Bayesian Sensitivity Analysis for Offline Policy Evaluation

Jongbin Jung
jongbin@stanford.edu
Stanford University

Avi Feller
UC-Berkeley
afeller@berkeley.edu

Ravi Shroff
ravi.shroff@nyu.edu
New York University

Sharad Goel
scgoel@stanford.edu
Stanford University

ABSTRACT

On a variety of complex decision-making tasks, from doctors prescribing treatment to judges setting bail, machine learning algorithms have been shown to outperform expert human judgments. One complication, however, is that it is often difficult to anticipate the effects of algorithmic policies prior to deployment, as one generally cannot use historical data to directly observe what would have happened had the actions recommended by the algorithm been taken. A common strategy is to model potential outcomes for alternative decisions assuming that there are no unmeasured confounders (i.e., to assume ignorability). But if this ignorability assumption is violated, the predicted and actual effects of an algorithmic policy can diverge sharply. In this paper we present a flexible Bayesian approach to gauge the sensitivity of predicted policy outcomes to unmeasured confounders. In particular, and in contrast to past work, our modeling framework easily enables confounders to vary with the observed covariates. We demonstrate the efficacy of our method on a large dataset of judicial actions, in which one must decide whether defendants awaiting trial should be required to pay bail or can be released without payment.

1 INTRODUCTION

Machine learning algorithms are increasingly used by employers, judges, lenders, and other experts to guide high-stakes decisions [2, 3, 5, 6, 22]. These algorithms, for example, can be used to help determine which job candidates are interviewed, which defendants are required to pay bail, and which loan applicants are granted credit. When determining whether to implement a proposed policy, it is important to anticipate its likely effects using only historical data, as ex-post evaluation may be expensive or otherwise impractical. This general problem is known as offline policy evaluation. One approach to this problem is to first assume that treatment assignment is ignorable given the observed covariates, after which a variety of modeling techniques are theoretically justified, including regression, matching, and doubly robust estimation [1, 10, 23, 24].

As the assumption of ignorability is often unrealistic in practice, sensitivity analysis methods seek to gauge the effects of unmeasured confounding on predicted outcomes. The literature on sensitivity analysis dates back at least to the work of Cornfield et al. [8] on the link between smoking and lung cancer. In a seminal paper, Rosenbaum and Rubin [21] proposed a framework for sensitivity analysis in which a binary unmeasured confounder affects both a binary response variable and a binary treatment decision. Some extensions of the Rosenbaum and Rubin approach to sensitivity analysis include allowing for non-binary response variables, and incorporating machine learning methods into the estimation process [4, 9, 15].

A complementary line of work extends classical sensitivity analysis by taking a Bayesian perspective and averaging, rather than sweeping, over values of the sensitivity parameters [19, 20]. In this setup, even a weakly informative prior distribution over the sensitivity parameters can exclude estimates that are not supported by the data. Several authors have recently extended traditional methods for sensitivity analysis to offline policy evaluation [16, 17], though we do not know of any existing Bayesian approaches to the policy evaluation problem.

In this paper we develop a flexible Bayesian method to evaluate decision algorithms in the presence of unmeasured confounding. In contrast to most past work on sensitivity analysis, our approach conveniently accommodates latent confounding that varies with the observed covariates. This flexibility, coupled with a full Bayesian specification, helps to automatically exclude values of the latent variables that are not supported by the data and model of confounding.

To illustrate our method, we consider judicial bail decisions, starting with a detailed dataset of over 165,000 judgments in a large,
urban jurisdiction. Based on this information, we create realistic synthetic datasets for which potential outcomes and selection mechanisms are, by construction, fully known. We apply our method to these synthetic datasets, selectively masking potential outcomes and the data-generation process to mimic real-world applications. Under a variety of conditions, we find that our approach accurately recovers true policy effects despite the presence of unmeasured confounding. In these simulations, our approach outperforms classical, non-Bayesian sensitivity analysis techniques that have recently been adapted to the algorithmic decision-making context [16, 21]. Further, when applied to the special case of estimating average treatment effects, our general method performs on par with existing, state-of-the-art approaches that are tailored to that specific problem.

