Credible Assumption Mixtures:

Combining Point and Partial Identification Assumptions for Interpretable Bayesian Causal Inference

Marco Morucci

Abstract

Bayesian causal inference is a powerful and elegant tool as it allows analysts to directly model unobserved counterfactual parameters via prior distributions. Choosing these priors is, however, often hard and either requires overly strong assumptions or leads to choices of vague prior with no substantive interpretation. In this paper we introduce a methodology to specify counterfactual priors by adapting common point and partial identification assumptions used in frequentist causal inference. Our method is based on a Bayesian mixture model for the counterfactual outcome and uses several assumptions at once to produce a single, robust, result for the average treatment effect. Priors specified in the way we suggest are easily interpretable as substantive statements on the problem studied, as well as a level of importance of each assumption employed in producing the result. A natural extension of the proposed method as a form of sensitivity analysis is also outlined in this paper. The effectiveness of the proposed methodology is evaluated through simulation studies and through an application to the relationship between electoral support and targeted public goods in Malawi.

1 Introduction

Much of existing research in political science, economics and quantitative social science in general deals with trying to establish causal relationships between two factors of interest, a treatment T and an outcome Y. As is widely known, this can never be done with certainty as counterfactual scenarios (how would have WWII ended had Germany not attempted to invade Russia?) are never observed by definition. Experimental research deals with this problem by actively manipulating treatment values, but this is not possible in observational settings, where treatment values cannot be directly changed. Bayesian causal inference deals with this problem by directly modeling counterfactual parameters with a prior distribution (Rubin, 1978). If treatment is randomized, then this distribution can be updated with the observed likelihood, but in observational inference settings this is not always true. How should such a distribution then be chosen? Counterfactual parameters are unobserved, and it is particularly hard to model them explicitly without either making unrealistic assumptions, such as pre-specifying location, scale and shape of a counterfactual distribution, or without being too vague, such as assuming that the counterfactual parameter of interest could lie anywhere in its support with equal probability.

In this paper we introduce a modeling strategy aimed at dealing with this issue. Our method uses point and partial identification assumptions common in frequentist causal inference to specify priors on counterfactual parameters that are informative only as far as they are interpretable, that is to say, any restrictions on

the prior distribution of the counterfactual parameter have a clear and direct interpretation in the proposed framework. Our approach is useful in at least three ways:

- Prior distributions specified with our approach have clear and direct interpretations as assumptions on the relationship between treatment, response and control covariates.
- Our approach allows analysts to specify multiple identification assumptions at once and to combine them into one result. This improves over the current standard in frequentist causal inference, where only one assumption at a time is used in virtually all settings.
- Our approach makes the level of importance of each assumption in the prior explicit: how much a result relies on each assumption is clearly quantified, allowing robust data-driven results to be easily distinguished from assumption-driven results.

The proposed approach works by combining point-mass distributions with uniform or truncated distributions into a mixture prior for the counterfactual parameter, with each component of the mixture being defined by one frequentist identification assumption. Common causal quantities of interest can be obtained easily in our framework with widely used MCMC tools. Another advantage of our proposed methodology is that it can naturally be used to assess the sensitivity of a result to relaxations of the assumptions used to obtain it: we introduce a way to explicitly use our proposed prior specification as a method for sensitivity analysis by obtaining the optimal set of mixture weights that produce a pre-specified credible interval while putting the least amount of weight on the strongest assumption. Finally, our proposed methodology can be easily applied to common questions in empirical political science: we show this by modeling strategic allocation of public goods to electoral districts by politicians in Malawi: we show that two different sets of identification assumptions lead to similar results, and that affirmative conclusions are almost entirely dependent on the identification assumptions employed.

The paper is organized as follows: Section 2 introduces notation, assumptions, quantities, and the model commonly used in Bayesian causal inference. Section 3 introduces our prior specification for counterfactual quantities, and gives prior formulations for several popular point and partial identification assumptions. Section 4 presents a way to use our prior mixtures as a tool for sensitivity analysis. Section 5 presents results of applying our method to simulated datasets, and Section 6 presents an application of our method to allocation of public goods in Malawi.

1.1 Issues of Model Interpretability and Identification in Causal Inference

We highlight three issues with Bayesian causal inference and causal inference in general that this paper aims to deal with:

First, most of the practice in Bayesian observational causal inference relies on conditional ignorability of treatment assignment (Rubin, 1978): this allows analysts to use the same distribution for both the observed and counterfactual outcomes, and allows the model to be updated with the observed data in the canonical way. This approach, however, fails to take full advantage of the flexibility of the Bayesian framework for causal inference: by specifying different distributions for the observed and counterfactual parameters, the

uncertainty around the second can be quantified more precisely and transparently, leading to more robust and trustworthy results. Having to specify priors on counterfactual parameters that do not rely on conditional ignorability often leads to choices of priors that are exceedingly vague and not directly interpretable as statements about the data generating process.

Second, most applied papers in frequentist observational research are centered around one specific identification assumption, often with large chunks of the paper devoted to defending its validity. Because of this, there is often little to no discussion of other alternative identification assumptions that could be used to obtain values for the counterfactual of interest. Alternative assumptions could offer different, but still informative conclusions on the same data, sometimes even strengthening the results produced under the authors' chosen assumption.

Finally, there is often little to no indication of how reliant the results produced are on the specific identification assumption chosen, both in frequentist and in Bayesian causal inference. The assumption in question is always required to hold fully on the data and results are presented conditional on this fact. However results obtained under a weak assumption could be similar to those obtained under a stronger one. This would lend credence to the fact that results are driven by the data, instead of the identification assumptions chosen. Tools from classical and modern sensitivity analysis (see, e.g. Rosenbaum and Rubin, 1983; VanderWeele and Ding, 2017; Cinelli and Hazlett, 2018) offer plenty of ways to produce summaries of the reliance of results on counterfactual assumptions, but are rarely used in practice in much empirical social science research. The few political science works that employ partial identification assumptions include Mebane and Poast (2013), and Coppock et al. (2017).

This paper introduces a method that deals with all the issues presented above by allowing researchers to mix several different identification assumptions with pre-assigned levels of reliance into one unique result. This allows for multiple identification strategies to be considered at once, and statements about the reliance of results on each of the assumptions to be produced simulateneously. When results are shown to not rely too much on strong assumptions they can be thought of as being drawn from the data and, as such, to likely hold true even if the chosen assumption is not perfectly justified. This allows analysts to highlight strong results that hold in practice when assumptions are weak without having to focus too heavily on defending the chosen identification assumption. All these features naturally lead to priors on counterfactual parameters that are informative and easily interpretable without requiring assumptions that are too stringent.

2 Bayesian Causal Inference, Assumptions and Partial Identification

In this section we present the problem of causal inference from a Bayesian perspective, define the notion of identification assumption, and outline the inference target that the rest of the paper will focus on. We adopt the potential outcomes framework and notation: let $Y(t) \in \mathbb{R}$ be a potential outcome function of some discrete treatment $t \in \{1, \ldots, M_T\}$, and a discrete set of P covariates, $X \in \mathcal{X}$, such that $Y(t)$ is drawn from some distribution with CDF $F_{Y(t)|X}$ and PDF $f_{Y(t)|X}$. Throughout the rest of the paper we use f_x and F_x to denote PDF and CDF of any random variable x respectively. We observe, for a sample of N units, the variable $T \in \{1, \ldots, M_T\}$, which describes which of the treatments each of the N units is

assigned to. The observed outcome for each unit i, will then be $Y_i = \sum_{t=1}^{M_T} Y(t)\mathbb{I}[T_i = t]$, where $\mathbb{I}(E)$ is the indicator function for event E, i.e. we never observe counterfactual potential outcomes for every unit: we only see the outcome for the treatment a unit is assigned to. Our target for inference is the Conditional Average Treatment Effect (CATE) for some two treatments $t, s \in \{1, \ldots, M_T\}$ and a value of the covariates $x \in \mathcal{X}$, which we denote by $\tau(t, s, x)$:

$$\tau(t, s, x) = \mathbb{E}[Y(t) - Y(s)|X = x] = \mathbb{E}[Y(t)|X = x] - \mathbb{E}[Y(s)|X = x].$$

Note that, if the CATE for an arbitrary value of X can be estimated, then aggregate functions of it, such as the Average Treatment Effect (ATE), denoted by $\tau(t, s)$, can also be estimated easily:

$$\tau(t, s) = \mathbb{E}[\tau(t, s, X)] = \mathbb{E}[\mathbb{E}[Y(t)|X] - \mathbb{E}[Y(s)|X]],$$

where the outside expectation is taken over the distribution of X.

We begin with the problem of estimating the CATE for two treatments, t and s. We opt to not model this quantity parametrically, and therefore we wish to estimate the response functions $\mathbb{E}[Y(t)|X = x]$ and $\mathbb{E}[Y(s)|X = x]$ separately. Consider treatment level t: by the law of total expectation, $\mathbb{E}[Y(t)|X = x]$ can be decomposed as:

$$\begin{aligned}\mathbb{E}[Y(t)|X = x] &= \mathbb{E}[Y(t)|X = x, T = t] \Pr(T = t|X = x) + \mathbb{E}[Y(t)|X = x, T \neq t] \Pr(T \neq t|X = x) \\ &= \phi(t, x)e(t, x) + \theta(t, x)(1 - e(t, x)),\end{aligned}$$

where $\phi(t, x) = \mathbb{E}[Y(t)|T = t, X = x]$, $\theta(t, x) = \mathbb{E}[Y(t)|T \neq t, X = x]$ and $e(t, x) = \Pr(T = t|X = x)$. Clearly, our quantity of interest is made up of two parts: first $\phi(t, x)$, which we observe information on and can estimate with our data, since, by definition of Y:

$$\begin{aligned}\mathbb{E}[Y|T = t, X = x] &= \sum_{s=1}^{M_T} \mathbb{E}[Y(t)\mathbb{I}(t = s)|T = t, X = x] \\ &= \sum_{s \neq t} \mathbb{E}[Y(s) \times 0|T = t, X = x] + \mathbb{E}[Y(t) \times 1|T = t, X = x] \\ &= \mathbb{E}[Y(t)|T = t, X = x].\end{aligned}$$

Second, we have $\theta(t, x) = \mathbb{E}[Y(t)|T \neq t, X = x]$: this quantity represents the average response to treatment t in the event in which treatment t is not administered to the experimental units. Clearly this is impossible to observe directly and is only expressible as a potential response to a treatment. In a classical statistical sense, we can also express the issue above by saying that $\phi(t, x)$ is identified since it is equal to $\mathbb{E}[Y|X = x, T = t]$ a quantity that can be consistently estimated with the observed data, and $\theta(t, x)$ is unidentified. Furthermore, since $\mathbb{E}[Y(t)|X = x]$ contains an unidentified part, then it is also unidentified.

