Friday, September 9, 2022

Replication materials for this manuscript are available at https://github.com/ajnafa/Latent-Bayesian-MSM/tree/apsa-paper

Taking Uncertainty Seriously[*]

Bayesian Marginal Structural Models for Causal Inference in Political Science

A. Jordan Nafa[†]
University of North Texas

Andrew Heiss
Georgia State University

ABSTRACT The past two decades have been characterized by considerable progress in developing approaches to causal inference in situations where true experimental manipulation is either impractical or impossible. With few exceptions, however, commonly employed techniques in political science have developed largely within a frequentist framework. In this article, we argue that common approaches rest fundamentally upon assumptions that are difficult to defend in many areas of political research and highlight the benefits of quantifying uncertainty in the estimation of causal effects. Extending the approach to causal inference for cross-sectional time series and panel data under selection on observables introduced by Blackwell and Glynn (2018), we develop a two-step Bayesian pseudo-likelihood method for estimating marginal structural models. We demonstrate our proposed procedure via a simulation study and two empirical examples. Finally, we provide flexible open-source software implementing the proposed method.

[*] This manuscript was prepared for the American Political Science Association's Annual Meeting in Montreal, Quebec, September 15–18, 2022.

[†] Corresponding author.

KEYWORDS Bayesian statistics; causal inference; panel data

Research in political science and public policy is often concerned with causal effects. Does a specific piece of legislation reduce intergenerational poverty? Does a US Supreme Court ruling cause increased polarization? Do gender-based legislative quotas cause improved gender-based economic outcomes? Do international economic sanctions prevent states from going to war? Quantitatively estimating these causal effects, however, is difficult task fraught with methodological pitfalls. Across scientific disciplines, randomized controlled trials have long been held up as the gold standard in causal thinking. If treatment conditions are randomly assigned across comparable samples of a population, differences in those samples' outcomes can be attributed directly to the treatment and researchers can safely tell causal stories. For many—if not most-questions in political science, however, estimating causal effects through experiments can be difficult, unethical, or impossible. Researchers cannot randomly assign countries to go to war, randomly assign states to adopt specific policies, or randomly assign legislators to win or lose elections.

In the absence of experimental data on most political phenomena, researchers must work with observational data. However, existing observational data reflects already-realized outcomes. For instance, a country's level of level of democracy, legislative system, and other political choices are all influenced by decades of prior institutional, political, social, and economic choices, as well as broader geographic and historical trends. Similarly, the behavior of individuals such as residents, voters, legislators, and political leaders is influenced by a host of other external factors. Observations in a dataset—be they individuals, Census blocks, states, or countries—self-select into (or out of) treatment conditions. As a result, we cannot directly compare observations that chose specific treatments.

A robust cross-disciplinary body of methods has emerged in the past decades to tackle the problem of observational causal inference. Econometricians have focused on quasi-experimental methods such as difference-in-difference analysis, regression discontinuity designs, instrumental variable approaches, and synthetic controls (Angrist and Pischke 2009). Epidemiologists and biostatisticians, on the other hand, have developed matching and weighting techniques based on causal models and *do*-calculus (Pearl, Glymour, and Jewell 2016; Pearl and Mackenzie 2018; Robins, Hernán, and Wasserman 2015). Each approach is designed to account for endogenous self-selection bias and recover causal effects in non-experimental data.

In parallel to developments in observational causal thinking, advances in computational power in the past decade have led to the broader adoption of Bayesian statistical methods. Bayesian analysis provides an alternative to more more standard null hypothesis significance testing (NHST), where researchers test for the probability of observed data given a null hypothesis, or $P(\text{data} \mid H_0)$. In contrast, Bayesian analysis determines the probability of a specific hypothesis given the existing data, or $P(H \mid \text{data})$. When measuring the uncertainty of estimates under NHST, we calculate confidence intervals (often at a 95% level), which indicate that if we collected our data many more times, 95% of those confidence intervals would contain the true population parameter we are interested in. With Bayesian methods, uncertainty can be described with an entire posterior distribution of plausible values that can be summarized in various ways. We can use the posterior distribution to calculate a credible interval (at 95%, for instance) that would allow us to say that there is a 95% probability that the true population parameter falls in the range of the interval. This Bayesian estimand captures what researchers are most often interested in—the probability of a hypothesis being true, rather than the probability of seeing an effect of a certain size in a world where there is a null effect.

This headway in both observational causal inference and Bayesian methods lays the foundation for analysis of causal effects that is more robust, deals more directly with uncertainty, and is more easily interpretable. Most statistical approaches for causal inference, such as difference-in-differences, regression discontinuity designs, and other quasi-experimental designs, map easily into a Bayesian paradigm. Unfortunately, this is not the case with research designs that rely on matching and weighting. Due to mathematical and philosophical incompatibilities, causal inference methods that rely on propensity multiple stages of models (i.e. a design model that estimates the probability of an observation selecting into treatment, which we then use to generate propensity scores and weights that we then use in an outcome model to estimate the effect of treatment) cannot use a Bayesian approach.

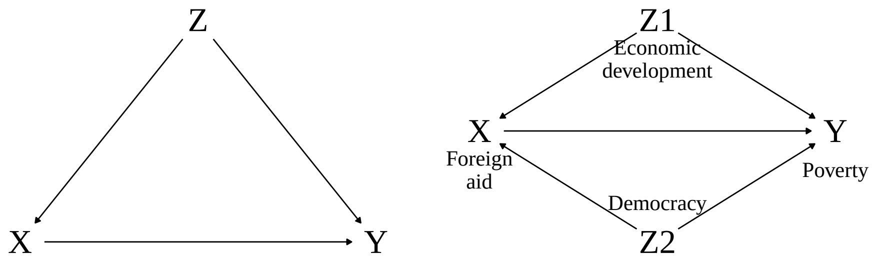
Because propensity scores and weights are not part of the formal data-generating process for the relationship between a treatment and an outcome, we cannot model them with Bayesian methods. In a critique of attempts in bio-statistics at calculating propensity scores with Bayesian methods, Robins, Hernán, and Wasserman conclude that "Bayesian inference must ignore the propensity score" (2015, p. 297).

In this article, we seek to reconcile these parallel developments, merging Bayesian analysis with propensity score-based causal estimation methods. We introduce a method for working with propensity scores in a pseudo-Bayesian manner, allowing researchers to work with posterior distributions of causal effects and make causal inferences without the need for null hypotheses. In particular, we provide a general method for incorporating propensity score-based weights from a treatment model into an outcome model in the spirit of Bayes, propagating the uncertainty of these weights from the treatment model to the outcome model. The approach applies to situations where confounding is addressed through inverse probability weighting, both in a simple single-time-period setting and in a time-series cross-sectional panel data setting with marginal structural models (Robins, Hernán, and Brumback 2000; Blackwell and Glynn 2018). We begin with a brief overview of confounding and endogeneity, introducing concepts from the language of causal graphs and *do*-calculus to help isolate causal effects (Pearl, Glymour, and Jewell 2016; Pearl and Mackenzie 2018) and describe how to statistically adjust for confounding with both inverse probability weighting and marginal structural models. We then explore the methodological reasons why Bayesian analysis is incompatible with these propensity score-based approaches to causal inference and propose an alternative (and compatible) pseudo-Bayesian estimator. Finally, we compare this estimator to existing approaches and demonstrate its results through simulation and by replicating previous research.