2 METHODOLOGY

A common goal of causal inference is to estimate the average effect of a binary treatment $T \in \{0, 1\}$ on a response $Y \in \mathbb{R}$. That is, one often seeks to estimate $\mathbb{E}[Y(1) - Y(0)]$, where $Y(0)$ and $Y(1)$ denote the potential outcomes under the two possible treatment values. Here we consider a generalization of this problem that arises in many applied settings. Namely, given a policy $\pi : \mathbb{R}^m \rightarrow \{0, 1\}$ that assigns treatment based on individual characteristics $X \in \mathbb{R}^m$, we seek to estimate the average response $V^\pi = \mathbb{E}[Y(\pi(X))]$. In our judicial bail application (described in detail in following sections), $\pi$ is an algorithmic rule that determines which defendants are released on their own recognizance (RoR) and which are required to pay bail, $Y \in \{0, 1\}$ indicates whether a defendant appears at trial, $X$ is a vector of observed covariates, and $V^\pi$ is the expected proportion of defendants who fail to appear in court when the rule is followed.

2.1 Policy evaluation without unmeasured confounding

To motivate the standard approach for estimating $V^\pi$, we decompose it into the sum of two terms:

$$V^\pi = \mathbb{E}[Y(T) \mid \pi(X) = T] \cdot \text{Pr}[\pi(X) = T] + \mathbb{E}[Y(1 - T) \mid \pi(X) \neq T] \cdot \text{Pr}[\pi(X) \neq T].$$

The proposed and observed policies are the same for the first term but differ for the second term. Thus estimating the first term is straightforward, since we directly observe the outcomes for this group, $Y^{\text{obs}} = Y(T)$. The key statistical difficulty is estimating the second term, for which we must impute the missing potential outcomes, $Y^{\text{mis}} = Y(1 - T)$. That is, in these cases we must estimate what would have happened had the proposed action been followed, rather than the observed action. This challenge is known as the fundamental problem of causal inference [14].

To estimate $V^\pi$, it is common to start with a sample of historical data $\{(x_i, t_i, y_i(t_i))\}_{i=1}^n$ on individual covariates, treatment decisions, and observed outcomes. We can then estimate $V^\pi$ via:

$$\hat{V}^\pi = \frac{1}{n} \sum \frac{y_i^{\text{obs}}}{n} + \sum \frac{y_i^{\text{mis}}}{n},$$

where $y_i^{\text{obs}}$ is the observed outcome for individual $i$ when the proposed and observed policies agree. A simple approach to estimating the unobserved potential outcomes $y_i^{\text{mis}}$ in the second term is to directly model the outcomes conditional on treatment and observed covariates—a strategy sometimes referred to as response surface modeling [13].

This direct modeling approach implicitly assumes that the treatment is ignorable given the observed covariates (i.e., that there is no unmeasured confounding). Formally, ignorability means that

$$Y(0), Y(1) \perp \perp T \mid X.$$  

In other words, conditional on the observed covariates, those who receive treatment are similar to those who do not. In many realistic situations, ignorability is a strong assumption. The main contribution of this paper is a Bayesian approach to assessing the sensitivity of estimated policy outcomes $V^\pi$ to violations of ignorability, which we present next.

2.2 Policy evaluation with unmeasured confounding

When ignorability does not hold, the resulting estimates of $V^\pi$ can be strongly biased. To address this issue, the sensitivity literature typically starts by assuming that there is an unmeasured confounder $U \in \{0, 1\}$ (or, more generally, $U \in \mathbb{R}$) for each individual such that ignorability holds given both $X$ and $U$:

$$Y(0), Y(1) \perp \perp T \mid X, U.$$  

One then generally imposes additional structural assumptions on the form of $U$ and its relationship to decisions and outcomes. We follow this basic template to obtain sensitivity estimates for policy outcomes.

