Bayesian Causal Inference: Introduction and Examples

Jizhou Kang and Shuangjie Zhang[1]

Department of Statistics, UC Santa Cruz[1]

05/02/2023

This is the written report of the presentation we gave at the journal club. The authors would like to thank the audiences for their active participation, and Prof. Juhee Lee for providing references and valuable discussions.

1 Introduction

This report examines the Bayesian approach to causal inference under the potential outcomes framework. We closely follows the review paper by Li et al. (2023). We recapitulate the key insights of the paper, while providing examples from other references. The rest of the report is organized as follows. We introduce the potential outcome setup in Section 2 and explain the Bayesian context in Section 3. Bayesian model specification is provided in Section 4. Why and how to use the propensity score in Bayesian causal inference is discussed in Section 5. Section 6 is for sensitivity analysis. In Section 7, we investigate Bayesian causal inference under complex assignment settings through illustrative examples. Finally, we present the concluding remarks in Section 8.

2 Potential outcome setup

Li et al. (2023) focus on the potential outcome framework for causal inference in the paper. For each unit i, we observe the outcome $Y_i(Z_i)$ with treatment $Z_i = 0, 1$ and covariate X_i, where $Z_i = 0$ represents the control group and $Z_i = 1$ represents the treatment group. This results in two potential outcomes $Y_i(0)$ and $Y_i(0)$ for each unit. The main interest in causal inference literature is causal effects defined based on $Y_i(0)$ and $Y_i(1)$. Five main causal effects are listed in Table 1.

Individual treatment effect (ITE) is the treatment effect for each unit. Sample average treatment effect (SATE) is the sample average of ITE in a finite sample. Population average treatment effect (PATE) is a population average level of ITE in a population sense, which require an assumption of the density of covariates X. MATE tries to avoid making this assumption by using the empirical density of covariates from the data. It leads to the sample average of CATE in the finite sample. CATE is the conditional treatment effect, where one

Table 1: Five main causal effects of interest

Causal effect	Formula
Individual Treatment Effect (ITE)	$\tau_i = Y_i(1) - Y_i(0)$
Sample Average Treatment Effect (SATE)	$\tau^S = N^{-1} \sum_{i=1}^N (Y_i(1) - Y_i(0))$
Conditional Average Treatment Effect (CATE)	$\tau(x) = E(Y_i(1) - Y_i(0) \mid X_i = x)$
Population Average Treatment Effect (PATE)	$\tau^P = E(Y_i(1) - Y_i(0)) =$ $E(\tau(X_i)) = \int \tau(x) F_X(dx)$
Mixed Average Treatment Effect (MATE)	$\tau^M = \int \tau(x) \hat{F}_X(dx) = N^{-1} \sum_{i=1}^N \tau(X_i)$

usually puts fancy models on. Linear regressions, trees and random forests among others will be discussed in Section 4. Different options of causal effects are considered subject to the context, questions and dataset.

Under the potential outcome framework, there's only one observed outcome either $Y_i(0)$ or $Y_i(1)$, which needs additional assumptions to identify the causal effects. Most causal studies assume the ignorability assumption in Assumption 1.

Assumption 1. *(Ignorability). (a) Unconfoundness.* $\Pr(Z_i \mid Y_i(0), Y_i(1), X_i) = \Pr(Z_i \mid X_i)$ *or* $Z_i \perp \{Y_i(1), Y_i(0)\} \mid X_i$. *(b) Overlap.* $0 < e(X_i) < 1$, *where propensity score is defined as* $e(X_i) \equiv \Pr(Z_i \mid Y_i(0), Y_i(1), X_i) = \Pr(Z_i \mid X_i)$ *in Rosenbaum and Rubin (1983).*