Dealing with confounding

Recent developments in causal inference provide a standardized vocabulary and systematized calculus for discussing causal effects through directed acyclic graphs (DAGs) (Pearl, Glymour, and Jewell 2016; Pearl and Mackenzie 2018; Robins, Hernán, and Wasserman 2015). In the left panel of Figure 1, we can see the relationship between some treatment or intervention X (e.g. legislative quotas, foreign aid, economic sanctions, etc.) and an outcome Y (e.g. improved minority representation in parliament, reduced poverty, decreased likelihood of war, etc.). The causal effect of X on Y is represented with an arrow connecting the two nodes. A third node Z exerts a causal effect on both the treatment X and the outcome Y. This variable confounds the $X \rightarrow Y$ relationship and opens up an alternative path between the treatment and outcome. For example, if we are interested in measuring the causal effect of foreign aid (X) on poverty (Y) (see the right panel of Figure 1) and we observe a positive correlation between the two in a dataset, that correlation could be caused by some other confounding factor, like a country's level of economic development (Z_1), or improvements in a country's overall level of democracy or economic development (Z_2).

Figure 1: (L) DAG showing the causal effect of of X on Y, confounded by Z; (R) DAG showing the causal effect of foreign aid on poverty, confounded by $Z1$ (economic development) and $Z2$ (democracy)



We can remove the effect of confounders like Z, Z_1, and Z_2 through statistical adjustment. For instance, if we compare countries with similar (or even identical) levels GDP per capita and democracy, we mitigate the confounding effects of economic development democratic development.

We can use a variety of methods to deal with backdoor confounding. Quasi-experimental research designs like difference-in-differences, regression discontinuity, instrumental variables, and synthetic controls each use specific real-world situations to approximate treatment and control groups to remove the effect of confounders on the relationship between treatment and outcome. Alternatively, we can adjust for confounders through matching and weighting. One common approach—particularly in epidemiology and biostatistics—is to adjust for confounding using inverse probability of treatment weights (IPTWs).

Inverse probability weighting

Causal inference using inverse probability weighting involves a two-stage estimation process. In the first stage, often called the *treatment stage* or *design stage*, we create a model that predicts an observation's choice to receive the treatment based on all confounders identified in a causal graph. We then use the design model to calculate a propensity score for each observation. Next, we calculate inverse probability of treatment weights for each observation based on its propensity of treatment. Weights are calculable for both binary and continuous treatments:

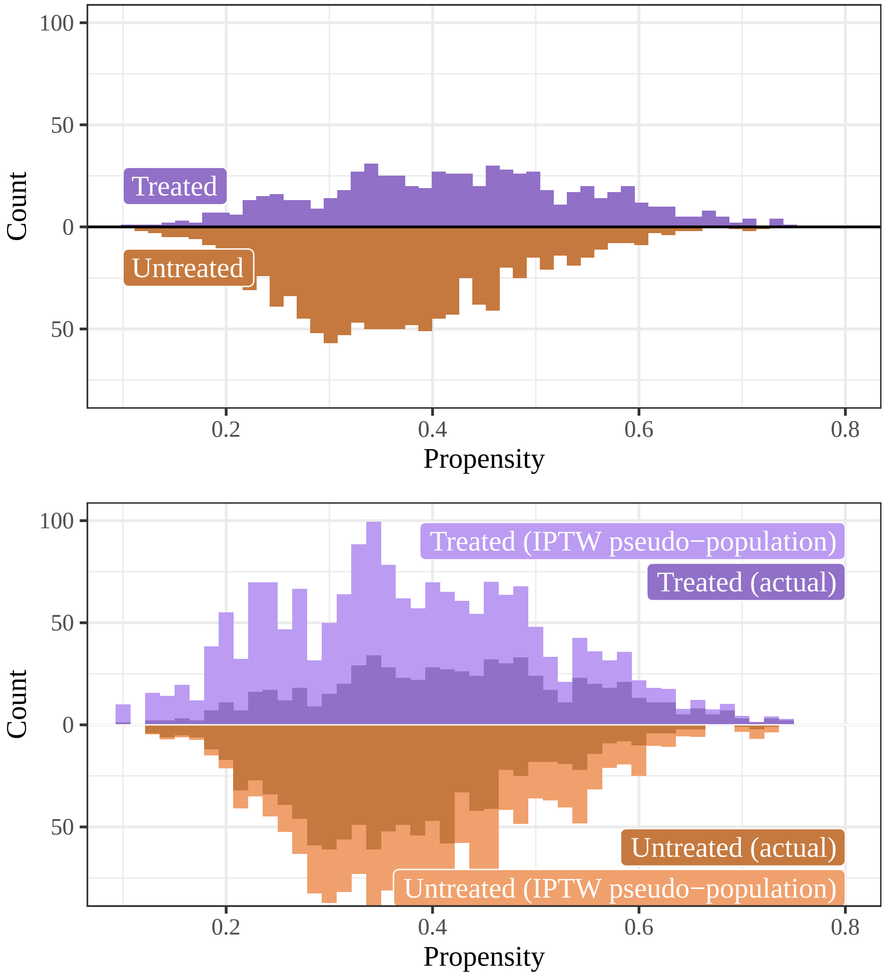
$$w_{\text{binary}} = \frac{\overbrace{X_i}^{\text{Treatment}}}{\underbrace{\pi_i}_{\text{Propensity}}} + \frac{1 - X_i}{1 - \pi_i} \tag{1}$$

$$w_{\text{continuous}} = \frac{\overbrace{f_{X_i}(X_i; \mu, \sigma^2)}^{\text{Probability distribution of treatment } X}}{\underbrace{f_{X_i|Z_i}(X_i \mid Z_i; \mu, \sigma^2)}_{\substack{\text{Probability distribution of treatment } X \\ \text{given confounders } Z}}} \tag{2}$$

The ultimate purpose of these weights IPTWs (w) is to create pseudo-populations of treated and untreated observations that are comparable across all levels of confounders. We can give less weight to observations with a low probability of being treated and who subsequently were not treated and more statistical weight to observations with a high probability of being treated but who were not. Conversely, we can give more weight to treated observations with a low probability of being treated and less weight to treated observations with a high probability of being treated. After scaling each observation by this IPTW, we can create comparable treated and untreated pseudo populations.

