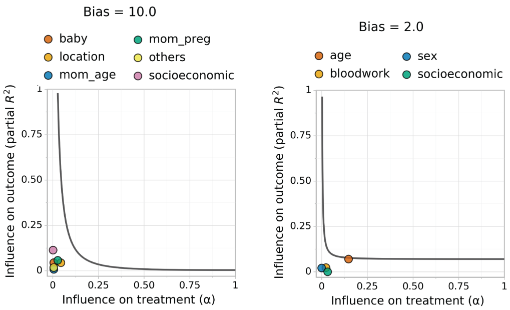

Figure 3: Austen plots are informative when applied to real data analysis. The left-hand plot is for the estimated effect of IHDP participation level on child IQ. The conclusions of this study seem robust to unobserved confounding—even the observed covariate groups do not have sufficient influence to undo the qualitative conclusion of the model. The right-hand plot is for the estimated effect of combination treatment on diastolic blood pressure. In this case, whether the study conclusions are reliable depends on whether an unobserved confounder as influential as age is credible—we should consult with an expert. In both cases, we model the outcome with Bayesian Additive Regression Trees, and the propensity score with logistic regression.

The Austen plots are informative for these examples. In the first case, the Austen plot increases our confidence in the qualitative result. In the second case, it suggests we should be cautious about the conclusions unless unobserved confounders as strong as age are deemed unlikely.

Table 1: The sensitivity model tends to be conservative in its bias estimates. Bias estimates for leaving out a confounding covariate are computed according to the sensitivity model (using the left-out covariate data) and by comparing non-parametric effect estimates from the full data (τ_x), and the left-out covariate data ($\tau_{x\setminus z}$). In all cases, the sensitivity model estimate is larger.

Study:	LaLonde Restricted	Blood Pressure	IHDP
Omitted covariate:	Education	Age	Socioeconomic
τ_x	2508.63	-2.33	12.72
$\tau_{x\setminus z}$	1982.54	-2.86	13.35
Nonparametric $\texttt{bias}$	526.09	0.53	-0.63
Sensitivity Model $\lvert\texttt{bias}\rvert$	986.90	1.91	0.75

Sensitivity model conservatism Any sensitivity analysis must be predicated on some assumption about the influence of the unobserved confounder. The bias curves and influence estimates in Austen plots are contingent on the assumed sensitivity model, Assumption 1. We motivated our particular choice by simplicity and tractibility. We also expect that our associated sensitivity model will often yield conservative values for bias; i.e., the bias anticipated by the sensitivity model is higher than the true bias induced by the real, physical, mechanism of confounding. The reason is that bias is monotonically increasing in both treatment and outcome influence. In reality, hidden confounders can have more complicated relationships that ‘cancel out’. For example, the effect of age in the blood pressure example might be: blood pressure increases with age, but young patients don’t take their medication (preferring diet and exercise), middle age patients take it at a base rate, and old patients don’t take the medication (fatalism). These effects cancel out somewhat, reducing the bias induced by failing to adjust for age. Assumption 1 does not allow for such cancellations.

To test conservativism, we create deliberately confounded datasets by removing an observed confounder from our baseline data. We compute the bias anticipated by our model, $\texttt{bias}(R^2_{Y,X\setminus Z}, \alpha_{X\setminus Z})$, using the measured influence strength of the covariate. We compute a non-parametric estimate of the bias by estimating the effect with the full data, estimating the effect with the deliberately confounded data, and taking the difference. The results are shown in table 1, and confirm the conservatism-in-practice. This increases our confidence that when an Austen plot suggests robustness to unobserved confounding we do indeed have such robustness.

4 Related Work

There are a wealth of approaches to sensitivity analysis. The most closely related approaches to ours are sensitivity analysis based on parametric models in the style of Imbens [Imb03]. These typically assume some relatively simple parametric latent variable model, where the latent variable is the unobserved confounder. Dorie et al. [Dor+16] extends an Imbens-like approach to accomodate BART as the outcome model. Cinelli and Hazlett [CH20] allow for arbitrary kinds of confounders and propensity score models, but require that the outcome is modeled with linear regression. Cinelli et al. [Cin+19] make concrete assumptions about the edges of a causal DAG and use causal identification tools to translate those assumptions into effect (hence, bias) estimates. However, they assume that all relationships in the DAG are linear. Rosenbaum and Rubin [RR83] assume a categorical covariate and a binary confounder. They don’t impose any explicit additional constraints on the propensity score or outcome model, but their general approach requires 4 sensitivity parameters for each level of the observed covariate; making the sensitivity analysis practical requires further assumptions.