The unconfoundedness assumption states that there is no unmeasured confounding, and covariates X explain all the confounding. Additionally, it implies that the assignment mechanism does not depend on the outcome Y. It advocates Bayesian methods with the missing at random mechanism. We will see this in Section 3. During the discussion after the presentation, we discuss whether we could assume a more complex case by jointly modeling (Y, Z) conditional on covariate X. We will see two examples of complex assignment mechanisms in Section 7. Besides these two, there is another area called regression discontinuity design (RDD) in applied economics and causal inference, where the joint modeling of (Y, Z) is popular. The policy/treatment Z impact the conditional density of $Y \mid X$. For example, a pass/fail policy may impact the density of grades Y and create a discontinuity at the threshold. We refer to a summary paper for RDD by Jales and Yu (2016).

The overlap assumption states that each unit has non-zero probability of being assigned to each treatment condition. In general, as the two groups overlap more and become more balanced, the causal estimates become less sensitive to the estimate strategy and model specification. We will see one example later to illustrate the big uncertainty by Bayesian models when there is poor overlap. The propensity score is usually used for balancing, matching, or weighting. And we will further elaborate on this later in Section 5.

With the ignorability assumption, the conditional distribution of the potential outcomes is identifiable from observed data. We can assume flexible mean models as $\mu_z(x) \equiv E(Y_i(z) \mid X_i = x) = \mathbb{E}(Y_i \mid Z_i = z, X_i = x)$, then CATE is identified as $\mu_1(x) - \mu_0(x)$, and PATE becomes $\tau^P = \mathbb{E}\{\mu_1(X_i) - \mu_0(X_i)\}$.

3 Bayesian causal inference

From the Bayesian point of view, the potential outcome framework is essentially a missing data problem. Under the assumption of ignorable assignment mechanism, it lies in the missing at random regime. For each unit i, only three out of four quantities $\{Y_i(0), Y_i(1), X_i, Z_i\}$ are observed. There's a one to one relationship between $(Y_i(0), Y_i(1))$ and $(Y_i^{\text{miss}}, Y_i^{\text{obs}})$ that $Y_i^{\text{obs}} = Y_i(Z_i), Y_i^{\text{miss}} = Y_i(1 - Z_i)$. We assume a joint distribution on the complete data likelihood of $\{Y_i(0), Y_i(1), X_i, Z_i\}$ conditional on some parameters θ

$$\Pr(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{Z} \mid \theta) = \prod_{i=1}^{N} \Pr(Y_i(0), Y_i(1), X_i, Z_i \mid \theta). \tag{1}$$

The parameters θ can be decomposed into three aspects $\theta = (\theta_X, \theta_Y, \theta_Z)$. Then we can further write the joint likelihood (1) into

$$\prod_{i=1}^{N} \underbrace{\Pr(Z_i \mid Y_i(0), Y_i(1), X_i; \theta_Z)}_{\text{Propensity score } \Pr(Z_i|X_i;\theta_Z)} \underbrace{\Pr(Y_i(0), Y_i(1) \mid X_i; \theta_Y)}_{\text{Potential outcomes}} \underbrace{\Pr(X_i; \theta_X)}_{\text{Covariates}}.$$

As proposed in the discussion, the factorization order does not change the final result, since under ignorability assumption, $\Pr(Y_i(0), Y_i(1) \mid Z_i, X_i; \theta_Y) = \Pr(Y_i(0), Y_i(1) \mid X_i; \theta_Y)$. We notice that the posterior distribution of θ_X and θ_Y does not depend on the propensity score, and that's why propensity score is ignorable under the Bayesian causal inference context. We will discuss how to incorporate the propensity score into the model in Section 5.

Before we discuss the model specification in Section 4, we first introduce the independent prior assumption which is commonly adopted. The independent prior assumption indicates that θ_X, θ_Y and θ_Z are priori distinct and independent. Under this assumption, the full conditional distribution of θ_X, θ_Y and θ_Z is proportional to

$$\Pr(\theta_X) \prod_{i=1}^{N} \Pr(X_i \mid \theta_X) \Pr(\theta_Y) \prod_{i=1}^{N} \Pr(Y_i(1), Y_i(0) \mid \theta_X; \theta_Y) \Pr(\theta_Z) \prod_{i=1}^{N} \Pr(Z_i \mid X_i; \theta_Z).$$

As mentioned before, the posterior distribution of θ_X and θ_Y does not depend on the propensity score or the assignment mechanism part.