Figure 2 demonstrates the intuition visually using a simulated binary treatment. In the top panel of Figure 2, fewer observations received the

Figure 2: Distribution of both original and weighted propensity scores; weighted scores represent comparable pseudo-populations of treated and untreated observations



treatment than not, and those who did not had a lower average propensity of treatment, visible in the cluster in the lower range of propensities. This represents selection bias—those who did not receive the treatment already had a low probability of seeking out the treatment in the first place. As a result, the treated and untreated populations are not comparable. The bottom panel of Figure 2 shows the distribution of propensity scores after weighting by the IPTWs. The lighter distributions represent pseudo-populations of treated and untreated observations, and these two groups now mirror each other fairly well. Both low-propensity treated observations and high-propensity untreated observations are scaled up and receive more statistical weight to improve cross-group comparability.

Having calculated IPTWs and created comparable pseudo-populations, the final stage in causal estimation is to create an outcome model the estimates the effect of the treatment on the outcome, weighted by w. The resulting effect of X on Y represents a causal effect, with all observable confounding accounted for and adjusted away.

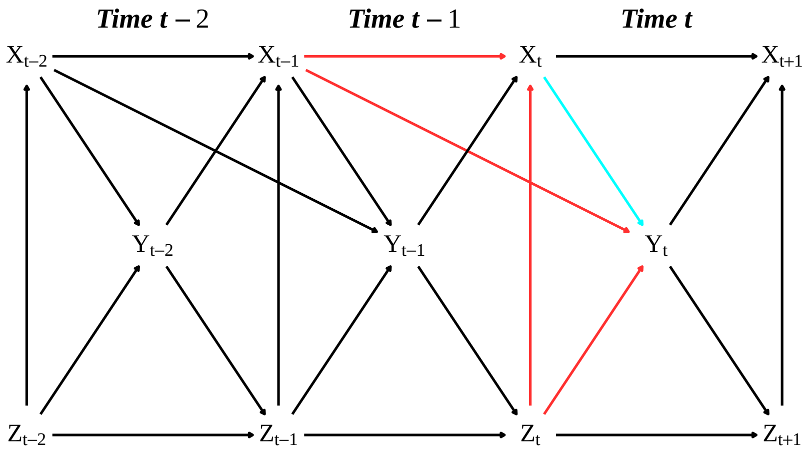
Marginal structural models

The IPTW approach to adjusting for confounding can be extended to more complex data generating processes where treatments, outcomes, and confounders vary over time. In these cases, marginal structural models allow us to adjust for time-invariant confounders, time varying confounders, previous levels of the outcome, and treatment history (Robins 1997; Robins, Hernán, and Brumback 2000; Cole and Hernán 2008; Imai and Ratkovic 2015; Blackwell and Glynn 2018; Saarela et al. 2015).

Figure 3 illustrates one possible DAG showing the relationship between treatment X, outcome Y, and confounders Z in three different time periods (t, $t-1$, and $t-2$). Following the logic of *do*-calculus, if we are interested in measuring and isolating the contemporaneous effect of X_t on Y_t, we no longer need to adjust only for Z_t, as we did in the simpler IPTW example above. The previous value of X, or the treatment history X_{t-1} is now also a confounder that needs to be accounted for statisticially.

When creating IPTWs for data generating processes with repeated measures, as in Figure 3, the resulting weights need to account for the temporal nature of the data and incorporate weights from previous time

Figure 3: DAG showing the contemporaneous effect of X_t on Y_t, given contemporaneous confounders Z_t and treatment history X_{t-1}



periods. This can be done by taking the cumulative product of each observation's weights over time, both for binary and continuous weights:

$$w_{it;\text{ binary}} = \prod_{t=1}^{t} \frac{\Pr[X_{it} \mid X_{it-1},\ C_i]}{\Pr[X_{it} \mid \underbrace{Z_{it}}_{\substack{\text{Time-varying}\\ \text{confounders}}},\ X_{it-1},\ Y_{it-1},\ \underbrace{C_i}_{\substack{\text{Time-invariant}\\ \text{confounders}}}]} \tag{3}$$

$$w_{it;\text{ continuous}} = \prod_{t=1}^{t} \frac{\overbrace{f_{X_{it}|X_{it-1},C_i}[(X_{it} \mid X_{it-1},\ C_i);\ \mu,\ \sigma^2]}^{\substack{\text{Probability distribution of treatment } X \text{ given past}\\ \text{treatment } X_{t-1} \text{ and time-invariant confounders } C}}}{\underbrace{f_{X_{it}|Z_{it},X_{it-1},Y_{it-1},C_i}[(X_{it} \mid Z_{it},\ X_{it-1},\ Y_{it-1},\ C_i);\ \mu,\ \sigma^2]}_{\substack{\text{Probability distribution of treatment } X \text{ given time-varying confounders } Z,\\ \text{past treatment } X_{t-1}, \text{ past outcome } Y_{t-1}, \text{ and time-invariant confounders } C}}} \tag{4}$$

The process for estimating a causal effect using a marginal structural model follows the same two-stage procedure as standard IPTW adjustment. We first use a design stage model to predict treatment status using all backdoor confounders identified in a causal graph, including past treatment status, past level of the outcome, time-invariant covariates, and time-varying covariates. We then generate IPTWs using either Equation 3 or Equation 4, depending on the nature of the treatment variable. We finally fit an outcome model weighted by w to estimate the adjusted effect of the treatment on the outcome.

Bayesian propensity scores

Adjustment through inverse probability weighting—both with single time periods and with more complex marginal structural models—is typically done with frequentist statistical methods. The design stage model is ordinarily fit using logistic regression, while the outcome model is fit using weighted least squares regression with standard errors adjusted post-estimation through bootstrapping (Hernán and Robins 2020, 152). In the case of marginal structural models, the outcome stage typically uses generalized estimating equations (GEE) to account for the panel structure of the data (Thoemmes and Ong 2016; Hernán and Robins 2020, 147). Why bother developing a Bayesian procedure in the first place then if these alternative approaches already exist?

First, in practical terms there is no shortage of examples demonstrating Bayesian estimators often have superior long-run properties and provide more accurate estimates than their frequentist alternatives in many common causal modeling applications including, though not necessarily limited to, matching and propensity score methods (Alvarez and Levin 2021; Capistrano, Moodie, and Schmidt 2019; Kaplan and Chen 2012; Liao and Zigler 2020; Zigler and Dominici 2014), g-computation (Keil et al. 2017), instrumental variable estimation (Hollenbach, Montgomery, and Crespo-Tenorio 2018), and in the presence of heterogeneous treatment effects and high dimensional applications more broadly (Antonelli, Papadogeorgou, and Dominici 2020; Pang, Liu, and Xu 2021; Hahn, Murray, and Carvalho 2020). If, as researchers, we aspire to be *less wrong*, it is worth thinking seriously about a proper approach to quantifying uncertainty in our causal estimands.