We wish to perform Bayesian inference on the response function. Let $\mathcal{D} = \{(X_i, Y_i, T_i)\}_{i=1}^N$ be the observed data, and let $\mu(t, x) = \mathbb{E}[Y(t)|X = x]$ be the response function that we wish to estimate for

treatment t. Note that since we are in a Bayesian setting, this is now a random variable. We are interested in finding $\Pr(\mu(t,x)|\mathcal{D})$, the posterior distribution of μ_t conditional on the observed data. We employ the following model:

$$\begin{aligned} Y_i|T_i = t, X_i = x, \phi(t,x), \boldsymbol{\eta} &\overset{iid}{\sim} f_Y(t,x,\phi(t,x),\boldsymbol{\eta}) \\ \phi(t,x)|\boldsymbol{\eta} &\sim g(t,x,\boldsymbol{\eta}) \\ \theta(t,x)|\boldsymbol{\eta} &\sim h(t,x,\boldsymbol{\eta}) \\ \mu(t,x) &= \phi(t,x)e(t,x) + \theta(t,x)(1-e(t,x)), \end{aligned}$$

where f_Y is some distribution on the observed data, dependent on the identified parameter, $\phi(t,x)$, and with CDF F_Y; g is a prior on the identified parameter $\phi(t,x)$, with CDF G; h is a prior on the unidentified parameter, $\theta(t,x)$, with CDF H, and $\boldsymbol{\eta}$ is a vector of hyperparameters. Throughout the rest of this paper we will assume that $e(t,x)$ and quantities that depend exclusively on it are known.[1]

Consider now the posterior for $\phi(t,x), \theta(t,x)$ we know that it takes form (Gustafson, 2010):

$$\begin{aligned} \Pr(\phi(t,x), \theta(t,x)|\mathcal{D}) &= \frac{f_Y(\mathbf{Y};t,x,\phi(t,x),\boldsymbol{\eta})g(\phi(t,x);t,x,\boldsymbol{\eta})h(\theta(t,x),t,x,\boldsymbol{\eta})}{\int f_Y(\mathbf{Y};t,x,\phi(t,x),\boldsymbol{\eta})d\phi(t,x)} \\ &= g_{\mathcal{D}}(\phi(t,x);t,x,\boldsymbol{\eta})h(\theta(t,x);t,x,\boldsymbol{\eta}), \end{aligned}$$

where $g_{\mathcal{D}}(\phi(t,x),t,x,\boldsymbol{\eta})$ is the posterior pdf of $\phi(t,x)$ given the observed data. This posterior form shows the main feature of a partially identified model such as this: the only distribution being updated by the data is g, the distribution of the observed parameter. The distribution for the unobserved part of the model is instead not updated and has the same form of the prior. The main inference task is to compute two distributions: $g_{\mathcal{D}}(t,x,\boldsymbol{\eta})$ first, and $h(t,x,\boldsymbol{\eta})$ second. The first can be computed with any of the ordinary Bayesian Inference tools, such as the parametric regression models of the outcome variable that are common in political science. The main concern here is choice of h: we will shortly see that this prior on the unobserved counterfactual can be chosen in a way that allows us to employ posterior information on $\phi(t,x)$ to construct h.

Obtaining the posterior density of $\tau(t,s,x)$ from this model is straightforward. We first write the posterior pdf of $\mu(t,x)$ evaluated at a point u as a convolution of the distributions of the identified and unidentified parts of the model. The distribution of $\tau(t,s,x)$ is another standard quantity in Bayesian causal inference that can be found in a similar way.[2] In most cases it will only be feasible to compute these quantities via MCMC integration. Fortunately, it is straightforward in this framework to quickly sample from all the densities involved in the computation for most commonly used models of Y.

[1] The model above can be easily extended to incorporate inference for $e(t,x)$ and Bayesian inference can be performed easily even with this extended model. Otherwise $e(t,x)$ can be estimated with any of the conventional methods.

[2] Exact formulations for these quantities are given in the supplementary materials in equations (9) and (10) respectively.

3 Credible Assumption Mixtures

How should we choose the counterfactual prior, h? On the one hand, it would be tempting to be as vague as possible so as to not make any assumptions about the unobserved part of our model. This is, however, dangerous because first, a prior that is too uninformative will likely lead to an uninformative posterior, and second, because a prior that's too diffuse will lack interpretability. On the other hand, we would like to avoid imposing restrictions on the distributions of $\theta(t, x)$ without theoretical or empirical justification. The suggestion that we make in this paper is that different we can use frequentist identification assumptions to inform h with estimated values of $\phi(t, x)$.

In frequentist inference, an identification assumption is simply a way to relate $\phi(t, x)$ to $\theta(t, x)$, implying that knowledge of this relationship and of $\phi(t, x)$ gives us some knowledge of $\theta(t, x)$. Formally, an identification assumption can be thought of a map from the space of the observed parameter to the space of the unobserved one. There are two main types of identification assumptions commonly employed in causal inference: point identification assumptions that link one value of $\phi(t, x)$ to one value of $\theta(t, x)$, and partial identification assumptions, that link one value of $\phi(t, x)$ to several values of $\theta(t, x)$.

The framework introduced above offers a natural way to employ multiple identification assumptions at once by encoding them into a specific form for h: a mixture of different distributions, each one corresponding to one identification assumption, that is we suggest:

$$h(t, x, \phi(t, x), \boldsymbol{\eta}) = \sum_{l=1}^{L} \pi_l h_l(t, x, \phi(t, x), \boldsymbol{\eta}). \tag{1}$$

In the above $\pi_l \in (0, 1)$ for each component l, with $\sum_{l=1}^{L} \pi_l = 1$. These proportions are commonly referred to as mixture weights Bishop (2006) and they represent the proportion of $h(t, x, \phi(t, x), \boldsymbol{\eta})$ that should be drawn from each of the component. A smaller value of π_l implies that a component has a lower overall weight in the final mixture. Section 4 is devoted to interpreting and choosing π_l in the context of mixtures of identification assumptions. To summarize the proposed interpretation we suggest viewing each π_l as a proportion of the data in which each of the identification assumptions holds. This suggest that modeling g as a mixture of priors is tantamount to assuming that the population of interest is composed of multiple subpopulations, where in each one a different identification assumption plausibly holds. We will give ways to express some popular point and partial identification assumptions as mixture component in the next section.

Note further that this strategy is preferable to simply modeling g as a continuous distribution over the space of $\theta(t, x)$. This is for at least two reasons: first, researchers often do not know enough about the full shape of the distribution of $\theta(t, x)$ to suggest one fully, rather identification assumptions often either identify a single point in the parameter space, or define bounds around the counterfactual parameter Manski (2009). Second, the nature and definition of these assumptions is often that they are based on discrete qualitative facts about the process being modeled. For example, conditional ignorability is based around knowledge of what observables influence the treatment selection process: it would be hard and conceptually impractical to ask analysts to define and justify a continuous version of this assumption. [3]

[3] It is unclear what the continuous dimension should be (an infinite space of unobservables?), and even if this dimension was

3.1 Point and Partial Identification Assumptions: Some Examples

In this section we provide some example of priors that can be used in an assumption mixture based on common point and partial identification assumptions. We suggest using point mass distributions to express point identification assumption as priors, and uniform distributions with specific endpoints for partial identification assumptions.

Conditional Ignorability: This is the canonical assumption behind much observational research in quantitative social science. Informally it states that all the potential factors that affect both the treatment and the outcome are measured and observed. Formally, Conditional Ignorability holds if $Y \perp\!\!\!\perp T|X$. Using this fact we can see that: $\mathbb{E}[Y(t)|T = t, X = x] = \mathbb{E}[Y(t)|T \neq t, X = x]$, that is the average observed response is equal to the counterfactual response under this assumption. In our framework this translates to the following form for h_l:

$$h_l(t, x, \phi(t, x), \boldsymbol{\eta}) = \delta_{[\phi(t,x)]},$$

where $\delta_{[a]}$ is the distribution that puts mass 1 at a and 0 everywhere else. This is a point mass prior, which is also commonly used in variable selection applications (Gelman et al., 2013). Combined with, for example, a bound on the parameter it leads to the common spike and slab prior (George and McCulloch, 1993; Ishwaran et al., 2005), where values for the unobserved parameter θ are shrunk towards the observed one ϕ. In general this is a very strong assumption. It requires analysts to believe that all potential confounders of the causal relationship between T and Y are being held constant, which is often a claim that's hard to substantiate in real world scenarios.