Having posterior samples in hand, we can revisit the estimates for the causal effects of interest. MATE is the easiest one since it utilizes the empirical densities of covariates X. $\tau^M = \int \tau(x; \theta_Y) \hat{F}_X(dx)$, where $\hat{F}_X(dx)$ is the empirical cdf of observed covariates. $\hat{\tau}^M = \mathbb{E}_{\hat{f}_X}(\tau(x; \hat{\theta}_Y^{\text{pos}}))$. PATE needs to specify the model $\Pr(X_i \mid \theta_X)$ on covariates X. Or one can draw X from a Bayesian bootstrap(Rubin, 1981). After obtaining the posterior estimates of $\hat{\theta}_X^{\text{pos}}$, $\hat{\tau}^P = \mathbb{E}_{f_X(\hat{\theta}_X^{\text{pos}})}(\tau(x; \hat{\theta}_Y^{\text{pos}}))$. Both MATE and PATE do not need the dependence between $Y_i(0)$ and $Y_i(1)$ since we don't impute the missing outcomes. However, for SATE $\tau^S = N^{-1} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$, we need to have posterior draws of either $Y(0) \mid Y(1), X_i, \theta_Y$ or $Y(1) \mid Y(0), X_i, \theta_Y$ when one outcome is naturally missing in causal inference. This inference highly depends on the model structure and the prior on the correlation parameter in the model specification due to the lack of knowledge from observed data. That is, we never jointly observe two outcomes $Y(0)$ and $Y(1)$. To conclude, the uncertainty of the three average effects is: $\tau^P > \tau^M > \tau^S$ and the numerical difference is usually very small in real applications.

4 Bayesian model specification

Now we can move to the potential outcome models and see how we build a model for $\mu_z(x)$. There are two general categories of models: S(ingle) learners and T(wo) learners. S learners put a single model on $\mu_z(x)$, but T learners lay two separate models on $\mu_1(x)$ and $\mu_2(x)$. They are the same only under the linear models. Flexible models such as BART(Hill, 2011), Bayesian random forest(Hahn et al., 2020), Gaussian Process(Ray and van der Vaart, 2020) and DP mixture (Karabatsos and Walker, 2012) have been applied.

However, flexible models may output large uncertainty when there is poor overlap, even in low-dimensional cases. Example 4.1 in the original paper indicates that the linear model is too much confident about the area where there's no overlap. And GP has large uncertainty and sacrifices the accuracy of the average treatment effects. BART also has its own issue having the same uncertainty everywhere. Soft decision trees have been developed to solve the BART problem.

Another problem for datasets with high dimensional covariates is that sparsity-inducing prior acts as an informative prior for the model(Linero, 2021). Many Bayesian regularization priors would induce a high spike at zero for the selection bias $\delta_z = \mathbb{E}(Y_i \mid Z_i = z) - \mathbb{E}(Y_i(z))$. When we have a large spike at 0 for δ_z, this implies that we could ignore the covariates X_i and directly estimate $\mathbb{E}(Y_i(z))$ by estimating $\mathbb{E}(Y_i \mid Z_i = z)$.