A different line of work relaxes parametric assumptions at the price of requiring the analyst to make judgments about more abstract sensitivity parameters [e.g., Fra+19; She+11; VA11; DV15]. For example, Franks et al. [Fra+19] allow arbitrary models to be used for the initial analysis. Their sensitivity model is adapted from the missing data literature, and requires the analyst to specify $\mathrm{P}(T = t \mid Y(1-t), X)$—the posterior belief about probability of treatment assignment had the counterfactual outcome under no-treatment been observed. The sensitivity parameters used by these methods are more abstruse than the ones used in parametric-model-based sensitivity analysis. However, the subjective judgments required for each analysis are quite different, and these alternative approaches may be easier in some scenarios. In this sense, these methods are complimentary to the sensitivity analysis approach proposed in this paper.

5 Societal Consequences

This paper addressed sensitivity analysis for causal inference. We have extended Imbens' approach to allow the use of arbitrary machine-learning methods for the data modeling. Austen plots provide an entirely post-hoc and blackbox manner of conducting sensitivity analysis. In particular, they make it substantially simpler to perform sensitivity analysis. This is because the initial analysis can be performed without have a sensitivty analysis already in mind, and because producing the sensitivity plots only requires predictions from models that the practitioner has fit anyways.

The ideal positive consequence is that routine use of Austen plots will improve the credibility of machine-learning based causal inferences from observational data. Austen plots allow us to both use state-of-the-art models for the observed part of the data, and to reason coherently about the causal effects of potential unobserved confounders. The availability of such a tool may speed the adoption of machine-learning based causal inference for important real-world applications (where, so far, adoption has been slow).

On the negative side, an accelerated adoption of machine-learning methods into causal practice may be undesirable. This is simply because the standards of evidence and evaluation used in common machine-learning practice do not fully reflect the needs of causal practice. Austen plots partially bridge this gap, but they just one of the elements required to establish credibility.

References

[BG+92] J. Brooks-Gunn, F. ruey Liaw, and P. K. Klebanov. "Effects of early intervention on cognitive function of low birth weight preterm infants". In: *The Journal of Pediatrics* 3 (1992).

[Chi+10] H. A. Chipman, E. I. George, and R. E. McCulloch. "Bart: bayesian additive regression trees". In: *Ann. Appl. Stat.* 1 (2010).

[CH20] C. Cinelli and C. Hazlett. "Making sense of sensitivity: extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1 (2020).

[Cin+19] C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim. "Sensitivity analysis of linear structural causal models". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019.

[DV15] P. Ding and T. J. VanderWeele. "Sensitivity analysis without assumptions." In: *Epidemiology* (2015).

[Dor+16] V. Dorie, M. Harada, N. B. Carnegie, and J. Hill. "A flexible, interpretable framework for assessing sensitivity to unmeasured confounding". In: *Statistics in Medicine* 20 (2016).

[Fra+19] A. M. Franks, A. DAmour, and A. Feller. "Flexible sensitivity analysis for observational studies without observable implications". In: *Journal of the American Statistical Association* 0 (2019).

[Hil11] J. L. Hill. "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 1 (2011).

[Imb03] G. Imbens. "Sensitivity to exogeneity assumptions in program evaluation". In: *The American Economic Review* (2003).

[LaL86] R. J. LaLonde. "Evaluating the econometric evaluations of training programs with experimental data". In: *The American Economic Review* 4 (1986).

[RR83] P. R. Rosenbaum and D. B. Rubin. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 2 (1983).

[She+11] C. Shen, X. Li, L. Li, and M. C. Were. "Sensitivity analysis for causal inference using inverse probability weighting". In: *Biometrical Journal* 5 (2011).

[VA11] T. J. VanderWeele and O. A. Arah. "Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders". In: *Epidemiology* 1 (2011).

Appendix

Anonymous Author(s)
Affiliation
Address
`email`

A Average Treatment Effect on the Treated

In many situations, the average treatment effect (ATT) on the treated is a more convenient estimand than the ATE.