Partial Identification: We now concentrate on partial identification assumptions: assumptions that permit us to identify a set of potential values for θ given ϕ. This implies, practically, that they lead to bounds on the value of θ, rather than to an actual point estimate. In a very general sense, consider two fixed and known functions, $L(x, t, \phi(x, t), \boldsymbol{\eta})$ and $U(x, t, \phi(x, t), \boldsymbol{\eta})$, such that we can assume that $L(x, t, \phi(x, t), \boldsymbol{\eta}) \leq \theta(x, t) \leq U(x, t, \phi(x, t), \boldsymbol{\eta})$, conditionally on values of $x, t, \phi(x, t)$ and $\boldsymbol{\eta}$, then one possibility is using a uniform distribution with these two values as bounds as a mixture component for h:

$$h_l(x, t, \phi(x, t), \boldsymbol{\eta}) = Uniform(L(x, t, \phi(x, t), \boldsymbol{\eta}), U(x, t, \phi(x, t), \boldsymbol{\eta})).$$

Uniform priors are often unused in objective bayesian statistics because they are not transformation-invariant (Hoff, 2009), however in our case the density of $\theta(t, x)$ is not normally updated with a likelihood, and is instead simply re-computed with updated values of $\phi(t, x)$ as parameter values. Because of this the problem of transformation invariance does not present itself anymore and an uniform density can be used as a choice for h_l. An alternative to choosing a uniform density would be to truncate another distribution at the bounds. This is the approach taken by Hahn et al. (2016) in proposing a truncated prior for a partially identified

well defined, it would be hard to justify the assumption with qualitative or quantitative additional evidence.

model. In general, we will use $q_{[L,U]}(t, x, \boldsymbol{\eta})$ to denote any distribution that has domain bounded between L and U, that is, under which values outside of $[L, U]$ have probability 0.

Natural Bounds and Objective Priors: If the outcome of interest is known to be bounded a natural choice for one of the components of the mixture is a uniform prior between the bounds, that is, if $L_0 \leq Y \leq U_0$ then a choice for h_l is:

$$h_l(L_0, U_0) = q_{l[L_0,U_0]}(t, x, \boldsymbol{\eta}). \tag{2}$$

In case the outcome is not known to be formally bounded it is possible to choose constants that are known to be far away from the observed data in the parameter space. One possibility is to use a uniform prior with large bounds, another possibility is to use a weakly informative prior as suggested by Gelman et al. (2008). Another option is to rescale the outcome variable of interest so that it takes value within a bounded set. This can be accomplished with any of the known mappings between the entire real line and one of its bounded subsets. In some cases, natural bounds can also be derived from contextual knowledge (Hahn et al., 2016).

It is always recommended to include a component that spans the full space of potential values of θ, this is so that no points in the space are excluded a-priori and without a specific assumption justifying their exclusion. The possibility of always including a component spanning the whole domain of $\theta(t, x)$ in the assumption mixture is one of the main advantages of the approach suggested here as it allows to mix knowledge that comes "from the data alone" (Manski, 1993) with identification assumption. This weakens the required belief in any other assumption while still allowing for the production of informative results.

Throughout the rest of this paper we assume that two such constants can always be found for the outcome of interest, all the assumptions introduced below are, in principle, well specified even in cases in which $L_0 = -\infty$ and $U_0 = \infty$, although they may not be substantially meaningful anymore.

Monotone Treatment Selection: This assumption, originally introduced by Manski and Pepper (2000), posits that, for any two treatment levels t, s such that $t \geq s$, we have

$$\mathbb{E}[Y(t)|T = t, X = x] \geq \mathbb{E}[Y(t)|T = s, X = x].$$

Note that this assumption requires the treatment levels to have some kind of meaningful ordering, so that the relationship $t \geq s$ is well-defined. Informally, this assumption states that units that self-select into some greater level of treatment will also exhibit a higher value of the potential response under that treatment than units who did not select into that treatment. Examples of this are income as an outcome and years of schooling as a treatment variable: we can plausibly assume that individuals who choose to undergo more education will earn more than individuals who do not. The key here is that we do not make any assumptions as to *why*: it could either be because education is having a causal effect on earnings or because some unobserved (propensity to work hard, pre-education skills, passion for the profession,...) is actually causing the individual to both select into more education and to earn more. Let $P(t, x) = \Pr(T < t|X = x)$, using

this assumption $\theta(t,x)$ can be bounded in terms of identifiable quantities as follows:

$$\begin{aligned}\theta(t,x) &= \mathbb{E}[Y(t)|T \neq t] \\ &= \mathbb{E}[Y(t)|T > t, X = x] \Pr(T > t|X = x) + \mathbb{E}[Y(t)|T < t, X = x] \Pr(T < t|X = x) \\ &\leq U_0(1 - P(t,x) - e(t,x)) + \phi(t,x)P(t,x) = U_{MTS}(t,x)\end{aligned}$$

and,

$$\theta(t,x) \geq L_0 P(t,x) + \phi(t,x)(1 - P(t,x) - e(t,x)) = L_{MTS}(t,x).$$

We can use this bound to construct a uniform prior on the counterfactual:

$$h_l(x,t,\phi(x,t),\boldsymbol{\eta}) = q_{l[L_{MTS}(t,x),U_{MTS}(t,x)]}(t,x,\boldsymbol{\eta}).$$

Monotone Treatment Response: Manski and Pepper (2000) go one step further from MTS and also study a scenario in which we can make the following assumption, for any two treatment levels t, s such that $t \geq s$:

$$Y(t) \geq Y(s).$$

This assumption states that the direction of the treatment response is known: we can assume that Y varies monotonically with the treatment a unit is administered; again, this presupposes the treatment to take value in some ordered set. Going back to the returns to schooling example, it is plausible to assume that more years of schooling should lead to weakly greater income compared to fewer years for the same individual. Note that this assumption is commonly made with respect to treatment take up in instrumental variable estimation of LATEs (Imbens and Rubin, 2015). With this assumption Manski and Pepper (2000) derive the following bounds on the counterfactual:

$$\begin{aligned}\theta(t,x) &\leq \sum_{s>t} \phi(s,x)e(s,x) + U_0 P(t,x) = U_{MTR}(t,x) \\ \theta(t,x) &\geq \sum_{s<t} \phi(s,x)e(s,x) + L_0(1 - P(t,x) - e(t,x)) = L_{MTR}(t,x).\end{aligned}$$

Note that the form of these bounds follows from results in Manski and Pepper (2009). Again, these bounds can be implemented with the mixture component choice:

$$h_l(x,t,\phi(t,x),\boldsymbol{\eta}) = q_{l[L_{MTR}(t,x),U_{MTR}(t,x)]}(t,x,\boldsymbol{\eta})$$

. It is important to keep in mind that MTR excludes the possibility of a treatment effect being on either side of the zero line: once it is assumed, the bound on the TE obtained with it is either an upper or lower bound, with zero being the other bound (Manski and Pepper, 2000).

If MTS and MTR are assumed to hold at the same time in the same population, then they can be com-

bined into the following tighter bound:

$$\theta(t,x) \leq \sum_{s>t} \phi(s,x)e(s,x) + \phi(t,x)P(s,x) = U_{MTS-MTR}(t,x)$$
$$\theta(t,x) \geq \sum_{s<t} \phi(s,x)e(s,x) + \phi(t,x)(1 - P(s,x) - e(s,x)) = L_{MTS-MTR}(t,x).$$

This result is given in Proposition 2.2 of Manski and Pepper (2009). Note that assuming MTS-MTR at the same time is different from including two components in the prior mixture, one for MTS and one for MTR. This is, conceptually, because in the first case the two assumptions hold together in the same population, while in the second case we are assuming that, if in a portion of the population MTS holds then it is certainly the case that MTR does not and vice versa. For this reason it makes sense to combine components with MTS, MTR only and MTS-MTR in the same prior mixture.

Monotone Instrumental Variable: A more general version of the MTS assumption can be introduced by assuming that the response is weakly increasing conditionally on a particular covariate, the *Monotone Instrument*. Let $Z \in \{1, \ldots, M_Z\}$ represent the monotone instrument under consideration. The MIV assumption states that, for two values of Z, z, v such that $z > v$:

$$\mathbb{E}[Y(t)|Z = z, X = x] \geq \mathbb{E}[Y(t)|Z = v, X = x].$$

Informally, this assumption states that outcome means are increasing in the monotone instrument. For a practical example of this type of instrument, consider the effect of a training program on work performance as treatment and outcome respectively, and let education of the employee be the monotone instrument. It is plausible that, an employee with a greater level of education would perform better than one with a lower education level after undergoing the same training program. To introduce bounds under this assumption, we define the following notation: $\phi(t,x,z) = \mathbb{E}[Y(t)|T = t, X = x, Z = z], e(t,x,z) = \Pr(T = t|X = x, Z = z), P(t,x,z) = \Pr(T < t|X = x, Z = z), e_Z(z,x) = \Pr(Z = z|X = x)$. Using this assumption and the notation just introduced, bounds usable in the proposed prior mixture are as follows:

$$\begin{aligned}
\theta(t,x) &\leq \sum_{z=1}^{M_Z} \left[\min_{v>z} \phi(t,x,v)e(t,x,v) + U_0(1 - e(t,x,v)) - \phi(t,x,z)e(t,x,z)\right] \frac{e_Z(z,x)}{1 - e(t,x)} \\
&= U_{MIV}(t,x,Z) \\
\theta(t,x) &\geq \sum_{z=1}^{M_Z} \left[\max_{v<z} \phi(t,x,v)e(t,x,v) + L_0(1 - e(t,x,v)) - \phi(t,x,z)e(t,x,z)\right] \frac{e_Z(z,x)}{1 - e(t,x)} \\
&= L_{MIV}(t,x,Z).
\end{aligned}$$

The forms of these bounds follows directly from proposition 2.1 of Manski and Pepper (2000). To obtain tighter bounds, MIV can be combined with MTR, which is equivalent to assuming that the two propositions

hold at once on the same observation. Under MIV-MTR the bounds become:

$$\begin{aligned}
\theta(t,x) &\leq \sum_{z=1}^{M_Z}\Bigg[\min_{v>z}\left(\sum_{s\geq t}\phi(s,x,v)e(s,x,v)\right)(P(t,x,v)+e(t,x,v))+U_0(1-P(t,x,v)-e(t,x,v)) \\
&\qquad -\phi(t,x,z)e(t,x,z)\Bigg]\frac{e_Z(z,x)}{1-e(t,x)} \\
&= U_{MIV-MTR}(t,x,Z) \\
\theta(t,x) &\geq \sum_{z=1}^{M_Z}\Bigg[\max_{v<z}\left(\sum_{s\leq t}\phi(s,x,v)e(s,x,v)\right)(1-P(t,x,v))+L_0P(t,x,v) \\
&\qquad -\phi(t,x,z)e(t,x,z)\Bigg]\frac{e_Z(z,x)}{1-e(t,x)} \\
&= L_{MIV-MTR}(t,x,Z).
\end{aligned}$$

This follows from the same proposition of Manski and Pepper (2000). These are only some of the possible assumptions that could be used to partially identify $\theta(t,x)$, several other common assumptions, such as natural experiments or parametric models, can also be expressed as choices of components for the prior mixture h_l.