5 Propensity score and robustness

Under the Frequentist domain, propensity score plays a central role in causal inference. Nonetheless, we have seen in Section 3 that under the ignorability and prior independence assumptions, the propensity score drops out from the likelihood and thus its value appears to be irrelevant in Bayesian causal inference. Seemingly paradoxical, propensity score is essential in Bayesian causal inference as well, especially in obtaining robust causal effect estimation. The fundamental problem of causal inference is that $Y_i(0)$ and $Y_i(1)$ can never be jointly observed. In other words, the observed data provides no information about the pair-wised correlation $\rho = \text{corr}(Y_i(0), Y_i(1))$. When performing Bayesian inference, we can place a prior on ρ to proceed to the inference. Obviously the result will be sensitive to the choice of prior and outcome model, unless we ensure certain level of overlap and balance in the design stage. The propensity score is essential in ensuring such overlap and balance. Thus, it is the key to reduce the sensitivity to the model specification and to achieve robustness.

Existing literature focuses on incorporating the propensity score in the analysis stage. Three strategies have been proposed, (i) including the propensity score as a covariate in the outcome model (Zigler et al., 2013); (ii) specifying priors of outcome model that are dependent on the propensity score (Wang et al., 2012); (iii) plugging in posterior draws of the propensity score and the potential outcomes into the doubly-robust estimator (Saarela et al., 2016). Despite their success in ensuring robustness, criticism have been raised towards it not been dogmatically Bayesian, or not been a general solution.

Incorporating the propensity score in analysis stage is not the only solution for robust Bayesian causal inference. One can also use propensity score in the design stage to ensure overlap and balance in observational studies. Another approach is to use flexible outcome models, such as Bayesian nonparametric models, adaptively quantify the uncertainty according to the degree of overlap.

6 Sensitivity Analysis

Sensitivity analysis refers to assessing the sensitivity of the results with respect to unmeasured confounding in an observational study. To perform sensitivity analysis, one can either obtain the result over a plausible range of values of the sensitivity parameters, or derive theoretical threshold for the sensitivity parameters that would explain away the observed treatment-outcome association. The identification of the sensitivity parameters depends on the parameterization of confounding. A conventional approach is to involve the unmeasured confounders in the joint model, and the sensitivity parameters are the ones in the probabilistic model regarding the unmeasured confounders. Alternatively, since the unconfoundedness assumption have the mathematically equivalent representation: $\Pr[Y(z) \mid Z=1, X] = \Pr[Y(z) \mid Z=0, X]$, for $z=0,1$, we can model the difference between the distributions $\Pr[Y(z) \mid Z=1, X]$ and $\Pr[Y(z) \mid Z=0, X]$. Then the sensitivity analysis is performed on the parameters that directly influence this difference. Despite numerous successful approaches in the literature about performing sensitivity analysis, a general criticism to it is, to assess untestable unconfoundedness, one make even more untestable assumptions.

7 Complex assignment mechanisms

We have presented Bayesian causal inference framework under the simplest setting of an ignorable treatment at one time point. The basic formulation can be extended to complex assignment mechanisms. We examine two important extensions in this section, targeting on lifting the ignorable assumption (Section 7.1) and generalizing to multiple time points (Section 7.2), respectively.

7.1 Instrumental variable

Instrumental variable (IV) is used in settings where there are unmeasured confounders. To be a valid IV, a variable should occur before a treatment, be independent to the treatment-outcome confoudning, and most importantly, affect the outcome only through its effects on the treatment assignment. Conventionally, IVs are chosen as the variable that is believed to be randomized in nature, or can be thought of as a randomized encouragement to receive treatment. Therefore, it brings a source of exogenous variation that helps identify causal effects. Given a valid IV, one can extract the causal effects of the treatment on an outcome by a two-stage least-squares (2SLS) estimator (Angrist and Pischke, 2009).

IV methods perform causal inference on a track that differs from the one induced by the potential outcome framework, because they are based on a set of assumptions alternative to ignorability. Angrist et al. (1996) connects IV to the potential outcomes framework in the setting of randomized experiments with binary treatment. Because their discussion reflects key ideas in causal inference, we investigate it in detail here.

