

# Variable selection and estimation in causal inference using Bayesian spike and slab priors

Journal Title  
XX(X):2–27  
©The Author(s) 2018  
Reprints and permission:  
[sagepub.co.uk/journalsPermissions.nav](http://sagepub.co.uk/journalsPermissions.nav)  
DOI: 10.1177/ToBeAssigned  
[www.sagepub.com/](http://www.sagepub.com/)  


**Brandon Koch<sup>1</sup>, David M. Vock<sup>2</sup>, Julian Wolfson<sup>2</sup>, Laura Boehm Vock<sup>3</sup>**

## Abstract

Unbiased estimation of causal effects with observational data requires adjustment for confounding variables that are related to both the outcome and treatment assignment. Standard variable selection techniques aim to maximize predictive ability of the outcome model, but they ignore covariate associations with treatment and may not adjust for important confounders weakly associated to outcome. We propose two methods for estimating causal effects that simultaneously consider models for both outcome and treatment using a Bayesian formulation with spike and slab priors on each covariate coefficient; the Spike and Slab Causal Estimator (SSCE) aims to achieve minimum bias of the causal effect estimator while the Bilevel SSCE (BSSCE) aims to minimize its mean squared error (MSE). Simulations show SSCE can greatly reduce bias over a similar approach that does not consider a model for treatment, while BSSCE can substantially reduce MSE over SSCE and a recent popular approach. Simulations also show SSCE and BSSCE perform well with large numbers of covariates, including situations where the number of covariates is greater than the sample size. We illustrate SSCE and BSSCE by estimating the causal effect of vasoactive therapy versus fluid resuscitation on hypotensive episode length for patients in the MIMIC-III critical care database.

---

## **Keywords**

Bayesian methods, Causal inference, High dimensional data, Spike and slab, Variable selection

## **1 Introduction**

Inferring the causal effect of a treatment, exposure, or intervention (hereafter referred to as “treatment”) on some outcome or response is often the primary goal of a study. Randomizing treatment assignment is the gold standard for estimating causal treatment effects but is unethical, infeasible, or not cost-effective in many situations. When treatment is not randomized, confounding variables – those associated with both treatment and outcome – can induce bias if unaccounted for in the treatment effect estimator.<sup>1</sup> There are many ways to adjust for confounding variables, including modeling the outcome as a function of treatment and covariates or modeling the treatment as a function of covariates (or both).<sup>2,3</sup>

In the absence of unmeasured confounding (admittedly a strong assumption in many cases), an unbiased estimator of the causal treatment effect can be obtained from a model that correctly specifies the conditional mean of the outcome as a function of treatment and covariates.<sup>4</sup> Controlling for all measured covariates prevents confounder omission and, therefore, protects against bias. However, adjusting for covariates that are unrelated to the outcome can increase the variance of the treatment effect estimator without reducing bias, so the “all-inclusive” approach can be suboptimal for estimating treatment effects.<sup>5,6</sup> The potential loss of efficiency is particularly worrisome when many measured covariates are under consideration. One possible approach is to use variable selection and adjust for only the covariates that are related to the outcome (which includes confounding variables). But traditional variable selection techniques, such as the lasso, select covariates based on their associations in only the outcome model and may not select important confounding variables that are weakly related to outcome but strongly

---

<sup>1</sup>School of Community Health Sciences, University of Nevada, Reno, NV, 89557, USA

<sup>2</sup>Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

<sup>3</sup>Department of Mathematics, Computer Science, and Statistics, Gustavus Adolphus College, St. Peter, MN 56082, USA

**Corresponding author:**

Brandon Koch, School of Community Health Sciences, University of Nevada, Reno, NV, 89557, USA.  
Email: bkoch@unr.edu

---

associated with treatment.<sup>7</sup> Using standard variable selection techniques on only the outcome (or treatment) model can, therefore, bias the treatment effect estimator.<sup>8</sup>

Employing Bayesian methods for causal treatment effect estimation has become quite popular. Bayesian model averaging (BMA) proposes taking a weighted average of the effect estimates across models with different covariates included where the weights are determined by the posterior model probability.<sup>9</sup> However, like traditional variable selection, standard BMA tends to prioritize models which include covariates strongly associated with outcome and may assign significant weight to models that include only a subset of the necessary confounders, resulting in biased causal treatment effect estimation.<sup>10</sup> Crainiceanu et al.<sup>10</sup> introduce a two-stage BMA method that forces strong predictors of treatment that are identified in a first stage to be included in the outcome model in a second stage, and then strong predictors of outcome that are not identified in stage one are identified in stage two. Wang et al.<sup>11</sup> propose Bayesian Adjustment for Confounding (BAC), a Bayesian model averaging method on the outcome model with an informative prior obtained separately from the treatment model (see Wang et al.<sup>12</sup> for extensions; see Zigler and Dominici<sup>13</sup> for related Bayesian methods that select variables for propensity score estimation; see Hahn et al.<sup>14</sup> for a recent Bayesian approach that uses a simultaneous regression model to select variables when the treatment is continuous rather than binary). BAC contains a prior dependence parameter,  $\omega$ , ranging from 1 to  $\infty$  that links the treatment model to the outcome model. If  $\omega = \infty$ , all covariates with associations in the treatment model are forced into the outcome model, whereas  $\omega = 1$  treats the two models independently and is equivalent to standard BMA on the outcome model.

Two approaches have been predominantly used to select  $\omega$ : (1) setting  $\omega$  equal to  $\infty$  as the default and (2) selecting  $\omega$  data-adaptively to minimize mean squared error (MSE) or other criterion. As described below, each approach has limitations.