Second, in cases where the data comprise an apparent population, for example all sovereign countries between 1945 and 2020, uncertainty estimates and test statistics derived from asymptotic assumptions of repeated sampling and exact long-run replications that form the foundation of the null hypothesis significance testing (NHST) framework and valid interpretations of confidence intervals are logically difficult to defend (Berk, Western, and Weiss 1995; Gill and Heuberger 2020; Western and Jackman 1994). Under such circumstances a Bayesian framework provides a principled and logically consistent alternative that allows us to quantify the uncertainty in our parameter estimates conditional on the observed data and prior assumptions about the universe of effect sizes we believe to be theoretically possible (Gelman and Shalizi 2012; Jackman 2004).

Finally, a Bayesian framework also provides us with the ability to acknowledge that we are virtually always uncertain about the set of confounders we need to adjust for in observational settings and employ model averaging or stacking based approaches to average across different specifications for the design-stage model (Yao et al. 2018; Hollenbach and Montgomery 2020). We can then derive a distribution of weights as illustrated below and propagate the uncertainty inherent in design stage estimation on to the outcome stage model. Where residual correlation within units is a concern, Bayesian hierarchical approaches provide an alternative to GEE models as it is possible to obtain a population-averaged estimate by integrating out the group-level effects.

The impossibility of true two-stage Bayesian analysis

While Bayesian approaches to causal inference provide a wealth of information about the uncertainty and distributions of estimates, Bayesian regression is not directly applicable to two-stage models involving propensity scores or inverse probability weights, for both statistical and philosophical reasons (Robins, Hernán, and Wasserman 2015). To explore this incompatibility, we can define the average treatment effect (ATE) of a binary intervention with the estimand in Equation 5. Here we subtract the average outcome Y when treatment T is both 0 and 1, following adjustment for confounders X.

$$\Delta_{\text{ATE}} = E[\quad \overbrace{E\left(Y_i \mid T_i = 1, X_i\right)}^{\substack{\text{Average outcome } Y \text{ when} \\ \text{treated, given confounders } X}} \quad - \quad \overbrace{E\left(Y_i \mid T_i = 0, X_i\right)}^{\substack{\text{Average outcome } Y \text{ when} \\ \text{not treated, given confounders } X}} \quad] \tag{5}$$

Expressed more generally, the ATE can be calculated by some arbitrary function f that incorporates information from $\mathbf{T}$, $\mathbf{X}$, and $\mathbf{Y}$, as seen in Equation 6.

$$f(\Delta \mid \mathbf{T}, \mathbf{X}, \mathbf{Y}) \tag{6}$$

This function f can represent any kind of estimation approach, including two-stage inverse probability weighting, matching, or design-based quasi-experimential strategies. This more general definition of the ATE also fits well in a Bayesian paradigm. Since Δ is unknown, we can can build a Bayesian model to estimate an unknown θ parameter and set $\theta = \Delta$, conditional on a likelihood for $(\mathbf{T}, \mathbf{X}, \mathbf{Y})$ and a prior distribution for θ (Liao and Zigler 2020). We can thus express Equation 6 using Bayes' formula, as seen in Equation 7. With observed data $\mathbf{T}$, $\mathbf{X}$, and $\mathbf{Y}$, we can proceed with model fitting and sampling and obtain an estimate for θ, or our ATE Δ.

$$\overbrace{P[\theta \mid (\mathbf{T}, \mathbf{X}, \mathbf{Y})]}^{\substack{\text{Posterior estimate} \\ \text{of } \theta \text{, given data}}} \propto \overbrace{P[(\mathbf{T}, \mathbf{X}, \mathbf{Y}) \mid \theta]}^{\substack{\text{Likelihood for existing} \\ \text{data, given unknown } \theta}} \times \overbrace{P[\theta]}^{\substack{\text{Prior} \\ \text{for } \theta}} \tag{7}$$

Crucially, however, Equation 7 does not include any propensity scores or weights. Inverse probability of treatment weights are not part of the data-generating process for θ and thus are not part of either the likelihood or the prior. Weights are a part of the estimation process, not reality. We use weights solely for approximating pseudo populations of treated and untreated observations—these weights do not determine observations' actual behavior or cause changes in the outcome, and therefore are not defined as part of the formal model for θ. The absence of propensity scores or weights in the likelihood is the foundation for Robins, Hernán, and Wasserman (2015)'s critique of Bayesian attempts at causal inference with inverse probability weighting. They explain that "Bayesian logic is rigidly defined: given a likelihood and a prior, one turns the Bayesian crank to obtain a posterior" (297). There is no place for weights in the Bayesian engine—since weights do not fit in either the likelihood or the prior, the posterior estimate of θ cannot reflect the pseudo-populations required for statistical adjustment of confounders. True Bayesians therefore cannot use inverse probability weights in causal inference.

This is disappointing, given the advantages of Bayesian analysis noted earlier. The ability to make inferences with entire posterior distributions of causal effects rather than frequentist null hypotheses can provide us with richer details about those effects, and modeling the uncertainty of our estimates allows us to better explore the robustness of our findings.

Incorporating propensity scores and weights into Bayesian models

At the end of their critique, Robins, Hernán, and Wasserman (2015) state that though it is not possible to use a fully Bayesian approach with two-stage causal inference model, it is possible to adopt a "Bayes-frequentist" compromise in order to better analyze and work with the uncertainty inherent in causal estimation. Liao and Zigler (2020) propose one such compromise and outline a method of incorporating propensity scores into Bayesian estimation of causal effects. To do so, they represent propensity scores as a new parameter ν that is used by a general function that estimates the causal effect given $\mathbf{T}$, $\mathbf{X}$, $\mathbf{Y}$, and ν, representing the propensity score-based weights. This ν parameter is estimated using a design stage model using both treatment status $\mathbf{T}$ and confounders $\mathbf{X}$. By marginalizing over the distribution of the product of the outcome model and the design model, we can eliminate the ν term, resulting in an estimand in Equation 8 that is identical to Equation 6.

$$\overbrace{f(\Delta \mid \mathbf{T}, \mathbf{X}, \mathbf{Y})}^{\substack{\text{Estimand for} \\ \text{the ATE, without } \nu}} = \int_{\nu} \overbrace{f(\Delta \mid \mathbf{T}, \mathbf{X}, \mathbf{Y}, \nu)}^{\substack{\text{Outcome model} \\ \text{with } \nu}} \overbrace{f(\nu \mid \mathbf{T}, \mathbf{X})}^{\substack{\text{Design model} \\ \text{creating propensity} \\ \text{scores with } T \text{ and } X}} \, \mathrm{d}\nu \tag{8}$$

ν represents the posterior distribution of propensity scores or treatment weights, and it contains complete information about the uncertainty in these weights. By incorporating the posterior distribution of ν into the outcome stage of the model, we're able to propagate the variation in weights into the final estimation. We can thus overcome the main shortcoming of Bayesian approaches to two-stage estimation—weights are now a formal parameter in the model (see Equation 9).