3.2 General Inference via MCMC

Under some of the choices of prior introduced above, we have that the posterior distribution of $\theta(t,x)$ depends on $\phi(t,x)$ as the latter is used to bound the distribution of the former. This can make analytically obtaining forms for the distribution of $\theta(t,x)$ complicated, however MCMC inference is straightforward for all the choices of h considered in this paper. This is because of the observation that $\Pr(\theta(t,x),\phi(t,x)|\mathbf{Y}) = \Pr(\phi(t,x)|\mathbf{Y})\Pr(\theta(t,x)|\phi(t,x))$. This follows from the fact that $\theta(t,x)$ does not depend on the data (Kadane, 1975). Thanks to this, Bayesian inference on $\Pr(\mu(t,x)|\mathcal{D})$ can be performed easily and generally, and this enables easy inference for $\Pr(\tau(t,s,x)|\mathcal{D})$. The following is a valid general MCMC sampling scheme for $\mu(t,x)$ and $\tau(t,s,x)$:

1. Obtain a vector of S samples from $g_{\mathcal{D}}(t,x,\boldsymbol{\eta})$
2. For each sample $\phi^{(s)}(t,x)$, $s=1,\ldots,S$:
   (a) With probability π_l sample $\theta^{(s)}(t,x)$ from $h_l(t,x,\phi^{(s)}(t,x),\boldsymbol{\eta})$.
   (b) Compute $\mu^{(s)}(t,x) = \phi(t,x)^{(s)}e(t,x) + \theta^{(s)}(t,x)(1-e(t,x))$
3. Repeat from step 1 for treatment value s to generate a vector of S samples of $\mu(s,x)$.
4. For each sample $s=1,\ldots,S$ compute $\tau(t,s,x)^{(s)} = \mu^{(s)}(t,x) - \mu^{(s)}(s,x)$.

This is a very general scheme that can be adapted to the needs of the specific choices of g and h made in every instance. In general, this method should highlight the fact that it is easy both mathematically and

computationally to obtain values of the partially identified posterior under the proposed assumption mixture. Common functionals of $\Pr(\tau(t,s,x)|\mathcal{D})$ that are of interest can be computed easily from the samples with any of the standard MCMC integration tools.

3.3 Interpreting the Mixture Weights

Interpreting the weight π_l assigned to each assumption is a complex issue. In this section we outline two possible interpretation for these weights that can guide analysts in thinking about credible assumptions. Before introducing these two interpretations it is important to clarify one thing that they have in common: each assumption is considered to hold separately from any other in the mixture, for example, if two assumptions are added together in a mixture this is equivalent to believing that one assumption holds with a certain probability but not the other, and vice-versa. If we wish to encode the idea that two assumptions should hold at once in our model we need to do so explicitly as another mixture component. Mixture components are mutually exclusive in what they imply[4], and this is common to both possible interpretation of the weights.

The first possible interpretation for the mixture weights is that they represent a level of belief in each assumption held by the analyst. Saying that an assumption (a mixture component) has a weight greater than another one is equivalent to saying that we believe that one assumption (and that assumption only) is more likely to hold in the population of interest than the other. This interpretation is useful when trying to convey the strength of the results produced with our method: if the weights on the strongest assumptions in a mixture are small then we can conclude that the results would hold even at a low level of belief in those assumptions, i.e. the assumptions are likely not what's driving the result. This is analogous to the use of common sensitivity analysis (Imbens and Rubin, 2015).

The second possible interpretation of the assumption mixture is as a composition of populations, in each of which only one of the assumptions exclusively holds. In this view the mixture weights are simply the proportion of each population that make up the superpopulation from which the observational data is drawn. This interpretation is useful because it allows analysts to think of their data as a mix of different groups, and evidence for the fact that each assumption holds can to be gathered separately for each group. For example, if the data comes from different obvservational contexts, such as geographic locations or events, the mixture weights can be the proportions of the data that come from each context. This is useful because each assumption now has to be defended only within the subgroup it is assumed to hold on, which potentially makes gathering evidence in support of each assumption easier.

4 Optimally Credible Weights

How should the mixture weights $\pi_1, \ldots, \pi_L$ be specified? Clearly, the data cannot help us in determining how much weight should be given to each assumption, since these assumptions represent a level of belief they are a purely theoretical construct. Analysts have to specify a value of $\pi_1, \ldots, \pi_l$ a priori. Here we propose an approach based on a sufficiency criterion: what is the most credible set of weights that produces

[4]Note that this does not preclude that their content, such as the portions of the counterfactual domain they place positive probability on, could overlap.

a $1-\alpha$-credible interval that excludes 0? Given a ranking for each of the assumptions in the mixture in terms of their strength, our approach returns the set of weights that puts the most strength on the weakest assumption, subject to the resulting credible interval for the ATE being of the desired size and direction.

Optimally credible weights are of interest to potential analysts because, they represent a boundary condition on the importance of each assumption, that is, they let analysts know the minimum amount of belief that they need to hold in each assumption in order to produce conclusive results at a certain confidence level. Once the weights have been found by the proposed algorithm, then analysts can decide whether they are willing to assign similar levels of belief to each assumption, or if the requirements on each assumption's importance are too stringent. In order to compute these weights, we propose a simple quadratic programming approach, which only requires knowledge of the size of the credible intervals for each assumption. As shown in Section 3.2, we can sample from the posterior distribution of τ separately for each component of our assumption mixture, therefore a credible interval at the desired level can always be computed for each component separately. Effectively, this represents a form of sensitivity analysis as it quantifies the reliance of a desired result on the strongest assumptions made.

Consider the problem of having to obtain optimal weights for the CATE at two two treatment levels t and s. Let the assumption mixture for $\theta(t,x)$ have L components $h_{t1},\dots,h_{tL}$ and let the assumption mixture for $\theta(s,x)$ have R components $h_{s1},\dots,h_{sR}$. Fix a desired credible interval a,b, noting that one of the two terms could a the natural bound on the domain of $\tau(t,s,x)$. To compute a 1-α credible interval we need the probability that the CATE of interest is between two values, this is given by the integral of $f_{\tau(t,s,x)|\mathcal{D}}$: between the two values in question.[5]

This probability can be computed easily with samples obtained from $f_{\tau(t,s,x)|\mathcal{D}}$ with the method described in the previous section. Define the following probability for each r,s pair:

$$w_{lr} = \int_a^b \int_u \int_v g_{\mathcal{D}}\left(\frac{v}{e(t,x)}; x,t,\boldsymbol{\eta}\right) h_{tl}\left(\frac{\tau_{t,s}+u-v}{1-e(t,x)}; x,t,\phi(t,x),\boldsymbol{\eta}\right) dv \\ \times \int_v g_{\mathcal{D}}\left(\frac{v}{e(s,x)}; x,s,\boldsymbol{\eta}\right) h_{sr}\left(\frac{u-v}{1-e(s,x)}; x,s,\phi(s,x),\boldsymbol{\eta}\right) dv.$$

In order to compute the optimal weights for each l,r, w_{lr} needs to be computed via either explicit computation or sampling from the posterior of that case. Once these intervals are computed, we can define the following vectors and matrices:

$$\boldsymbol{\pi}_t = \begin{bmatrix} \pi_{t1} \\ \vdots \\ \pi_{tL} \end{bmatrix}, \quad \boldsymbol{\pi}_s = \begin{bmatrix} \pi_{s1} \\ \vdots \\ \pi_{sR} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} w_{11} & \dots & w_{1R} \\ \vdots & \ddots & \vdots \\ w_{L1} & \dots & w_{LR} \end{bmatrix}.$$

The probability interval on interest can be expressed as: $\Pr(a \le \tau(t,s) \le b|\mathcal{D}) = \boldsymbol{\pi}_t' \mathbf{W} \boldsymbol{\pi}_s$: this will help us in formulating a quadratic program to obtain optimally credible weights.

In order to find weights that minimize reliance on the strongest assumptions we must have some known strength values for each of the assumptions: $A_{t1},\dots,A_{tL}$ and $A_{s1},\dots,A_{sR}$. These values represent the

[5]The full definition of this quantity is given in Equation (11)in the supplementary materials.

level of strength of each assumption according to the analyst. The only requirement that these values must satisfy is that they have produce a valid ranking, potentially with ties, between the various component of the assumption mixture for each parameter. The actual values assigned to each of the A_l will not matter in the solution to the problem of finding the optimal weights. Because of this, simply assigning an integer representing the position of the assumption in the desired ranking is a valid way of specifying A_l.[6]

Once all these quantities are defined, optimally credible weights can be found by solving the following quadratic program:

$$\boldsymbol{\pi}_t^*, \boldsymbol{\pi}_s^* = \underset{\boldsymbol{\pi}_t, \boldsymbol{\pi}_s}{\arg\min} \sum_{l=1}^{L} \pi_{tl} A_{tl} + \sum_{r=1}^{R} \pi_{sr} A_{sr} \tag{3}$$

Subject to:

$$\boldsymbol{\pi}_t' \mathbf{W} \boldsymbol{\pi}_s = 1 - \alpha \tag{4}$$

$$0 \leq \pi_{tl} \leq 1 \qquad \text{for } l = 1, \ldots, L \tag{5}$$

$$0 \leq \pi_{sr} \leq 1 \qquad \text{for } r = 1, \ldots, R \tag{6}$$

$$\sum_{l=1}^{L} \pi_{tl} = 1 \tag{7}$$

$$\sum_{r=1}^{R} \pi_{sl} = 1. \tag{8}$$

Starting with Equation (3), the loss for the optimization problem is simply a weighted combination of the strength of each assumption. The constraints in Equations (5), (6), (7), and (8) define the weights as valid probabilities. The Constraint in (4) defines the desired $1 - \alpha$-credible interval. Note that the problem of obtaining optimally credible weights for the ATE is exacly the same except that the credible intervals w_{lr} must be computed by also averaging over observed values of X.