At a high level, we model the observed data $\{(x_i, t_i, y_i(t_i))\}_{i=1}^n$ as draws from parametric distributions depending on the measured covariates $x_i$ and the unmeasured, latent covariates $u_i$:

$$y_i(0) \sim f(x_i, u_i; \alpha)$$

$$y_i(1) \sim g(x_i, u_i; \beta)$$

$$t_i \sim h(x_i, u_i; \gamma)$$

$$y_i = y_i(t_i)$$

where $\alpha, \beta, \gamma, \text{and } u_i$ are latent parameters with weakly informative priors. The inferred joint posterior distribution on these parameters then yields estimates of $y_i(0)$ and $y_i(1)$ for each individual $i$. Finally, the posterior estimates of the potential outcomes yield a posterior estimate of $V^\pi$ via Eq. (1) that accounts for unmeasured confounding.

Our strategy, while conceptually straightforward, differs in three important ways from classical sensitivity methods. First, we directly model the treatment and potential outcomes, allowing for flexibility in the functional forms of $f$, $g$ and $h$. Second, our Bayesian approach automatically excludes values of the latent variables that are not supported by the data [19]. In contrast, previous methods generally require more extensive parameter tuning to obtain reasonable estimates. Finally, our focus is on the more general problem of estimating policy outcomes rather than the average treatment effects considered in past work.

We now describe the specific forms of $f$, $g$, and $h$ that we use throughout our analysis. First, we reduce the dimensionality of the observed covariate vectors $x$ down to three key quantities:
\[ \mu_0(x) = \mathbb{E}[Y(0) \mid X = x], \] the conditional average response if \( T = 0; \]
\[ \mu_1(x) = \mathbb{E}[Y(1) \mid X = x], \] the conditional average response if \( T = 1; \]
\[ e(x) = \text{Pr}(T = 1 \mid X = x), \] the propensity score. The first two quantities, \( \mu_0(x) \) and \( \mu_1(x) \), depend on potentially unobserved outcomes \( Y(0) \) and \( Y(1) \), and so typically cannot be perfectly estimated from available data—indeed, accurately estimating these terms is our ultimate goal. Thus, as a first step, we generate approximations \( \hat{\mu}_0(x) \) and \( \hat{\mu}_1(x) \) by assuming ignorability and fitting standard prediction models, like regularized regression, to subsets of the data with \( T = 0 \) and \( T = 1 \), respectively.

To support complex relationships between the predictors and outcomes, we divide the data into \( K \) approximately equally sized groups, ranking and binning by the estimated outcome \( \hat{\mu}_0 \). As Franks et al. [11] show, this flexible model form also plays an important theoretical role, helping to avoid issues of identifiability that are inherently at odds with the exercise of sensitivity analysis. Denoting the group membership of observation \( i \) by \( k[i] \in \{1, 2, \ldots, K\} \), and a Bernoulli distribution with mean \( p \) as \( \text{B}(p) \), we model the observed data as follows:

\[
\begin{align*}
y_i(0) &\sim \text{B} \left( \text{logit}^{-1}(\alpha_0, k[i] + \alpha_{u,i} k[i] \hat{u}(x)) + \alpha_{u,i} k[i] u[i] \right) \\
y_i(1) &\sim \text{B} \left( \text{logit}^{-1}(\beta_0, k[i] + \beta_{u,i} k[i] \hat{u}(x) + \beta_{u,i} k[i] u[i]) \right) \\
t_i &\sim \text{B} \left( \text{logit}^{-1}(\gamma_{0,i} k[i] + \gamma_{u,i} k[i] \hat{u}(x) + \gamma_{u,i} k[i] u[i]) \right) \\
y_i &\equiv y_i(t_i).
\end{align*}
\]

In each of the first three equations above, variables are modeled as draws from a Bernoulli distribution whose mean depends on both the reduced covariates—\( \hat{\mu}_0(x) \), \( \hat{\mu}_1(x) \), and \( \hat{e}(x) \)—and the unmeasured confounder \( u \). In the judicial context, for example, one can imagine that \( u \) corresponds to an unobserved risk of failing to appear, with judges more likely to demand bail from riskier defendants.