$$\overbrace{P[\theta \mid (\mathbf{T}, \mathbf{X}, \mathbf{Y}, \nu)]}^{\substack{\text{Posterior estimate} \\ \text{of } \theta \text{, given data and weights}}} \propto \overbrace{P[(\mathbf{T}, \mathbf{X}, \mathbf{Y}, \nu) \mid \theta]}^{\substack{\text{Likelihood for existing} \\ \text{data, given unknown } \theta}} \times \overbrace{P[\theta]}^{\substack{\text{Prior} \\ \text{for } \theta}} \tag{9}$$

Instead of calculating a single value of ν (i.e. a single set of propensity scores or weights) from the design stage of the model, we can incorporate a range of values of ν from the posterior distribution of the design model. To do so, we first fit a Bayesian design-stage model to calculate the posterior probabilities of treatment. We then generate K samples of propensity scores from the posterior distribution of the treatment. K can vary substantially, though it is often the number of posterior chains from the Bayesian model. Next, for each of the K samples of scores, we generate inverse probability weights and build an outcome model (either Bayesian or frequenist) using those weights. Finally, we combine and average the results from these many outcome models to calculate the final ν-free causal effect. The procedure is similar to multiple imputation or bootstrapping—we run the same outcome model many times using different variations of weights and then combine and average the results.

Importantly, this approach is not fully Bayesian, but pseudo-Bayesian. The parameters for the analysis are estimated using separate independent models: ν in the design stage and θ in the outcome stage. To qualify as a truly Bayesian approach, ν and θ would need to be estimated jointly and simultaneously.

Bayesian Pseudo-Likelihood Estimation

There are several different approaches one might take to accounting for uncertainty in the design stage weights when estimating the outcome stage of a marginal structural model. It is, for example, possible to pass a different random draw from the distribution of weights directly to the model at each iteration of the MCMC sampler though such an approach quickly becomes intractable in terms of computation. In this section we outline an alternative computationally efficient approach that requires passing only the location and scale of the design stage weights to the outcome model and propagates uncertainty by placing a prior on the scale component of the weights.

To implement our pseudo-likelihood estimator, we take as our starting point recent developments in the application of Bayesian methods for the analysis of complex survey designs (Savitsky and Toth 2016; Williams and Savitsky 2020a, 2020b). Following Savitsky and Toth (2016), we can express the Bayesian pseudo-posterior as

$$\hat{\pi}(\theta \mid y, \tilde{w}) \propto \left[\prod_{i=1}^{n} \Pr(y_i \mid \theta)^{\tilde{w}_i}\right] \pi(\theta) \tag{10}$$

where $\prod_{i=1}^{n} \Pr(y_i \mid \theta)^{\tilde{w}_i}$ represents the pseudo-likelihood of the observed data y and $\pi(\theta)$ is a prior on the unconstrained parameter space.

If we are interested in the average treatment effect of some binary treatment X at times t and $t-1$, the posterior predictive distribution of the stabilized inverse probability weights from the design stage SW is

$$\text{SW}_{it} = \prod_{t=1}^{t} \frac{\int \Pr(X_{it} \mid X_{it-1},\ C_i)\pi(\theta)d\theta}{\int \Pr(X_{it} \mid Z_{it},\ X_{it-1},\ Y_{it-1},\ C_i)\pi(\theta)d\theta} \tag{11}$$

where i and t index groups and periods respectively. The observed treatment status and outcome for the i^{th} group at each period t are represented by X and Y respectively. C is the observed value of the baseline time invariant confounders and Z is a set of time varying covariates that satisfies sequential ignorability (Blackwell and Glynn 2018). While we focus mainly on the average treatment effect at times t and $t-1$ here, it is possible to estimate longer lags, different estimands such as the average treatment effect in the treated, and continuous treatments.

We parameterize the regularized weights for each observation, denoted $\tilde{w}_i$ in Equation 10, in the outcome model as

$$\tilde{w}_i = \lambda_i + \delta_i \cdot \pi(\delta)$$

where λ and δ represent a vector containing the location the weights–typically the mean, though it may be preferable to use the posterior median under certain circumstances–and the scale of the posterior distribution of the stabilized weights for each observation $i \in \{1, 2, \dots, n\}$. At each iteration of the MCMC algorithm, $\pi(\delta)$ is a vector of length n sampled from a prior distribution on the scale of the weights that allows us to both model the uncertainty in the design stage weights and regularize their variance to stabilize computation by ruling out unrealistic or impossible values.

This of course requires the researcher to specify a proper Bayesian prior on the scale component of the weights to avoid severe convergence issues. In practice, we recommend a weakly to moderately informative prior distribution that concentrates the bulk of the prior density between 0 and 1.5. The exponential distribution with rate $\lambda > 3.5$ or Beta distribution with shape parameters $\alpha = 2$ and $\beta \geq 2$ tend to perform well in simulations. Although we do not consider such an approach here, one might also consider regularizing the location of the weights in a similar fashion in cases where there are a large number of observations with excessively large inverse probability weights.

This Bayesian pseudo-likelihood approach lends itself to relatively straight forward extensions such as multilevel regression as recently illustrated by Savitsky and Williams (2021) in the context of weighted survey designs. Perhaps more notably, it allows us to consider more than one possible specification for the design stage model via Bayesian weighting procedures such as model averaging or posterior stacking (Yao et al. 2018; Montgomery and Nyhan 2010; Hollenbach and Montgomery 2020). This may provide a way of reducing the degree to which the outcome stage model is sensitive to the design stage specification (Kaplan and Chen 2014; Zigler and Dominici 2014), a common criticism of propensity score based weighting estimators.

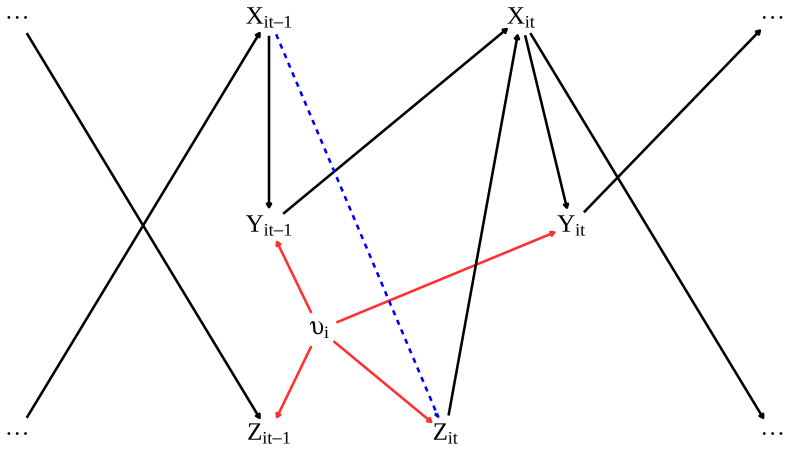
Simulation Study

To evaluate the performance of our proposed model in terms of its ability to recover the true parameter values, we employ a modified version of the simulation study in Blackwell and Glynn (2018). As depicted in Figure 4, we assume that values of X_i at time $t-1$ are independent of outcomes Y_i at time t and that X_i has only a contemporaneous treatment effect on Y_i at each time t–that is, the true lagged treatment effect of X_{it-1} on Y_{it} is 0. Furthermore, past values of Y_{it-1} are independent of Y_{it}, conditional on the treatment X_{it}. To identify the causal effect of X_{it-1} and X_{it} on Y_{it} we need to condition on the minimum adjustment set that blocks the unmeasured time invariant confounder υ_i, which in this case is $\{Y_{it-1}, Z_{it}\}$ and satisfies sequential ignorability for the causal path $X_{it} \longrightarrow Y_{it}$ (1077).