For unit i, let Z_i be the binary randomly assigned treatment, and W_i be the actual received treatment. Because W_i occurs post-assignment, it has two potential values $W_i(0)$ and $W_i(1)$. The joint potential treatment $\mathcal{U}_i = (W_i(1), W_i(0))$ leads to the principal stratification of the data, resulting in four strata corresponding to the four compliance types: always-takers (at), compliers (co), defiers (df), and never-takers (nt). The comparison between $Y_i(Z_i = 1)$ and $Y_i(Z_i = 0)$ within the stratum with same $\mathcal{U}_i$ have standard subgroup causal interpretations. It is termed the principal causal effects, denoted by $\tau_u = \mathbb{E}[Y_i(1) - Y_i(0)|\mathcal{U}_i = u]$. Marginalizing out $\mathcal{U}_i$, we obtain the conventional causal estimand $\mathbb{E}[Y_i(1) - Y_i(0)] = \sum_u \Pr(\mathcal{U}_i = u)\tau_u$, which is termed intention-to-treat effect. Due to the fundamental problem of causal inference, $\mathcal{U}_i$ is not observed. Consequently, the intention-to-treat effect is not identifiable. Angrist et al. (1996) formally stated the two assumptions, namely monotonicity and exclusion restriction, that rule out defiers, never-takers, and always-takers. Then the intention-to-treat effect reduces to the single term $\tau_{\text{co}} = \mathbb{E}[Y(1) - Y(0)|\mathcal{U} = \text{co}]$. It is termed the complier average causal effect, and is identifiable. Because it is the causal effect on the stratum of compliers, it is also the treatment-received effect.

The idea behind Angrist et al. (1996) is natural. The intention-to-treat effect considers the causal effect for four principal strata, so it can be thought as a "global" causal effect. While the complier average causal effect is the causal effect on the stratum $\mathcal{U}_i = \text{co}$ only, which suggests it is a "local" effect. Because of the unmeasured confounders, there is no hope to estimate the global causal effect. Nonetheless, we can estimate the local causal effect. Combined with reasonable assumptions, the local causal effect is enough for the problem we care about.

Bayesian implementation of IV methods resembles that of a mixture model, with $\mathcal{U}_i$ plays the role of a grouping configuration variable. Without further assumptions, the observed (Z, W) combination pattern consist of a mixture of units from more than one stratum. For example, units with observed $(Z_i = 1, W_i = 1)$ can be either always-takers or compliers. As a consequence, to perform Bayesian inference, we need to specify the compliance type model $P(\mathcal{U}_i|\mathbf{X}_i; \theta_{\mathcal{U}})$ and include the compliance type information $\mathcal{U}_i$ in the outcome model.

7.2 Time-varying treatment

Time-varying treatment refers to the setting in which subjects receiving treatments sequentially at multiple time points, and the treatment assignment at each time is affected by both time-varying confounders as well as the previous treatment. Denote the observed and hypothetical treatment sequence of length T by $\mathbf{Z}_{it} = (Z_{i1}, \cdots, Z_{iT})$ and $\mathbf{z}_T = (z_1, \cdots, z_T)$, respectively, and the sequence of time-varying confounders by $\mathbf{X}_{it} = (X_{i0}, X_{i1}, \cdots, X_{iT})$. The final observed outcome is $Y_i = Y_i(\mathbf{Z}_{iT})$. A common causal estimand could be the marginal effect comparing two specific treatment sequence, denoted as $\tau_{\mathbf{z}_T,\mathbf{z}'_T} = \mathbb{E}[Y_i(\mathbf{z}_T) - Y_i(\mathbf{z}'_T)]$. To estimate it, a fully Bayesian approach would specify a joint model for treatment assignment Z_t and time-varying confounders L_t at all time points, as well as all the potential outcomes $Y(\mathbf{z}_T)$. It soon becomes intractable as T and the number of confounders increases.