Finally, to complete the Bayesian model specification, we must describe the prior distribution on the parameters (additional detail is provided in the Appendix). On each of the unmeasured confounders \( u_i \), we set a \( \text{N}(0, 1) \) prior. On the coefficients \( \alpha, \beta, \) and \( \gamma \) we use a random-walk prior; intuitively, these priors ensure that adjacent groups have similar coefficient values, mitigating the dependence of results on the exact number of groups \( K \).

By definition, it is impossible to infer the degree of unmeasured confounding from the observed data alone. As such, most methods for sensitivity analysis involve one or more parameters that allow one to explicitly set the degree of confounding, which may, for example, be informed by domain knowledge. However, in our case, the necessary side information enters implicitly, from the structural form of our model of confounding, especially the choice of \( K \) discussed above. By positing smoothness and accordingly restricting the number of groups—which is reasonable in many applications—one can leverage the functional form of the model to avoid having to explicitly set the degree of confounding. This general approach is a valuable complement to the traditional perspective, but there is no free lunch: if the model of confounding is unreasonable for a particular application, the resulting sensitivity bounds will suffer. Below we discuss a strategy for appropriately setting the model parameters via simulation.

### 3 AN APPLICATION TO JUDICIAL DECISION MAKING

To demonstrate our general approach to offline policy evaluation, we now consider in detail the case of algorithms designed to aid judicial decisions [12, 16, 18]. In the U.S. court system, pretrial release determinations are among the most common and consequential decisions for criminal defendants. After arrest, a defendant is usually arraigned in court, where a prosecutor presents a written list of charges. If the case proceeds to trial, a judge must decide whether the defendant should be released on his or her own recognizance (RoR) or subject to money bail, where release is conditional on providing collateral meant to ensure appearance at trial. Defendants who are not RoR’d and who cannot post bail themselves may await trial in jail or pay a bail bondsman to post bail on their behalf. Judges must therefore balance the burdens of bail on a defendant and society against the risk that the defendant may fail to appear (FTA) for trial.

Here we consider algorithmic policies for assisting these judicial decisions, recommending either RoR or bail based on recorded characteristics of a defendant and case. The policy evaluation problem is to estimate, based only on historical data, the proportion of defendants who would fail to appear if algorithmic recommendations were followed. This is statistically challenging because one does not always observe what would have occurred had the algorithmic policy been followed. In particular, if the policy recommends releasing a defendant who was in reality detained, or recommends detaining a defendant who was in reality released, the relevant counterfactual is not observed. Further, since judges may—and likely do—base their decisions in part on information that is not recorded in the data, direct outcome models ignoring unmeasured confounding will likely be biased. We thus allow for a real-valued unobserved covariate \( u \) that affects both a judge’s decision (RoR or bail) and also the outcome (FTA or not) conditional on that decision. Our goal is to assess the sensitivity of flight risk estimates to such unmeasured confounding.

#### 3.1 Policy construction

Our analysis is based on 165,000 adult cases involving nonviolent offenses charged by a large urban prosecutor’s office and arraigned in criminal court between 2010 and 2015. These cases do not include instances where defendants accepted a plea deal at arraignment, where no pretrial release decision is necessary. For each case, we have 49 features describing characteristics of the current charges (e.g., theft, gun-related), and 15 features describing characteristics of the defendant (e.g., gender, age, prior arrests). We also observe whether the defendant was RoR’d, and whether the defendant failed to appear at any of his or her subsequent court dates. Applying notation introduced in the previous section, \( x_i \) refers to a vector of all observed characteristics of the \( i \)-th defendant, \( t_i = 1 \) if bail was set, and \( y_i = 1 \) if the defendant failed to appear at court.

To carry out our analysis, we first randomly select 10,000 cases from the full dataset which we set aside as our final test data. The remaining data are split into two folds of approximately equal size.

---

1Although we describe our method for binary potential outcomes, the general approach can accommodate real-valued outcomes as well.