Bias The same logic we used to derive an expression for the bias of the ATE can be used to derive an expression for the bias of the ATT. For the bias estimand, we just change take the expectation over X in Theorem 3 conditioned on $T = 1$. In practice, the bias can be estimated by taking the mean over only treated units. Note that the reparameterization calculation does not change.

Calibration using observed data Reference values for the ATT can be computed in exactly the same way as for the ATE—i.e., it is not required to restrict the expectations to only the treated units. This is because the bias expression is given in terms of 'full data' α and $R^2_{Y,\mathrm{par}}$.

B Proofs

Theorem 4. *Under our sensitivity model, Assumption 1, the outcome influence is*

$$R^2_{Y,\mathrm{par}}(\alpha, \delta) = \delta^2 \sum_{t=0}^{1} \frac{\mathbb{E}\big[\psi_1\big(g(X)^t(1-g(X))^{1-t}(1/\alpha - 1) + 1[T=t]\big)\big]}{\mathbb{E}[(Y - Q(T,X))^2]},$$

where ψ_1 is the trigamma function.

Proof. First, we write:

$$\begin{aligned}\mathbb{E}(Y - \mathbb{E}[Y \mid T, X, U])^2 &= \mathbb{E}(Y - Q(T,X))^2 \\ &\quad - 2\delta\mathbb{E}[(Y - Q(T,X))(\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T])] \\ &\quad + \delta^2\mathbb{E}((\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T]))^2 \\ &= \mathbb{E}(Y - Q(T,X))^2 - \delta^2\mathbb{E}[\operatorname{var}(\operatorname{logit}\tilde{g}(X,U) \mid X,T)]. \end{aligned} \tag{B.1}$$

Where we've used,

$$\begin{aligned}&\mathbb{E}[(Y - Q(T,X))(\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T])] \\ &= \mathbb{E}[\mathbb{E}[(Y - Q(T,X)) \mid T,X,U](\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T])]\end{aligned}$$

and other standard conditional expectation manipulations.

The usefulness of (B.1) is that $\operatorname{var}(\operatorname{logit}\tilde{g}(X,U) \mid X,T)$ has an analytic expression. Namely, by Beta-Bernoulli conjugacy, this is the variance of a logit-transformed Beta distribution. The analytic expression for this variance is,

$$\operatorname{var}(\operatorname{logit}\tilde{g}(X,U) \mid X,T) = \psi_1\big(g(X)(1/\alpha - 1) + T\big) + \psi_1\big((1-g(X))(1/\alpha - 1) + 1 - T\big),$$

where ψ_1 is the trigamma function. The claimed result follows by plugging in this expression into (B.1).

□

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Theorem 5. *Under our sensitivity model, Assumption 1,*

$$\alpha = 1 - \frac{\mathbb{E}[\tilde{g}(X,U)(1-\tilde{g}(X,U))]}{\mathbb{E}[g(X)(1-g(X))]}.$$

Proof. The key insight is:

$$\begin{aligned}\mathrm{var}(\tilde{g}) &= \mathbb{E}[\mathrm{var}(\tilde{g} \mid g)] + \mathrm{var}(\mathbb{E}[\tilde{g} \mid g]) \\ &= \mathbb{E}[\alpha g(1-g)] + \mathrm{var}(g),\end{aligned}$$

where the first line is the law of total variance, and the second line uses the assumed Beta distribution of $\tilde{g} \mid g$. Accordingly,

$$\alpha = \frac{\mathrm{var}(\tilde{g}) - \mathrm{var}(g)}{\mathbb{E}[g(1-g)]}.$$

Now, observe that by the law of total variance,

$$\begin{aligned}\mathrm{var}(T) &= \mathbb{E}[\mathrm{var}(T \mid g)] + \mathrm{var}(\mathbb{E}[T \mid g]) \\ &= \mathbb{E}[g(1-g)] + \mathrm{var}(g),\end{aligned}$$

where we have used that $T \mid g$ is Bernoulli. By the same logic,

$$\mathrm{var}(T) = \mathbb{E}[\tilde{g}(1-\tilde{g})] + \mathrm{var}(\tilde{g}).$$

Whence,

$$\mathrm{var}(\tilde{g}) - \mathrm{var}(g) = \mathbb{E}[g(1-g)] - \mathbb{E}[\tilde{g}(1-\tilde{g})].$$

The result follows immediately. □