Two Bayesian approaches have established success in causal inference estimation under a time-varying treatment setting, namely the Bayesian g-formula (referred as "BGF" hereinafter) and the Bayesin marginal structural model (referrd as "SSMK" hereinafter). The core idea of BGF (Rubin, 1986) is to probabilistically express the time-evolution of all variables, including both the treatment choice arising under a specific intervention and the treatment choice arising in the absence of any intervention. SSMK (Saarela et al., 2015) reconcile Bayesian inference with inverse probability of treatment weighting. It devised a Bayesian version of the marginal structural model via the Bayesian bootstrap based on modeling the treatment-choice mechanism. We illustrate these two approaches with the following example due to Gustafson (2015).

Consider the scenario with only two time points. At time point t, there is a binary confounder X_{it} and a binary treatment Z_{it}, $t = 1, 2$. We also observe a binary outcome Y_i after the treatment assigned at T. We simulate $N = 5000$ data from the following process

$$\begin{aligned}
&X_{i1} \sim \text{Bern}(0.25), \quad Z_{i1} \sim \text{Bern}(0.05 + \kappa_1 X_{i1}),\\
&X_{i2} \sim \begin{cases}\text{Bern}(0.25 - \kappa_2 Z_{i1}) & X_{i1} = 0\\ \text{Bern}(0.95) & X_{i1} = 1\end{cases}, \quad Z_{i2} \sim \begin{cases}\text{Bern}(0.1 + \kappa_1 X_{i2}) & Z_{i1} = 0\\ \text{Bern}(0.9) & Z_{i1} = 1\end{cases}\\
&Y_i \sim \text{MaxBern}(0.1 - \kappa_3 Z_{i1} + \kappa_4 X_{i1}, 0.2 - \kappa_3 Z_{i2} + \kappa_4 X_{i2})
\end{aligned}$$

where $\text{MaxBern}(a, b)$ is the maximum of independent $\text{Bern}(a)$ and $\text{Bern}(b)$ random variables. The hyperparameters $\{\kappa_1, \cdots, \kappa_4\}$ are carefully chosen to reflect the typical features. (The specific choice of them are presented in Gustafson (2015).) Let $\theta_{\zeta_1\zeta_2} := \mathbb{E}(Y(Z_1 = \zeta_1, Z_2 = \zeta_2))$ be the marginal expectation of the potential outcome corresponding to treatment sequence $(\zeta_1, \zeta_2) \in \{0, 1\}^2$, then a causal estimand could be $\theta_{\zeta_1\zeta_2} - \theta_{\zeta'_1\zeta'_2}$. Specifically for this example, we estimate $\theta_{11} - \theta_{00}$.

Before we present the detailed steps of BGF and SSMK estimation of the causal effect, we first introduce the Bayesian saturated binary regression tree (referred as BSAT hereinafter), which will served as a key model component in BGF and SSMK. BSAT is a flexible modeling approach for binary regression with binary covariates. For a binary response A and binary covariates $\mathbf{B} \in \{0, 1\}^p$, we place independent $\text{Unif}(0, 1)$ priors on $\lambda_{\mathbf{b}} := \Pr(A = 1|\mathbf{B} = \mathbf{b})$ for all 2^p possible values of $\mathbf{b}$. By conjugacy, the posterior distributions of $\{\lambda_{\mathbf{b}} : \mathbf{b} \in \{0, 1\}^p\}$ are independently beta-distributed, and directly sampling from them is trivial.