2In many jurisdictions, judges may also consider the risk that a defendant will commit a new crime if released when deciding whether or not to set bail, but not in the jurisdiction we consider.
We begin by constructing a family of algorithmic decision rules. On the first training fold, we fit an $L^1$-regularized (lasso) logistic regression model to estimate the probability of FTA given release, $\hat{\mu}_0(x_i)$. That is, we fit a logistic regression with the left-hand side indicating whether a defendant failed to appear at any court dates, and the right-hand side comprised of all available covariates for the subset of defendants that the judge released. We use the superscript $\pi$ to indicate that these estimates are computed on the first fold of data, and are used exclusively to define policies, not to evaluate them. With these estimates in hand, we construct a family of policies \{\pi_s\}, indexed by a risk threshold $s$ for releasing individuals:

$$\pi_s(x_i) = \begin{cases} 
1 & \text{if } \hat{\mu}_0(x_i) > s \\
0 & \text{if } \hat{\mu}_0(x_i) \leq s.
\end{cases}$$

These policies, which are based on a ranking of defendants by risk, are similar to pretrial algorithms used in practice [7].

Given the family of policies \{\pi_s\} defined above, we next estimate the proportion of defendants who would fail to appear under each policy, accounting for unmeasured confounding. To do so, on the second training fold, we fit three $L^1$-regularized logistic regression models to estimate each individual’s likelihood of failing to appear if RoR’d or required to pay bail—$\hat{\mu}_0(x_i)$ and $\hat{\mu}_1(x_i)$, respectively—as well as each individual’s likelihood of having bail set, $\hat{\ell}(x_i)$.

Finally, on the test fold consisting of 10,000 cases, we fit our Bayesian sensitivity model using the estimated quantities $\hat{\mu}_0(x_i)$, $\hat{\mu}_1(x_i)$, and $\hat{\ell}(x_i)$. But, to do so, we must first select appropriate parameters for our model of confounding, which we describe next.

### 3.2 Calibrating the model via simulation

As mentioned above, information about unmeasured confounding enters through the structural form of the model. In particular, the amount of unmeasured confounding is related to the number of observations per bin, $n/K$, with fewer observations per bin corresponding to more unmeasured confounding. We now describe how to calibrate our model of unmeasured confounding using synthetic datasets, before applying our method to real data.

We begin by creating a synthetic version of our original judicial dataset of judicial decisions where both potential outcomes for each defendant—$y_i(0)$ and $y_i(1)$—are known. We do this by first estimating $\hat{\mu}_0(x_i)$, $\hat{\mu}_1(x_i)$, and $\hat{\ell}(x_i)$ using $L^1$-regularized logistic regression, as above. Then, for each individual in the original dataset, we create synthetic potential outcomes and treatment assignments via independent Bernoulli draws based on these estimated probabilities. We denote the resulting synthetic (uncensored) dataset by $\Omega = \{(x_i, t_i', y_i'(0), y_i'(1))\}$. Because both potential outcomes are known, we can exactly calculate $T^\pi$ for any policy $\pi$ applied to $\Omega$. Moreover, our synthetic dataset satisfies ignorability by construction.

We next censor $\Omega$ in two ways. First, we restrict the observed covariates to a subset $x_i'$. Specifically, we consider three sets of restricted covariates: age; age and gender; and age, gender, and prior number of missed court appearances. Second, for each individual we remove the potential outcome for the action not taken, keeping only $y_i'(t_i')$. Thus, for each of our three choices of $x_i'$, we have a dataset of the form $\Omega' = \{(x_i', t_i', y_i'(t_i'))\}$. The covariates not included in $x_i'$ correspond to unmeasured confounding. Starting from these three synthetic datasets, we carry out the policy construction and sensitivity analysis procedure described above.

The results of our simulation study are shown in Figure 1, with each panel corresponding to a different choice of $x_i'$ and thus a different degree of unmeasured confounding. In all cases, we set $n/K = 1000$ (i.e., 1,000 observations per bin). The blue lines show estimates based on direct outcome modeling, ignoring unmeasured confounding, and the gray bands represent 50% and 95% credible intervals based on our sensitivity analysis. Importantly, because these results are derived from synthetic datasets, we can also compute the true policy outcomes, which are indicated in the plot by the red lines.