In our simulations we randomly vary whether the time varying covariate Z_{it} depends on past values of the treatment–that is, whether Z_{it}

is endogenous to the treatment X_{it-1}–since under such circumstances Z_{it} is post-treatment with respect to X_{it-1} and conditioning on it results in bias of unknowable direction and magnitude (Blackwell and Glynn 2018; Montgomery, Nyhan, and Torres 2018). A detailed explanation of the data generation process for the simulations and a discussion of additional considerations, interested readers may consult the online appendix.

Figure 4: DAG Depicting the Data Generation Process for the Simulations



Design

To assess how the model performs under different conditions and evaluate the asymptotic properties of our Bayesian pseudo-likelihood procedure, we vary both the number of groups–we consider 25, 45, 65, 85, and 100–and the number of periods per group–20 and 50–which results in $5 \times 2 \times 2$ unique period-group-condition combinations. For each combination, we repeat the simulation 100 times giving us 2,000 simulated data sets in total on which to evaluate the model and covering dimensions that are approximately representative of most cross-sectional time series applications in political science.

For each data set, we estimate the design stage models for the numerator and denominator of the weights via Bayesian logistic regression models of the form

$$\begin{aligned} \Pr(X_{it} = 1 \mid \theta_{it}) &\sim \mathit{Bernoulli}(\theta_{it}) \\ \theta_{it} &= \operatorname{logit}^{-1}(\alpha + X_n \beta_k) \\ &\text{with priors} \\ \alpha \sim \mathit{Normal}(0,\ 2) &\qquad \beta_k \sim \mathit{Normal}(0,\ 1) \end{aligned}$$

where α represents the global intercept, β is a vector of coefficients of length k, and X_n is an $n \times k$ matrix of predictors for the numerator and denominator of the weights. After estimating the location and scale of the distribution of the weights as discussed in the preceding section, we fit an outcome stage model of the form

$$\begin{aligned} y_{it} &\sim \mathit{Normal}(\mu_{it}, \epsilon^2)^{\tilde{w}_{it}} \\ \mu_{it} &= \alpha + \beta_1 X_{it} + \beta_2 X_{it-1} + \epsilon \\ &\text{where} \\ \tilde{w}_{it} &\sim \lambda_{it} + \delta_{it} \cdot \pi(\delta) \\ &\text{with priors} \\ \alpha \sim \mathit{Normal}(\bar{y},\ 2 \cdot \sigma_y) &\qquad \beta_k \sim \mathit{Normal}\left(0,\ 1.5 \cdot \frac{\sigma_y}{\sigma_x}\right) \\ \epsilon \sim \mathit{Exponential}\left(\frac{1}{\sigma_y}\right) &\qquad \delta_{it} \sim \mathit{Beta}(2,\ 5) \end{aligned}$$

The response y is assumed Gaussian with mean μ and variance σ^2 with the pseudo-likelihood of each observation being the product of the likelihood and the sampled weight $\tilde{w}_{it}$. Priors on the coefficients are assigned independent normal priors with mean 0 and standard deviation $1.5 \cdot \frac{\sigma_y}{\sigma_x}$ where σ_x and σ_y are the standard deviation of the predictor and response respectively. We place a slightly more diffuse prior on the global intercept α which is assumed normal with mean $\bar{y}$ and standard deviation $2 \cdot \sigma_y$. For the dispersion parameter ϵ we assign an exponential prior with rate $\frac{1}{\sigma_y}$. This approach automatically adjusts the scale of the priors to the data and can be regarded as weakly to moderately informative (Gelman, Hill, and Vehtari 2020, 124–126). At each iteration of the sampler, the prior on the scale of the weights is drawn from a regularizing Beta with shape parameters $\alpha = 2$ and $\beta = 5$ and the weights calculated for each observation as described in the preceding section.

In addition to our proposed marginal structural model, we also fit an auto-regressive distributed lag specification to each of the simulated data sets of the form

$$
\begin{aligned}
y_{it} &\sim Normal(\mu_{it}, \epsilon^2) \\
\mu_{it} &= \alpha + \beta_1 X_{it} + \beta_2 X_{it-1} + \beta_3 Y_{it-1} + \beta_4 Y_{it-2} + \\
&\quad \beta_5 Z_{it} + \beta_6 Z_{it-1} + \epsilon
\end{aligned}
$$

with priors

$$
\alpha \sim Normal(\bar{y},\ 2 \cdot \sigma_y) \qquad \beta_k \sim Normal\left(0,\ 1.5 \cdot \frac{\sigma_y}{\sigma_x}\right)
$$

$$
\epsilon \sim Exponential\left(\frac{1}{\sigma_y}\right)
$$

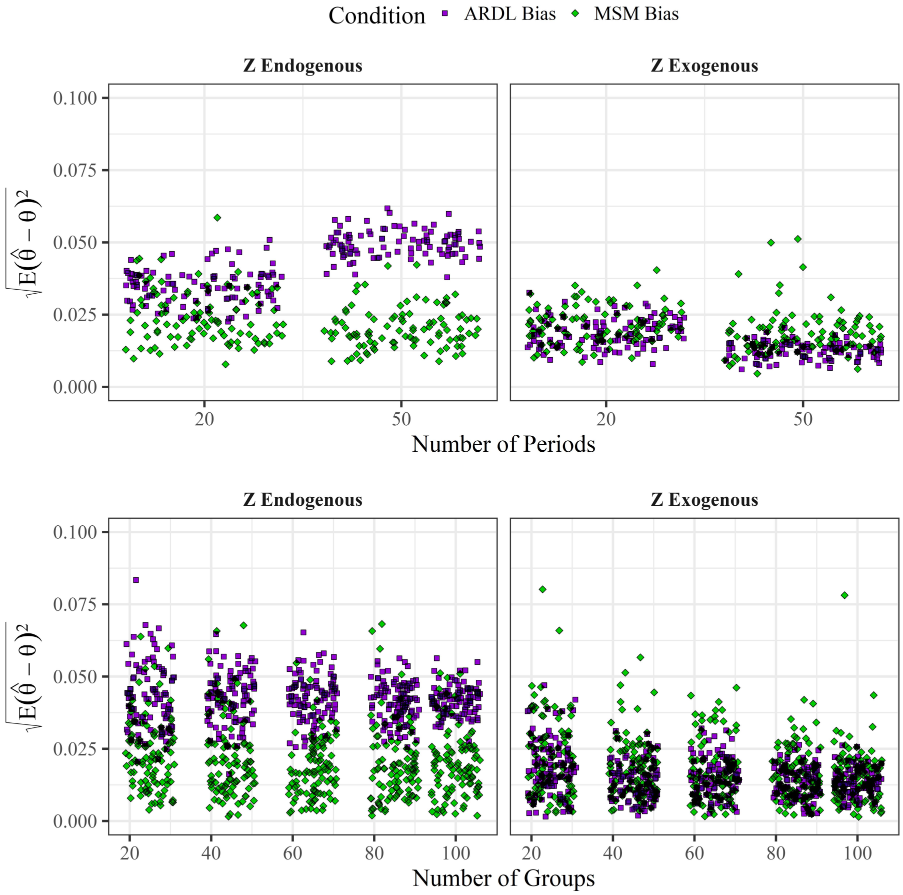
where the priors on the intercept, coefficients, and dispersion parameter are based on the same auto-scaling procedure discussed above.