This problem can be solved with any of the common quadratic programming solution software. Note also that, in most practical applications, the number of components in each assumption mixture will be relatively small and therefore the solution to this program can be computed in trivial time.

5 Simulation Results

In this section we present results from applying our methodology to several simulated datasets. We show how a mixture of assumptions is able to produce 95% credible intervals that both include the true treatment effect and are fully on one side of the 0 line. This permits researchers to draw conclusions on the direction of the treatment effect that are robust to reliance on strong identification assumptions. We apply our methodology to unconventional settings that are not handled by common methodologies, and show that our

[6] Another valid way would be to use the size of the domain of each assumption as a measure of its strength, with stronger assumption having a smaller domain. This would be feasible in the case of bounds on the parameters of interest being used as assumptions.

approach can produce informative results in these settings as well.

5.1 Binary Outcome

First we present a simple simulation implemented with a binary outcome $Y(t) \in \{0, 1\}$ and confounded treatment assignment. We simulate data from the following model:

$$
\begin{aligned}
Y_i(1)|T_i = 1 &\sim Bernoulli(\phi(1)) \\
Y_i(0)|T_i = 0 &\sim Bernoulli(\phi(0)) \\
Y_i(1)|T_i = 0 &\sim Bernoulli(\theta(1)) \\
Y_i(0)|T_i = 1 &\sim Bernoulli(\theta(0)),
\end{aligned}
$$

with $\phi(1) \neq \theta(1)$ and $\phi(0) \neq \theta(0)$: this implies ignorability does not hold in this model.[7] We use a CAM model with two assumptions: ignorability, and a uniform bound between 0 and 1, the natural domain of the outcome.

$$
\begin{aligned}
\mathbf{Y}(t) &\sim Bin(50, \phi(t)) \\
\phi(t) &\sim Beta(2, 2) \\
\theta(t) &\sim \pi\delta_{[\phi(t)]} + (1 - \pi)uniform(0, 1),
\end{aligned}
$$

for $t = 0, 1$. The whole model is estimated by sampling from the relative posterior distributions.

Results for the simulation are presented in Figure 1. The naive estimate in the figure is constructed by taking the posterior mean under ignorability (the red dots in the figure), and it results in an ATE estimate of about 0.7 with a credible interval ranging between 0.55 and 0.7. Not only is the ATE overestimated, but more importantly its true value is not contained in the credible interval. As shown by the figure, relaxing ignorability just slightly, i.e. going from 100% belief in it to 96% generates a credible interval that is wide enough to include the true value of the ATE. The figure also shows that a value of $\pi = .48$ or 48% for the weight on ignorability is the largest one for which the credible interval does not cross the 0 line: we have evidence of the treatment effect being positive even when we believe in CI only weakly, as a value of π even slightly larger than .48 ensures a credible interval that does not cross the 0 line.

[7] Parameter values for this simulation are given in Table 2 in the supplementary materials

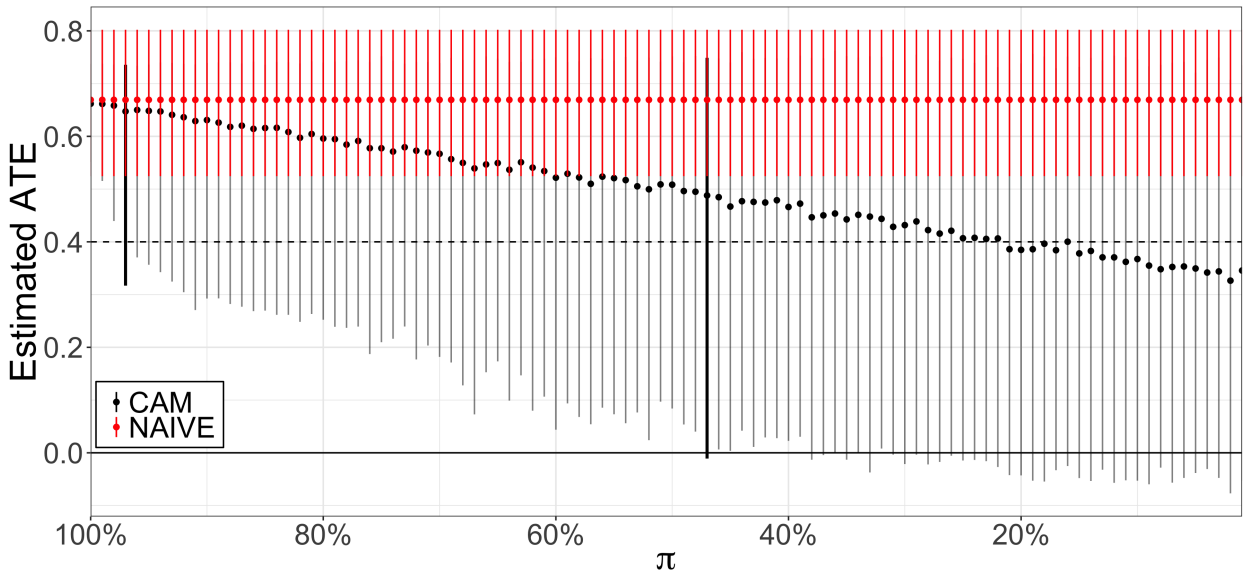

Figure 1: Results for a simulation with a binary outcome. The weight assigned to ignorability by the CAM model (π) is on the x-axis. Estimated treatment effects are on the y axis together with 95% credible intervals around them. The red band represents the naive posterior mean and 95% CI that assume full ignorability. The dashed line is the true treatment effect. The solid black credible intervals represent the beginning and end of the region in which the true treatment effect is included in the CI and the zero line is not.

5.2 Contaminated data

In this simulation we study a setting in which outcomes are generated from a set of ten observed and one unobserved covariates. The data generation process is:

$$
\begin{aligned}
Y_i &= \alpha + X_i\beta + T_i\tau + T_iU_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,3) \\
U_i &= \begin{cases} U_i^* & \text{if } D_i = 1 \\ 0 & \text{if } D_i = 0 \end{cases} \\
D_i &\sim Bernoulli(0.5).
\end{aligned}
$$

We only observe Y, X, T. We do, however, know that 50% of the data is contaminated by the unobservable quantity, while a different proportion is not. Note that we cannot restrict our analysis to only the uncontaminated portion because we do not know which units are contaminated, as we do not observe D_i. We use a mixture of conditional ignorability and a natural bound on θ_i to model this case:

$$
\begin{aligned}
Y_i|\phi(T_i, X_i) &\sim \mathcal{N}(\phi(T_i, X_i), \sigma^2) \\
\phi(T_i, X_i) &= X_i\beta + T_i\tau \\
\beta, \tau &\sim \mathcal{N}(0, \mathbf{I}) \\
\theta_i(t) &\sim \pi\delta_{[\phi(t,X_i)]} + (1-\pi)Uniform(-100, 100).
\end{aligned}
$$

This model handles unobservables with a natural bound on the counterfactual mixed with a point mass for those observation that are not contaminated. The values $(-100, 100)$ are chosen as bounds for θ because Y ranges roughly between -2 and 2: these values are large enough that can be reasonably chosen to bound it. [8] We compare the CAM model above to a naive Bayesian regression estimate produced by simply regressing X and T on the observed Y with normal priors on both the treatment and covariate coefficients.

Results are reported in Figure 2. We see that naive regression estimates overestimate the treatment effect by about 50%, with 95% credible intervals not covering the true ATE. This is expected, as the model is in violation of the canonical assumptions underpinning Bayesian regression. CAM generates 95% credible intervals that cover the true value of the treatment when conditional ignorability (or no omitted confounders) has a weight below 85% in the mixture. The CAM point estimate is closest to the true treatment effect when $\pi \approx .5$, which is explained by the true contaminated portion of the data being 50%.