Across the three levels of unmeasured confounding that we consider, our sensitivity bands cover the ground truth line. Moreover, the bands accurately reflect the true level of confounding, with wider bands when $x_i'$ consists only of age (i.e., there is extreme confounding), and narrower bands when $x_i'$ is comprised of age, gender, and prior number of missed court appearances (the confounding is less severe). Thus, based on this simulation, setting $n/K = 1000$ appears reasonable to account for the true level of unmeasured confounding in the synthetic datasets.

To illustrate the effect of our parameter choice on the results, we measure the average width of the 95% credible intervals in our synthetic datasets over a wide range of $n/K$ (i.e., observations per bin). The results are presented in Figure 2, which is based on a
dataset of 1000 observations. As the number of observations per bin increases—as a result of lowering the number of groups K—the credible interval widths decrease. Importantly, the credible intervals for all the settings we consider still cover the ground truth in the synthetic datasets, as in Figure 1.

3.3 Results on the observed data

Based on the above calibration exercise, we fit our Bayesian sensitivity model on the real data using $n/K = 1000$ observations per bin (i.e., with $K = 10$ bins on the held-out set of 10,000 data points). We note that we found qualitatively similar results when we repeated our analysis for $K = 5$ and $K = 100$ bins.

The results of our analysis are plotted in Figure 3. The blue dashed line shows, for each policy $\pi_k$, the proportion RoR’d and the estimated proportion that FTA under the policy, $\hat{V}_{\pi_k}$, estimated assuming no unobserved confounding. For reference, the black point shows the status quo: judges in our dataset release 69% of defendants, resulting in an overall FTA rate of 13%. The light and dark gray bands show, respectively, the 95% and 50% credible intervals that result from the sensitivity analysis procedure under our model of unobserved confounding.

Figure 3 illustrates three key points. First, for policies that RoR almost all defendants (toward the right-hand side of the plot), the blue line lies below our sensitivity bands, in line with expectations. If there is no unmeasured confounding, we would expect those who were in reality detained to be riskier than those who appear from the observed data alone; as a result, the direct outcome model underestimates the proportion of defendants that would fail to appear if all (or almost all) such defendants are RoR’d. Conversely, for policies that recommend bail for almost all defendants (toward the left-hand side of the plot), the blue line lies above the sensitivity bands, as we would expect, because those who were in reality released are likely less risky than they appear from the observed data alone. Second, as the policies move further from the status quo, toward the left- and right-hand extremes of the plot, the sensitivity bands grow in width, indicating greater uncertainty. This pattern reflects the fact that the data provide less direct evidence of flight risk as policies diverge from observations, heightening the potential impact of unmeasured confounding. Finally, even after accounting for unmeasured confounding, there are algorithmic policies that do substantially better than the status quo, in the sense that they RoR more defendants and simultaneously achieve a lower overall FTA rate.

4 COMPARISON WITH ALTERNATIVE METHODS

Sensitivity analysis for offline policy evaluation is relatively new [16, 17]. There are, however, several methods for assessing the sensitivity of average treatment effects to unmeasured confounding, which is a specific case of policy evaluation. Here we compare our approach to Jung et al.’s sensitivity method for offline policy evaluation, as well as to two recent Bayesian sensitivity methods designed for average treatment effects [9, 19]. For space, we omit the details of these methods, but note that we implement the two sensitivity analysis procedures exactly as they were originally described.

Figure 4 shows the results of the Jung et al. procedure applied to the three synthetic datasets described above. In particular, we compute the minimum and maximum policy estimates obtained by sweeping over two parameter regimes suggested in their paper. In the first regime (dark bands), we allow the unobserved confounder to double the odds of bail and the odds of failing to appear at trial, both if RoR’d or required to pay bail. We also consider a more extreme situation (lighter bands), where the unobserved confounding can increase the odds of being detained up to three times, and also triple the odds a defendant fails to appear. In each scenario, the red lines indicate the true outcomes of each policy, computed from the uncensored synthetic dataset, the blue lines show estimates based on direct outcome models ignoring unmeasured confounding, and the gray bands indicate the minimum and maximum values of each policy over all parameter settings in the corresponding regime.