Estimation is performed under version 2.30 of the probabilistic programming language Stan which implements the No-U-Turn sampler variant of Hamiltonian Markov Chain Monte Carlo (Carpenter et al. 2017; Hoffman and Gelman 2014). For each of the models, we run four markov chains in parallel for 2,000 iterations each, discarding the first 1,000 after the initial warm-up adaptation stage. This number proved sufficient for convergence and leaves us with 4,000 posterior samples per model for subsequent analysis. Stan code for each of these outcome models is provided in the appendix.

Results

The results of the simulations for each model are shown in Figure 5, which depicts estimates for the root mean square error (RMSE) by model and dimensions under each condition for the bias in the estimate of X_{it-1}. We see that our Bayesian pseudo-likelihood estimator performs quite

Figure 5: Simulation Results for the RMSE of the MSM and ARDL Models for the Lagged Treatment Effect



well overall, and as expected, tends to exhibit less bias when Z_{it} is endogenous to the treatment history X_{it-1} compared to the ARDL. Moreover, as the number of observed periods increases, the bias of the ARDL model grows while our MSM provides an approximately unbiased estimate of the average lagged treatment effect.

Figure 6: Distributions of the Posterior Means for the Contemporaneous and Lagged Treatment Effect Estimates from the MSM and ARDL Models

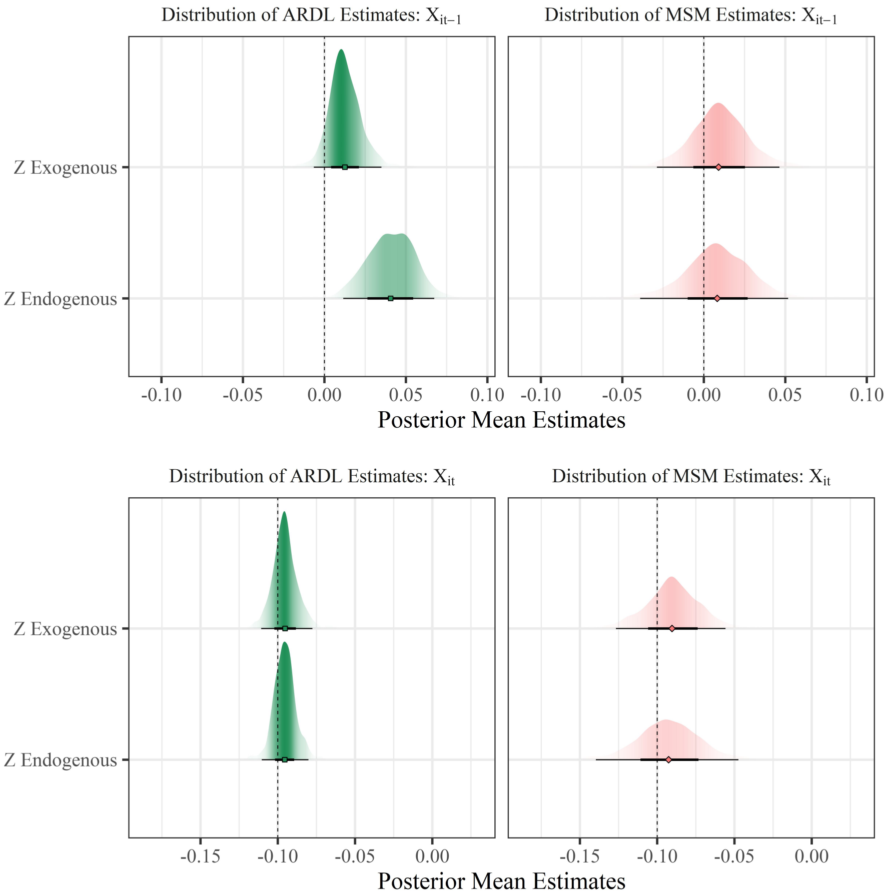


Figure 6 shows the distribution of posterior means for each model by condition and further illustrates that our Bayesian MSM approach performs reasonably well in terms of parameter recovery under both conditions while exhibiting substantially less bias than the ARDL approach in cases where time varying covariates are a function of past treatments. Overall, we see that our Bayesian pseudo-likelihood procedure performs quite well across a range of different scenarios that are roughly typical of data in political science and international relations.

Applied example

TODO: After APSA, apply this by replicating an existing paper

Conclusion

Computational and methodological advances in the past decade have laid the groundwork for substantial advances in quantitative political science and policy research. Developments in observational causal inference—in both quasi-experimental and propsensity score-based approaches—have enhanced the credibility and robustness of research findings, while increased use of Bayesian analysis has led to more results that are more interpretable and deal more directly with the uncertainty of estimates.

As we have demonstrated, however, Bayesian methods are incompatible with approaches to causal inference that rely on propensity scores. To overcome this incompatibility, we propose a pseudo-Bayesian approach to the calculation and use of inverse probability weights that allows researchers to work with posterior distributions that correctly capture and reflect uncertainty.

References

Alvarez, R. Michael, and Ines Levin. 2021. "Uncertain Neighbors: Bayesian Propensity Score Matching For Causal Inference." *arXiv,* 10.48550/arXiv.2105.02362.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Antonelli, Joseph, Georgia Papadogeorgou, and Francesca Dominici. 2020. "Causal Inference in High Dimensions: A Marriage Between Bayesian Modeling and Good Frequentist Properties." *Biometrics* 78 (1): 100–114. https://doi.org/10.1111/biom.13417.

Berk, Richard A., Bruce Western, and Robert E. Weiss. 1995. "Statistical Inference for Apparent Populations." *Sociological Methodology* 25:421. https://doi.org/10.2307/271073.