5.3 Regression with noisy bounded unobservables

In this simulation we study a regression model with an unobservable confounder. The outcome takes form:

$$
\begin{aligned}
Y_i &= \alpha + X_i\beta + T_i\tau + T_iU_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,3) \\
U_i &\sim Uniform(a_i, b_i).
\end{aligned}
$$

[8] Parameter values for the simulation are given in Table 3in the supplementary materials

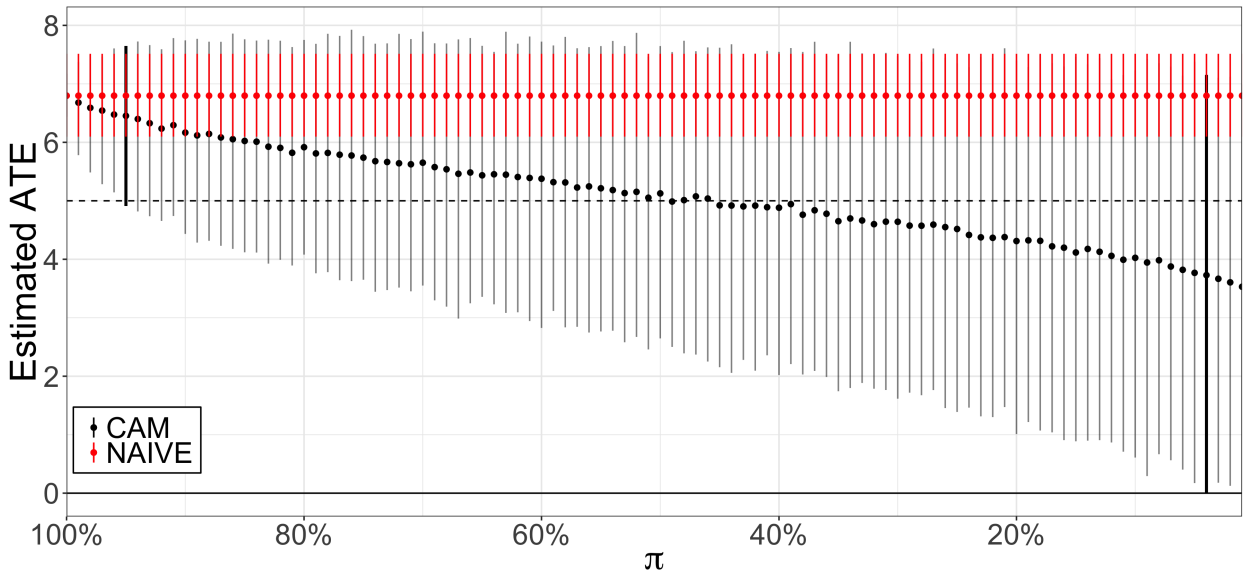

Figure 2: Results for a simulation with a regression with bounded unobservables. The importance of conditional ignorability in the CAM model (π) is on the x axis. Estimated treatment effects are on the y axis together with 95% credible intervals around them. The red band represents the naive estimate and 95% credible interval that was obtained with a Bayesian linear regression of X and T on Y. The dashed line is the true treatment effect. The solid credible intervals represent the beginning and end of the region in which the true treatment effect is included in the CI and the zero line is not.

We observe Y, T and X, but we do not observe U. We instead have bounds on it for each unit, represented by the quantities a_i and b_i. We also do not observe these bounds for all units, but only a version of them that is measured with error, that is, for each unit we observe:

$$A_i = \begin{cases} a_i & \text{if } D_i = 1 \\ \xi_i^a & \text{if } D_i = 0 \end{cases}, \text{ and } B_i = \begin{cases} b_i & \text{if } D_i = 1 \\ \xi_i^b & \text{if } D_i = 0 \end{cases},$$

where $D_i \sim Bernoulli(0.5)$ is a random variable independent of all other covariates for each unit. In sum, we have a regression with a bounded unobservable that has noisy bounds. Most existing statistical tools commonly used in the social sciences are not equipped to deal with this type of confounding.

We use the following CAM model for this simulation:

$$\begin{aligned} Y_i|\phi(T_i, X_i) &\sim \mathcal{N}(\phi(T_i, X_i), \sigma^2) \\ \phi(T_i, X_i) &= X_i\beta + T_i\tau \\ \beta, \tau &\sim \mathcal{N}(0, \mathbf{I}) \\ \theta_i(1, X_i) &\sim \pi Uniform(\phi(1, X_i) - B_i^*, \phi(1, X_i) - A_i^*) + (1-\pi)Uniform(-100, 100) \\ \theta_i(0, X_i) &\sim \pi Uniform(\phi(0, X_i) + A_i^*, \phi(0, X_i) + B_i^*) + (1-\pi)Uniform(-100, 100). \end{aligned}$$

We bound the counterfactual outcome for each observation with a mixture of a natural bound and the observed bounds for the unobservable value added and subtracted to the observed outcome. The natural bound is needed in the mixture to account for those cases in which observation of A_i and B_i is incorrect. The naive estimate in this simulation is the coefficient on T on a Bayesian linear regression of X, A, B, and T on Y.[9]

Results for this simulation are presented in Figure 3. Naive regression is clearly overestimating the treatment effect, and the 95% credible interval around its estimate includes the zero line. For the CAM model, we obtain a credible interval that covers the true value of the treatment by putting full weight on the observed bounds on θ, while credible intervals cross the zero line when weight on the natural bound goes below 17%. This simulation shows that, in theory, the bounding strategy for θ that uses one component could work without the need without the need for a mixture. In practice, however, mixing this bound with the natural, more conservative bound on θ is needed, as we cannot know if the true value of the ATE lies within the narrower observed bounds. Results

As an additional test, we simulated 50 different datasets from the generative processes of Simulation 2 and Simulation 3, and compared absolute estimation error and 95% coverage for three methods: CAM with the models specified for each of the two simulations, the naive Bayesian linear regression used in each simulation, and a nonparametric frequentist regression model: Kernel Regularized Least Squares (Hainmueller and Hazlett, 2014). For CAM we use a value of $\pi = 0.5$: this comes from the contextual knowledge we have in each simulation that half of the data is not generated according to the naive models. Results for this additional test can be found in Table 1. We see that CAM performs better than the two other methods it is compared to both in terms of absolute estimation error and coverage. This is because CAM makes use of

[9]Parameter values for this simulation are presented in Table 4in the supplementary materials.

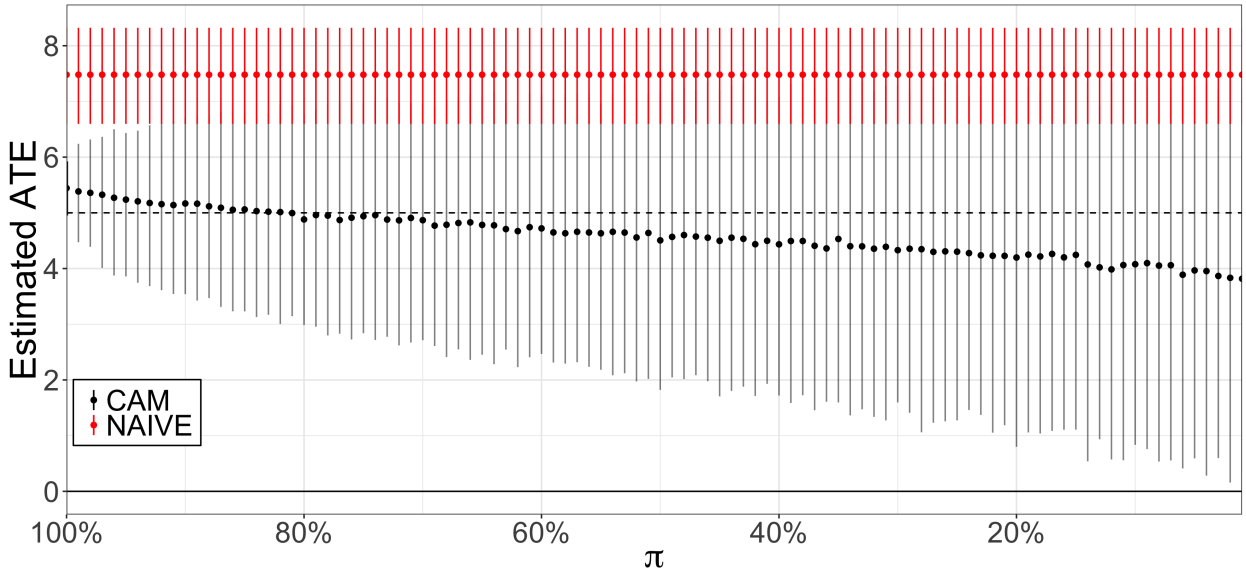

Figure 3: Results for a simulation with a regression with bounded unobservables. The weight of the observed bounds (π) in the CAM mixture is on the x axis. Estimated treatment effects are on the y axis together with 95% credible intervals around them. The red band represents the naive regression estimate and 95% CI. The dashed line is the true treatment effect. The solid credible intervals represent the beginning and end of the region in which the true treatment effect is included in the CI and the zero line is not.

additional information available for each simulation, such as the proportion of the data that is contaminated, and the observed bounds.

	Simulation 2			Simulation 3		
Method	Error mean	Error sd	Coverage	Error mean	Error sd	Coverage
CAM	0.61	0.23	1.00	0.36	0.31	1.00
NAIVE	2.46	0.30	0.00	2.52	0.44	0.00
KRLS	1.01	0.30	0.18	1.05	0.41	0.28

Table 1: Results from 50 simulated datasets with the settings and models in simulations 2 and 3. Coverage for KRLS is given by a 95% confidence interval. Coverage for CAM and naive is given by a 95% credible interval.

These simulations highlight the usefulness of CAM in two ways: first, they show how CAM can be used to incorporated known information about unobservables into a common regression framework, by explicitly putting a prior on the unobserved counterfactual. Second, it presents a scenario in which the stronger assumption holds relatively strongly and demonstrates the ability of the method to capture useful and correct results in this case as well.

6 Application: Do Incumbent Reward Electoral Supporters?

We apply our proposed methodology to a canonical question in political science: do incumbent politicians reward electoral districts that voted for them by selectively allocating public goods towards them? We test this proposition on data from elected officials in Malawi's National Assembly originally made available by Ejdemyr et al. (2018b). The cited paper provides evidence of the fact that incumbents selectively allocate more public goods to districts that are more ethnically segregated, as this allows them to more easily claim credit among the targeted groups. We exploit this result to construct a MIV bound on the effect of electoral margin of victory on allocation of public goods. We mix the bound with full ignorability conditional on a set of control covariates, and show that results in support of the hypothesis have to rely almost in full on the latter assumption to attain conventional levels of statistical significance.

Elected MPs in Malawi have both the incentives and the means to reward supporters with targeted public goods: on the incentives side, Malawi is a country in which parties are largely similar in terms of ideology and policies, and it is in contexts such as these that incumbents have the largest incentives to favor supporting constituencies (Cox and McCubbins, 1986; Dixit and Londregan, 1996). On the means side, elected MPs in Malawi's National Assembly have plenty of opportunities to sway the public good allocation process: MPs have been largely in charge of public good allocation until 2005, when district assemblies were first elected, and put in charge of the system (Chasukwa et al., 2014). After this election, MPs still retained ample influence district officials' allocation decisions, both through formal, and informal channels (O'Neil et al., 2014). For both these reasons, we have the expectation that an increase in margin of victory for an elected MP in a district will lead to an increase in public goods allocated to that district, as a reward.