In contrast to our own sensitivity analysis results, the sensitivity bands in Figure 4 often fail to capture the ground truth policy estimates. Further, and more importantly, the sensitivity bands do not
We have addressed the problem of offline policy evaluation by coupling classical statistical ideas on sensitivity analysis with modern methods of machine learning and large-scale Bayesian inference. The result is a conceptually simple yet powerful technique for evaluating algorithmic decision rules.

By definition, it is impossible to precisely quantify unmeasured confounding, and so all methods of sensitivity analysis require assumptions that are inherently untestable [11]. Traditional methods handle this situation by requiring practitioners to specify parameters describing the structure and scale of the assumed confounding, informed, for example, by domain expertise. In our case, the necessary side information enters through the assumed form of the data-generating process. By adopting an expressive model form that we calibrate through simulation, our approach balances the need to provide at least some information about the structure of the potential confounding with the impossibility of specifying it exactly. This middle ground appears to work well in practice, but it is useful to remember the conceptual underpinnings of our strategy when applying it to new domains.

Over the last two decades, sophisticated methods of machine learning have emerged and gained widespread adoption. More recently, these methods have been applied to traditional problems of causal inference and their modern incarnations, like offline policy evaluation. Prediction and causal inference are two sides of the same coin, but the links between the two are still under-developed. Here we have bridged one such gap by porting ideas from classical sensitivity analysis to algorithmic decision making. Looking forward, we hope that sensitivity analysis is more tightly integrated into machine-learning pipelines and, more generally, that our work spurs further connections between methods of causal inference and prediction.

REFERENCES

APPENDIX

In the main text, we described the likelihood function for our Bayesian model of unmeasured confounding. Here we complete the model specification by describing the prior distribution on the parameters.

As mentioned in the main text, we set a N(0,1) prior on each of the unmeasured confounders $u_i$. On the coefficients $\alpha$, $\beta$, and $\gamma$ we use a random-walk prior. Formally, the random-walk prior on $\alpha_0$ is given by

$$
\alpha_0 \sim N(0, 1)
$$

$$
\alpha_{0,j} \sim N(\alpha_{0,j-1}, \tau_{\alpha_0}^2) \quad \text{for } j \in \{2, 3, \ldots, K\}
$$

$$
\tau_{\alpha_0} \sim N_+(0, \sigma^2_\tau)
$$

where $N_+(0, \sigma^2_\tau)$ indicates the half-normal distribution with standard deviation $\sigma_\tau$. We analogously set priors for the parameters $\beta_0, \beta_{\hat{v}}, \gamma_0$, and $\gamma_{\hat{e}}$.

For the coefficients on the unobserved confounders $u_i$, we set random-walk priors with an additional constraint to ensure positivity, so that $Pr(T = 1)$, $Pr(Y(0) = 1)$, and $Pr(Y(1) = 1)$ all increase with $u_i$. This constraint, while not strictly necessary, facilitates estimation of the posterior distribution. Formally, for $\alpha_u$ we have

$$
\alpha_{u,1} \sim N_+(0, 1)
$$

$$
\alpha_{u,j} \sim N_+(\alpha_{u,j-1}, \tau_{\alpha_u}^2) \quad \text{for } j \in \{2, 3, \ldots, K\}
$$

$$
\tau_{\alpha_u} \sim N_+(0, \sigma^2_\tau)
$$

We similarly set sign-constrained random-walk priors on $\beta_u$ and $\gamma_u$. For the main results in this paper, we set values of $\sigma_\tau = 1$ and $K = 10$. However, the results are not substantially affected by the choice of $K$ or the prior distribution (parameterized by $\sigma_\tau$).