For the example at hand, the BGF approach proceed as follows. The outcome probability in the potential outcome world corresponding to the treatment sequence (ζ_1, ζ_2) are

$$\begin{aligned}
\theta_{\zeta_1\zeta_2} = \mathbb{E}(Y(Z_1 = \zeta_1, Z_2 = \zeta_2)) = \sum_{x_1=0}^{1}\sum_{x_2=0}^{1}[&\Pr(X_1 = x_1)\Pr(X_2 = x_2 \mid X_1 = x_1, Z_1 = \zeta_1)\\
&\times \Pr\{Y = 1 \mid \mathbf{X} = (x_1, x_2), \mathbf{Z} = (\zeta_1, \zeta_2)\}].
\end{aligned} \tag{2}$$

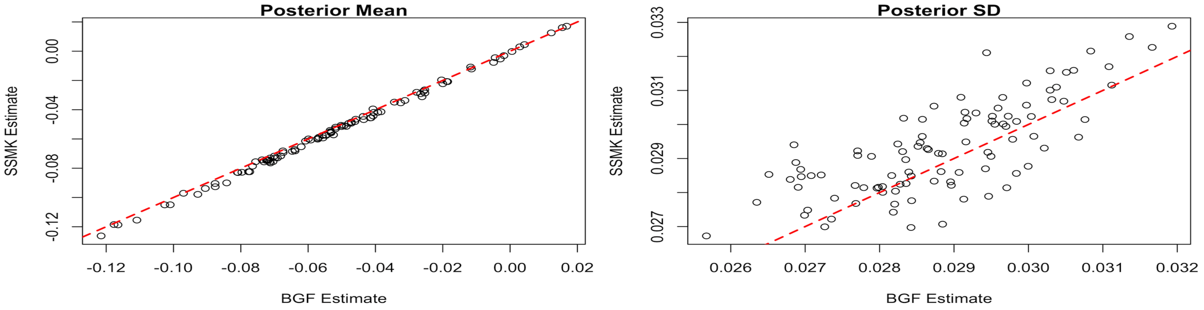

Figure 1: Comparing BGF and SSMK inferences on 100 simulated data sets. The left panel compares posterior means and the right panel compares posterior standard deviations.

We fit BSAT separately to (X_1), $(X_2 \mid X_1, Z_1)$, and $(Y \mid X_1, Z_1, X_2, Z_2)$, obtaining posterior samples of the corresponding model parameters. Combined with (2), we can obtain full posterior inference of $\theta_{11} - \theta_{00}$.

As for SSMK, notice that it mimics the inverse probability weighting estimator, the weight for subject i is given by

$$\omega_i = \frac{P(Z_{i1}; \boldsymbol{\phi}_1) P(Z_{i2} \mid Z_{i1}); \boldsymbol{\phi}_2}{P(Z_{i1} \mid X_{i1}; \boldsymbol{\phi}_3) P(Z_{i2} \mid X_{i1}, Z_{i1}, X_{i2}; \boldsymbol{\phi}_4)},$$

where $\{\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_4\}$ denote the corresponding model parameters. Hence, we first fit BSAT separately to (Z_1), $(Z_2 \mid Z_1)$, $(Z_1 \mid X_1)$, and $(Z_2 \mid X_1, Z_1, X_2)$ separately, and obtain L posterior samples of the model parameters, denoted as $\{(\boldsymbol{\phi}_1^{(\ell)}, \cdots, \boldsymbol{\phi}_4^{(\ell)}) : \ell = 1, \cdots, L\}$. Following (Saarela et al., 2015), the weight for each subject is then estimated by

$$\hat{\omega}_i = \frac{\sum_{\ell=1}^{L} P(Z_{i1}; \boldsymbol{\phi}_1^{(\ell)}) P(Z_{i2} \mid Z_{i1}); \boldsymbol{\phi}_2^{(\ell)}}{\sum_{\ell=1}^{L} P(Z_{i1} \mid X_{i1}; \boldsymbol{\phi}_3^{(\ell)}) P(Z_{i2} \mid X_{i1}, Z_{i1}, X_{i2}; \boldsymbol{\phi}_4^{(\ell)})}.$$