Our dependent variable is the difference in proportion of boreholes in district i in 1998, and the proportion of boreholes in district i in 2008. We focus on the relative proportion of boreholes, and not on the raw

count, because we are only interested in which districts get more boreholes relative to others. This captures allocation effects more directly than a raw count The treatment variable is the average incumbent's margin of victory in district i in from 1998 to 2008: districts with higher average margins, should see greater allocation of boreholes. The treatment variable, originally specified as a percentage, is divided into five quintiles, each corresponding to a treatment level. Our instrument, district ethnic segregation is measured by Ejdemyr et al. (2018a) with the Spatial Dissimilarity Index (Reardon and O'Sullivan, 2004), which we also divide into 5 quintiles, each representing a level.

Our counterfactual modeling strategy will make use of three different components for the prior mixture we will use. As we will see, these are clearly linked to empirical and theoretical statements about the question and setting of our application. First, we exploit the result in Ejdemyr et al. (2018a) to construct a MIV bound on the response function: $\mathbb{E}[Y_i(t)|X = x_i]$, where Y is the proportion of boreholes allocated to district i, t is the margin f victory of the incumbent candidate in that district, and X is a set of control covariates including measures of demand for boreholes in district i. Since Ejdemyr et al. show that more ethnically segregated districts are likely to be targeted with public goods by incumbent politicians, we can use s_i, the level of ethnic segregation in district i to make the MIV assumption: for $s \geq s'$: $\mathbb{E}[Y(t)|X = x_i, S = s] \geq \mathbb{E}[Y(t)|X = x_i, S = s']$, i.e., a more segregated district will, on average, be allocated more boreholes than a less segregated one, conditional on Ejdemyr et al.'s control covariates.

We also consider the possibility that, conditional on the control covariates in X, the incumbent's margin of victory in district i does not depend on the proportion of boreholes allocated to that district, i.e., conditional ignorability holds and we can assume $\mathbb{E}[Y(t)|T = t, X = x] \neq \mathbb{E}[Y(t)|T \neq t, X = x]$. There are two main confounders we control for. First, the actual demand for boreholes in each district: representatives could be allocating boreholes simply in response to this demand and not because of intent to reward supporters. This confounder is captured in the data employed with the inclusion of district measures of pre-treatment amount of boreholes, water and general aid projects, distance to the nearest urban center, and population measures. The second type of confounder we account for is the ethnicity of the representative in each district, as Ejdemyr et al. provide evidence that elected representatives tend to target more public goods towards districts that share their ethnicity. We address this problem by conditioning on whether or not the main ethnic group in each district matches the ethnicity of the elected representative. [10]

Finally, we use the fact that our dependent variable is a difference in proportions to bound it naturally between -1 and 1. Together, these three statements will be used to create a mixture of priors on the unobserved counterfactual of this scenario.

We choose to employ a mixture of a point-mass and uniform distributions as a prior on $\mathbb{E}[Y(t)|T \neq t, X = x]$, that is:

$$h = \pi_1 \delta_{[\phi(t,x)]} + \pi_2 Uniform(L_{MIV}(t,x), U_{MIV}(t,x)) + (1 - \pi_1 - \pi_2) Uniform(-1, 1).$$

We model the observed factual outcome parametrically with a Bayesian linear regression: $\phi(t,x) \sim \mathcal{N}(\alpha + x\beta + t\tau, \sigma^2)$, with standard normal priors on coefficients for treatment and control variables.

[10] More information on the control variables is available in the supplementary materials.

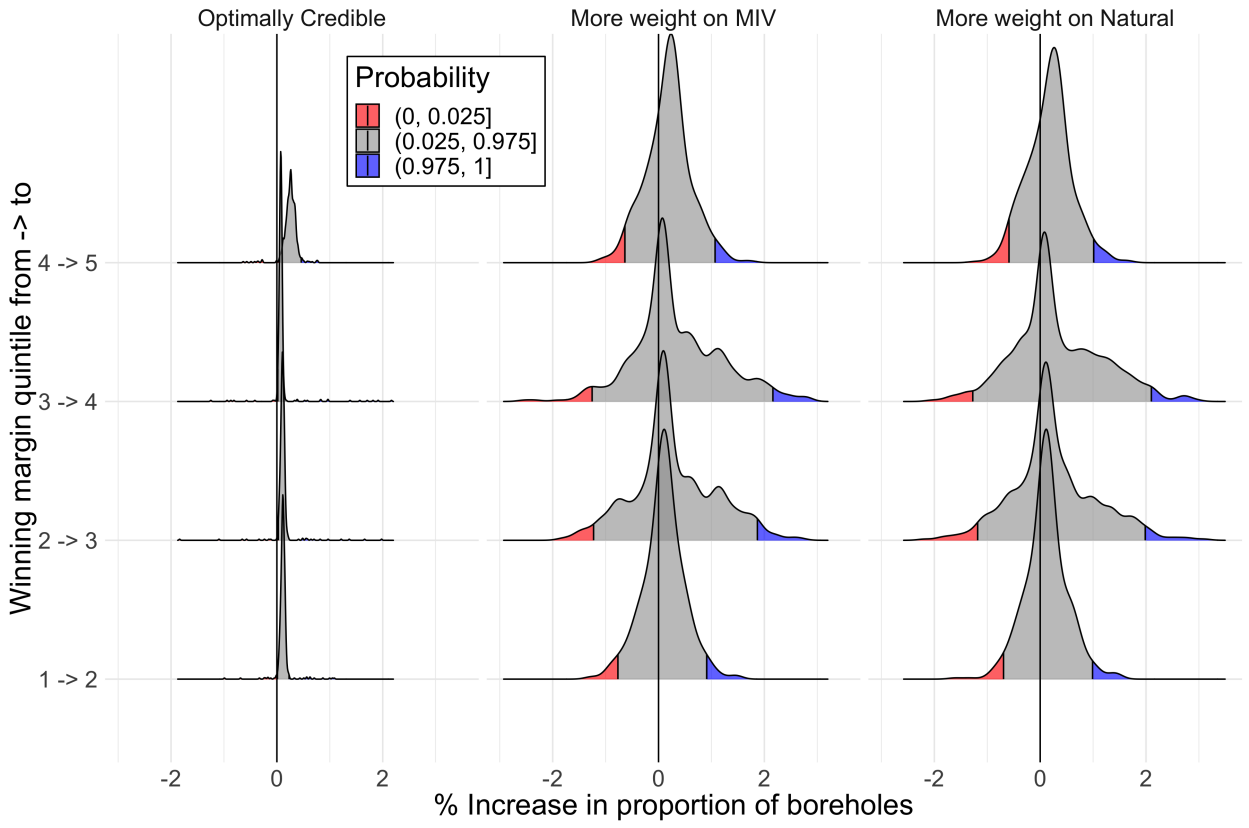

Figure 4: Posterior distribution of treatment effect of moving from one quintile of margin of victory to the next (y-axis) on change in proportion of boreholes in the district (x-axis), computed at three different values for the CAM weights. The optimally credible weights in the leftmost panel are $\pi_1 = 0.95, \pi_2 = 0, \pi_3 = 0.05$, and were obtained with the method described in Section 4. The weights in the second panel are $\pi_1 = 0.1, 0.7, 0.2$, and the weights in the third panel are $\pi_1 = 0.1, 0.1, 0.7$. When almost all weight is put on the conditional ignorability component of the mixture (leftmost panel), we obtain 95% credible intervals that are on the positive side of the zero line. This component consists of a point mass at $\phi(t, x)$, and therefore posteriors are extremely concentrated around it when a large weight is put on it. Relaxing conditional ignorability either by putting more weight on the MIV bound (center) or on the natural bound (right) leads to posteriors that are centered on the positive side of the zero line, but have substantial mass on either side.

Figure 4 shows the posterior distribution of the treatment effect of moving from one quintile of margin of victory to the next. On the leftmost panel, we have results when optimally credible weights are used: in order to obtain a credible interval that is fully on the positive side of the zero line, about 95% of the data must come from either the mixture component defined by conditional ignorability, bound, while the remaining 5% of the population can violate this assumption. On the other two panels, results are shown where more weight is put either on the MIV bound (center panel) or on the natural bound (right panel). Credible intervals in both these settings are almost symmetric around the zero line, suggesting that the posterior distributions induced on the ATE by these components are very similar to each other. In sum, these results show that concluding that these data supports the hypothesis that politicians reward electoral supporters would rely almost entirely on the assumption of conditional ignorability.

This application illustrates the usefulness of our proposed methodology in several ways: first, Bayesian causal inference is made possible with a set of easily interpretable priors, all linked to the data by clear statements that can be substantiated both theoretically and empirically. Second, it shows how similar results can be obtained with different combinations of assumptions: researchers can exploit this fact to guide their evidence collection in favor of one or the other assumption. Finally, it exemplifies how our methodology can be used to report several sets of results, instead of a single one, thus inducing more transparency in how results are reported.

7 Conclusion

In this paper we have set out to deal with three issues common in causal inference: first, the overreliance on a single identification assumption, as opposed to multiple possible ones. Second, the fact that the level of importance that the chosen identification assumption needs to have in order for the presented result to be believable is rarely transparently stated. Third, the fact that Bayesian causal inference either relies on counterfactual priors that are either too restrictive, or too vague and not interpretable.

We have introduced Credible Assumption Mixtures, a methodology to explicitly address these problems by first making explicit the amount of belief that we need to have in an identification assumption to support a certain result, and second by allowing analysts to combine several stronger and weaker identification assumptions into one robust result. Doing this, allows analysts to specify priors on counterfactual parameters that are easily interpretable as statements on the data generating process.