Blackwell, Matthew, and Adam N. Glynn. 2018. “How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables.” *American Political Science Review* 112 (4): 1067–1082. https://doi.org/10.1017/S0003055418000357.

Capistrano, Estelina S. M., Erica E. M. Moodie, and Alexandra M. Schmidt. 2019. “Bayesian Estimation of the Average Treatment Effect on the Treated using Inverse Weighting.” *Statistics in Medicine* 38 (13): 2447–2466. https://doi.org/10.1002/sim.8121.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. “Stan: A Probabilistic Programming Language.” *Journal of Statistical Software* 76 (1). https://doi.org/10.18637/jss.v076.i01.

Cole, Stephen R., and Miguel A. Hernán. 2008. “Constructing Inverse Probability Weights for Marginal Structural Models.” *American Journal of Epidemiology* 168 (6): 656–664. https://doi.org/10.1093/aje/kwn164.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories.* Cambridge University Press.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2012. “Philosophy and the Practice of Bayesian Statistics.” *British Journal of Mathematical and Statistical Psychology* 66 (1): 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x.

Gill, Jeff, and Simon Heuberger. 2020. “Bayesian Modeling and Inference: A Post-Modern Perspective.” In *The SAGE Handbook of Research Methods in Political Science and International Relations,* edited by Luigi Curini and Robert Franzese, 961–984. London, UK: SAGE.

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. 2020. “Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion).” *Bayesian Analysis* 15 (3). https://doi.org/10.1214/19-ba1195.

Hernán, Miguel A., and James M. Robins. 2020. *Causal Inference: What If.* Boca Raton, Florida: Chapman and Hall / CRC.

Hoffman, Matthew D., and Andrew Gelman. 2014. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.” *Journal of Machine Learning Research* 15 (1): 1593–1623.

Hollenbach, Florian M., and Jacob M. Montgomery. 2020. “Bayesian Model Selection, Model Comparison, and Model Averaging.” In *The SAGE Handbook of Research Methods in Political Science and International Relations,* edited by Luigi Curini and Robert Franzese, 937–960. SAGE Publications Ltd. https://doi.org/10.4135/9781526486387.

Hollenbach, Florian M., Jacob M. Montgomery, and Adriana Crespo-Tenorio. 2018. “Bayesian Versus Maximum Likelihood Estimation of Treatment Effects in Bivariate Probit Instrumental Variable Models.” *Political Science Research and Methods* 7 (3): 651–659. https://doi.org/10.1017/psrm.2018.15.

Imai, Kosuke, and Marc Ratkovic. 2015. “Robust Estimation of Inverse Probability Weights for Marginal Structural Models.” *Journal of the American Statistical Association* 110 (511): 1013–1023. https://doi.org/10.1080/01621459.2014.956872.

Jackman, Simon. 2004. “Bayesian Analysis for Political Research.” *Annual Review of Political Science* 7 (1): 483–505.

Kaplan, David, and Jianshen Chen. 2012. “A Two-Step Bayesian Approach for Propensity Score Analysis: Simulations and Case Study.” *Psychometrika* 77 (3): 581–609. https://doi.org/10.1007/s11336-012-9262-8.

———. 2014. “Bayesian Model Averaging for Propensity Score Analysis.” *Multivariate Behavioral Research* 49 (6): 505–517. https://doi.org/10.1080/00273171.2014.928492.

Keil, Alexander P., Eric J. Daza, Stephanie M. Engel, Jessie P. Buckley, and Jessie K. Edwards. 2017. “A Bayesian Approach to the g-formula.” *Statistical Methods in Medical Research* 27 (10): 3183–3204. https://doi.org/10.1177/0962280217694665.

Liao, Shirley X., and Corwin M. Zigler. 2020. “Uncertainty in the Design Stage of Two-Stage Bayesian Propensity Score Analysis.” *Statistics in Medicine* 39 (17): 2265–2290. https://doi.org/10.1002/sim.8486.

Montgomery, Jacob M., and Brendan Nyhan. 2010. “Bayesian Model Averaging: Theoretical Developments and Practical Applications.” *Political Analysis* 18 (2): 245–270. https://doi.org/10.1093/pan/mpq001.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science* 62 (3): 760–775. https://doi.org/10.1111/ajps.12357.

Pang, Xun, Licheng Liu, and Yiqing Xu. 2021. “A Bayesian Alternative to Synthetic Control for Comparative Case Studies.” *Political Analysis* 30 (2): 269–288. https://doi.org/10.1017/pan.2021.22.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer.* Hoboken, New Jersey: Wiley.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect.* New York: Basic Books.

Robins, James M. 1997. “Causal Inference from Complex Longitudinal Data.” In *Latent Variable Modeling and Applications to Causality,* edited by Maia Berkane, 120:69–117. Lecture Notes in Statistics. New York: Springer-Verlag.

Robins, James M., Miguel A. Hernán, and Larry Wasserman. 2015. “Discussion of ‘On Bayesian estimation of marginal structural models’.” *Biometrics* 71 (2): 296–299. https://doi.org/10.1111/biom.12273.

Robins, James M., Miguel Ángel Hernán, and Babette Brumback. 2000. “Marginal Structural Models and Causal Inference in Epidemiology.” *Epidemiology* 11 (5): 550–560. https://doi.org/10.1097/00001648-200009000-00011.

Saarela, Olli, David A. Stephens, Erica E. M. Moodie, and Marina B. Klein. 2015. “On Bayesian estimation of marginal structural models.” *Biometrics* 71 (2): 279–288. https://doi.org/10.1111/biom.12269.

Savitsky, Terrance D., and Daniell Toth. 2016. “Bayesian Estimation Under Informative Sampling.” *Electronic Journal of Statistics* 10 (1). https://doi.org/10.1214/16-EJS1153.

Savitsky, Terrance D., and Matthew R. Williams. 2021. “Pseudo Bayesian Mixed Models under Informative Sampling.” *arXiv,* https://doi.org/10.48550/ARXIV.1904.07680.

Thoemmes, Felix, and Anthony D. Ong. 2016. “A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models.” *Emerging Adulthood* 4, no. 1 (February): 40–59. https://doi.org/10.1177/2167696815621645.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88 (2): 412–423.

Williams, Matthew R., and Terrance D. Savitsky. 2020a. "Bayesian Estimation Under Informative Sampling with Unattenuated Dependence." *Bayesian Analysis* 15 (1). https://doi.org/10.1214/18-BA1143.

———. 2020b. "Uncertainty Estimation for Pseudo-Bayesian Inference Under Complex Sampling." *International Statistical Review* 89 (1): 72–107. https://doi.org/10.1111/insr.12376.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)." *Bayesian Analysis* 13 (3): 917–1007. https://doi.org/10.1214/17-ba1091.

Zigler, Corwin Matthew, and Francesca Dominici. 2014. "Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects." *Journal of the American Statistical Association* 109 (505): 95–107. https://doi.org/10.1080/01621459.2013.869498.