Notice that for different subjects with the same $(\mathbf{X}, \mathbf{Z})$ pattern, they will have the same weights. To obtain L posterior samples of $\theta_{\zeta_1\zeta_2}$, one need to perform Bayesian bootstrap first, and then reweight. Specifically, denote the ℓ-th draw of the bootstrap probability vector by $(\pi_1^{(\ell)}, \cdots, \pi_N^{(\ell)})$, we obtain the posterior sample $\theta_{\zeta_1\zeta_2}^{(\ell)}$ by

$$\theta_{\zeta_1\zeta_2}^{(\ell)} = \frac{\sum_{i=1}^{n} \pi_i^{(\ell)} \hat{\omega}_i I(Y_i = 1)}{\sum_{i=1}^{n} \pi_i^{(\ell)} \hat{\omega}_i}.$$

From here, obtaining posterior inference of $\theta_{11} - \theta_{00}$ is trivial.

Figure 1 presents the comparison of the posterior estimations obtained by BGF and SSMK on 100 simulated data sets. The posterior means are always essentially the same. The corresponding posterior standard deviations agree less closely, through both exhibit modest variation across repeated sampling. In conclusion, although the two methods start with different premises, and require modeling different parts of the joint distribution of observable, they give essentially the same estimate and close indication of uncertainty.

8 Concluding remarks

In this report, we investigate Bayesian causal inference, following the review paper Li et al. (2023). Despite the general virtue of Bayesian inference in obtaining uncertainty quantification, the benefits of using Bayesian methods in causal inference are summarized as follows: (i) we impute all missing potential outcomes, thus allowing straightforward inference of more complicated causal estimand; (ii) uncertainty quantification of any causal estimand can be obtained automatically from the posterior, which can be combined with decision theory for dynamic decision making; (iii) the well-established Bayesian methods for complex data structure make them particularly suitable for causal inference under complex settings; (iv) advanced Bayesian models, such as Bayesian nonparametrics, spatial-temporal models, and Bayesian model selection, enlarge the toolbox for performing causal inference.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**(434), 444–455.

Angrist, J. D. and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Gustafson, P. (2015), 'Discussion of 'on Bayesian estimation of marginal structural models'', *Biometrics* **71**(2), 291–293.

Hahn, P. R., Murray, J. S. and Carvalho, C. M. (2020), 'Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)', *Bayesian Analysis* **15**(3), 965–1056.

Hill, J. L. (2011), 'Bayesian nonparametric modeling for causal inference', *Journal of Computational and Graphical Statistics* **20**(1), 217–240.

Jales, H. and Yu, Z. (2016), 'Identification and estimation using a density discontinuity approach', *Adv. Econom* **38**.

Karabatsos, G. and Walker, S. G. (2012), 'A bayesian nonparametric causal model', *Journal of Statistical Planning and Inference* **142**(4), 925–934.

Li, F., Ding, P. and Mealli, F. (2023), 'Bayesian causal inference: a critical review', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **381**(2247), 20220153.

Linero, A. R. (2021), 'In nonparametric and high-dimensional models, bayesian ignorability is an informative prior', *arXiv preprint arXiv:2111.05137* .

Ray, K. and van der Vaart, A. (2020), 'Semiparametric bayesian causal inference'.

Rosenbaum, P. R. and Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rubin, D. B. (1981), 'The bayesian bootstrap', *The annals of statistics* pp. 130–134.

Rubin, J. (1986), 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect', *Mathematical Modelling* **7**, 1393–1512.

Saarela, O., Belzile, L. R. and Stephens, D. A. (2016), 'A bayesian view of doubly robust causal inference', *Biometrika* **103**(3), 667–681.

Saarela, O., Stephens, D. A., Moodie, E. E. M. and Klein, M. B. (2015), 'On Bayesian estimation of marginal structural models', *Biometrics* **71**(2), 279–288.

Wang, C., Parmigiani, G. and Dominici, F. (2012), 'Bayesian effect estimation accounting for adjustment uncertainty', *Biometrics* **68**(3), 661–671.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A. and Dominici, F. (2013), 'Model feedback in bayesian propensity score estimation', *Biometrics* **69**(1), 263–273.