We have given a fully bayesian formulation for our framework as well as tools for easy and general-case inference. We have presented several kinds of common identification assumptions that are found in the causal inference literature, but rarely used in applied research in political science. Simple ways to represent these assumptions as priors on the counterfactual parameter in our framework have been given.

A way to use our proposed method as a robustness check has been introduced as well, with a quadratic program formulation that returns the set of weights that puts the least amount of weight on the strongest assumptions, given a desired credible interval for the treatment effect of interest.

Simulation results, as well as a real-data application show that our method performs well in settings in which the data is unreliable or there are unobservables factors potentially confounding the causal relationship

of interest. These results show that our proposed methodology can be a useful and powerful tool to conduct more reliable and transparent causal inference in the social sciences.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Michael Chasukwa, Asiyati L Chiweza, and Mercy Chikapa-Jamali. Public participation in local councils in malawi in the absence of local elected representatives-political eliticism or pluralism? *Journal of Asian and African studies*, 49(6):705–720, 2014.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Working paper*, 2018.

Alexander Coppock, Alan S Gerber, Donald P Green, and Holger L Kern. Combining double sampling and bounds to address nonignorable missing outcomes in randomized experiments. *Political Analysis*, 25(2): 188–206, 2017.

Gary W Cox and Mathew D McCubbins. Electoral politics as a redistributive game. *The Journal of Politics*, 48(2):370–389, 1986.

Avinash Dixit and John Londregan. The determinants of success of special interests in redistributive politics. *the Journal of Politics*, 58(4):1132–1155, 1996.

Simon Ejdemyr, Eric Kramon, and Amanda Lea Robinson. Segregation, ethnic favoritism, and the strategic targeting of local public goods. *Comparative Political Studies*, 51(9):1111–1143, 2018a.

Simon Ejdemyr, Eric Kramon, and Amanda Lea Robinson. Replication data for: Segregation, ethnic favoritism, and the strategic targeting of local public goods. 2018b. doi: https://doi.org/10.1177%2F0010414017730079.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Paul Gustafson. Bayesian inference for partially identified models. *The International Journal of Biostatistics*, 6(2), 2010.

P Richard Hahn, Jared S Murray, and Ioanna Manolopoulou. A bayesian partial identification approach to inferring the prevalence of accounting misconduct. *Journal of the American Statistical Association*, 111 (513):14–26, 2016.

Jens Hainmueller and Chad Hazlett. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168, 2014.

Peter D Hoff. *A first course in Bayesian statistical methods*. Springer Science & Business Media, 2009.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Hemant Ishwaran, J Sunil Rao, et al. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

Joseph B Kadane. The role of identification in bayesian theory. *Studies in Bayesian econometrics and statistics*, 1975.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993.

Charles F Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

Charles F Manski and John V Pepper. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010, 2000.

Charles F Manski and John V Pepper. More on monotone instrumental variables. *The Econometrics Journal*, 12(s1), 2009.

Walter R Mebane and Paul Poast. Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Political Analysis*, 21(2):233–251, 2013.

Tam O'Neil, Diana Cammack, Edge Kanyongolo, Moir Walita Mkandawire, Tuntufye Mwalyambwire, Bryn Welham, and Leni Wild. Fragmented governance and local service delivery in malawi. *London: Overseas Development Institute*, 2014.

Sean F Reardon and David O'Sullivan. Measures of spatial segregation. *Sociological Methodology*, 34(1): 121–162, 2004.

Paul R Rosenbaum and Donald B Rubin. 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, 1983.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

Supplementary Material

8 Additional Equations

$$
\begin{aligned}
f_{\mu(t,x)|\mathcal{D}}(u) &= f_{\phi(t,x)e(t,x)+\theta(t,x)(1-e(t,x)|\mathcal{D}}(u) \\
&= \int_v f_{\phi(t,x),\theta(t,x)|\mathcal{D}}\left(\frac{v}{e(t,x)}, \frac{u-v}{1-e(t,x)}\right) dv \\
&= \int_v g_{\mathcal{D}}\left(\frac{v}{e(t,x)}; x, t, \boldsymbol{\eta}\right) h\left(\frac{u-v}{1-e(t,x)}; x, t, \frac{v}{e(t,x)}, \boldsymbol{\eta}\right) dv. \qquad (9)
\end{aligned}
$$

$$
\begin{aligned}
f_{\tau(t,s,x)|\mathcal{D}}(\tau) &= f_{\mu(t,x)-\mu(s,x)|\mathcal{D}}(\tau) \\
&= \int_u f_{\mu(t,x)|\mathcal{D}}(\tau+u) f_{\mu(s,x)|\mathcal{D}}(u) du \\
&= \int_u \int_v g_{\mathcal{D}}\left(\frac{v}{e(t,x)}; x, t, \boldsymbol{\eta}\right) h\left(\frac{\tau+u-v}{1-e(t,x)}; x, t, \frac{v}{e(t,x)}, \boldsymbol{\eta}\right) dv \\
&\quad \times \int_v g_{\mathcal{D}}\left(\frac{v}{e(s,x)}; x, s, \boldsymbol{\eta}\right) h\left(\frac{u-v}{1-e(s,x)}; x, s, \frac{v}{e(s,x)}, \boldsymbol{\eta}\right) dv. \qquad (10)
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(a \leq \tau(t,s,x) \leq b|\mathcal{D}) = \int_a^b f_{\tau(t,s,x)|\mathcal{D}}(\tau) d\tau \\
&= \int_a^b \int_u \int_v g_{\mathcal{D}}\left(\frac{v}{e(t,x)}; x, t, \boldsymbol{\eta}\right) \sum_{l=1}^{L} h_{tl}\left(\frac{\tau+u-v}{1-e(t,x)}; x, t, \phi(t,x), \boldsymbol{\eta}\right) dv \\
&\qquad \times \int_v g_{\mathcal{D}}\left(\frac{v}{e(s,x)}; x, s, \boldsymbol{\eta}\right) \sum_{r=1}^{R} h_{sr}\left(\frac{u-v}{1-e(s,x)}; x, s, \phi(s,x), \boldsymbol{\eta}\right) dv du d\tau. \\
&= \sum_{l=1}^{L} \pi_{tl} \sum_{r=1}^{R} \pi_{sr} \int_a^b \int_u \int_v g_{\mathcal{D}}\left(\frac{v}{e(t,x)}; x, t, \boldsymbol{\eta}\right) h_{tl}\left(\frac{\tau+u-v}{1-e(t,x)}; x, t, \phi(t,x), \boldsymbol{\eta}\right) dv \\
&\qquad \times \int_v g_{\mathcal{D}}\left(\frac{v}{e(s,x)}; x, s, \boldsymbol{\eta}\right) h_{sr}\left(\frac{u-v}{1-e(s,x)}; x, s, \phi(s,x), \boldsymbol{\eta}\right) dv du d\tau. \qquad (11)
\end{aligned}
$$

9 Simulation Details

Parameters that are drawn from some distribution were redrawn at each of the 50 simulations performed to obtain the results in Table 1.

Parameter	Value
ϕ_0	0.1
ϕ_1	0.9
θ_0	0.5
θ_1	0.5
$\Pr(T = 1)$	0.5
$Y(1)\|T = 1$	$Binomial(50, \phi_1)$
$Y(0)\|T = 0$	$Binomial(50, \phi_0)$
$ATE = 0.4$	

Table 2: Parameter values for Simulation 1.

Parameter	Value
N	500
P	10
β_j	Uniform(0, 1)
α	1
τ	5
X_{ij}	$\mathcal{N}(0, 1)$
D_i	$Bernoulli(0.5)$
U_i	$Uniform(0, 10)$
σ^2	3
ϵ_i	$\mathcal{N}(0, \sigma^2)$

Table 3: Parameter values for Simulation 2.

Parameter	Value
N	500
P	10
β_j	Uniform(0, 1)
λ_j	Uniform(0, 1)
τ	5
γ	-3
α	1
σ^2	3
ϵ_i	$\mathcal{N}(0, \sigma^2)$
X_{ij}	$\mathcal{N}(0, 1)$
D_i	$Bernoulli(0.5)$
U_i	$Uniform(-10, 10)$
a_i	$Uniform(-10, 0)$
b_i	$Uniform(0, 20)$
ξ_i^a	$Uniform(-2, 2)$
ξ_i^b	$Uniform(2, 4)$
U_i	$Uniform(D_i a_i + (1 - D_i)\xi_i^a +, D_i b_i + (1 - D_i)\xi_i^b)$

Table 4: Parameter values for Simulation 3.

10 Application Details

The outcome variable is defined as follows:

$$Y_i = \frac{\text{\# of boreholes in district } i \text{ in 1998}}{\sum_{i=1}^{191} \text{\# of boreholes in district } i \text{ in 1998}} - \frac{\text{\# of boreholes in district } i \text{ in 2008}}{\sum_{i=1}^{191} \text{\# of boreholes in district } i \text{ in 2008}}$$

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Ethno-Linguistic Fractionalization	191	0.409	0.243	0.013	0.186	0.599	0.841
Population Density	191	−1.583	0.920	−4.368	−2.081	−1.218	2.209
Urban Population	191	−5.342	2.268	−6.908	−6.908	−3.507	0.001
Area size	191	5.679	0.890	1.978	5.284	6.291	7.093
Boreholes per 10,000 people in 1998	191	0.578	0.821	0.000	0.000	0.897	5.196
Share of representative coethnics	191	0.612	0.305	0.013	0.341	0.899	0.992
Share of president's coethnics	191	0.149	0.174	0.0004	0.007	0.314	0.490
Distance to Nearest City (ln)	191	5.623	0.361	4.923	5.406	5.767	6.580
Water Aid Projects per 10,000 residents	191	0.082	0.277	0	0	0	3
All Aid Projects per 10,000 residents	191	1.533	2.662	0.000	0.258	1.460	22.749

Table 5: Descriptive statistics for control convariates in the application.