Setting  $\omega = \infty$  targets the set of covariates associated with treatment or outcome. However, BAC with  $\omega = \infty$  can have high inclusion probabilities for irrelevant covariates that are unrelated to outcome and treatment, particularly in smaller sample sizes. This can lead to inefficient estimation in moderate sample sizes compared to other variable selection approaches which target the same set of covariates (i.e., all variables associated with treatment or outcome). Furthermore, for a given sample size, selecting all covariates related to treatment and outcome may not be the set which leads to the most efficient estimator of the average causal effect. In Section 3.5.1 of this paper, the variance and bias of the treatment effect estimator is derived, and we find

that adjusting for covariates that are strongly related to treatment but unrelated (or very weakly related) to outcome can increase the asymptotic variance of the treatment effect estimator without substantially reducing its asymptotic bias. Consequently, to minimize MSE of the treatment effect estimator, it may be necessary to use models which do not include covariates related only to treatment (or those related to treatment but are very weakly related to the outcome) when estimating the treatment effect.

However, developing a data-adaptive approach to select  $\omega$  to minimize MSE has proved challenging. Lefebvre et al.<sup>15</sup> proposed using cross validation or the bootstrap to choose  $\omega$  with the aim of minimizing MSE of the treatment effect estimator, but they found that the performance of these procedures was sensitive to the underlying data generating mechanism and suggested that alternative approaches should be investigated. Even if cross validation or the bootstrap could be reliably used to choose  $\omega$ , such methods can be computationally intensive with large datasets. Further, BAC requires calculating the Bayesian Information Criterion at each posterior draw, which cannot be calculated when the number of potential confounders in the model exceeds the sample size.

In this paper, we propose the Spike and Slab Causal Estimator (SSCE) and the Bilevel SSCE (BSSCE), novel Bayesian methods that simultaneously consider models for outcome and treatment and use spike and slab priors on the covariate coefficients to encourage variable selection based on associations in both the outcome and treatment models. SSCE aims to minimize treatment effect bias by controlling only for covariates that are related to outcome or treatment (and removing irrelevant ones), while BSSCE adjusts for the subset of the covariates which minimize MSE of the treatment effect estimator. The proposed methods are implemented using fast Gibbs samplers that perform well with a large number of covariates, even when the number of covariates is greater than the sample size.

## 2 Preliminaries

### 2.1 Estimation of causal treatment effects

Suppose an observational study yields outcomes  $\mathbf{Y} = \{Y_1, \dots, Y_n\}$  with corresponding binary treatment indicators  $\mathbf{A} = \{A_1, \dots, A_n\}$  for independent subjects  $1, \dots, n$ , and we are interested in estimating the average causal effect of a binary treatment  $A$  on outcome  $Y$ , defined as

$$\Delta = E\{Y(1) - Y(0)\},$$

where  $Y(1)$  and  $Y(0)$  denote potential outcomes that would be observed from an arbitrary individual in the population if given treatment ( $A = 1$ ) and control ( $A = 0$ ), respectively. Other measures of the causal treatment effect, such as the average effect of treatment on the treated, could also be applied here. If we are willing to assume consistency and strong ignorability,<sup>1</sup> i.e., that measured covariates  $\mathbf{X} = X_1, \dots, X_p$  include all confounding variables so that  $\{Y(0), Y(1)\} \perp A|\mathbf{X}$ , then

$$E\{E(Y|A = 1, \mathbf{X})\} = E[E\{Y(1)|A = 1, \mathbf{X}\}] = E[E\{Y(1)|\mathbf{X}\}] = E\{Y(1)\},$$

and similarly,  $E\{E(Y|A = 0, \mathbf{X})\} = E\{Y(0)\}$ . This implies  $\Delta = E\{E(Y|A = 1, \mathbf{X}) - E(Y|A = 0, \mathbf{X})\}$ , so a correctly specified model for  $E(Y|A, \mathbf{X})$  can be used to consistently estimate  $\Delta$ . If we assume a linear regression model for  $E(Y|A, \mathbf{X})$ ,

$$Y_i|\mathbf{X}_i, A_i, \beta, \sigma^2 \sim N\{\mu(\mathbf{X}_i, A_i, \beta), \sigma^2\}, i = 1, \dots, n, \quad (1)$$

where

$$\mu(\mathbf{X}_i, A_i; \beta) = \beta_0 + \beta_A A_i + \beta^T \mathbf{X}_i,$$

a natural estimator for  $\Delta$  is

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^n \{\mu(\mathbf{X}_i, A_i = 1; \hat{\beta}) - \mu(\mathbf{X}_i, A_i = 0; \hat{\beta})\} = \hat{\beta}_A.$$

Since adjusting for measured covariates that are unrelated to  $E(Y|A, \mathbf{X})$  can increase the variability of the treatment effect estimator without reducing its bias, some type of dimension reduction to select covariates associated with  $E(Y|A, \mathbf{X})$  is necessary to achieve unbiased, efficient estimation of  $\beta_A$ . Throughout, we assume covariates  $\mathbf{X}$  are standardized to have marginal mean zero and unit variance so that covariate selection is invariant to the scale of the covariates.

## 2.2 Bayesian spike and slab lasso (BSSL)

The lasso<sup>16</sup> is a popular variable selection technique that can perform well when  $p$  is large, possibly even greater than  $n$ . Though Bayesian versions of the lasso also exist that shrink covariate coefficients toward zero, they do not yield posterior estimates that are exactly zero (see Park and Casella<sup>17</sup> and Kyung et al.<sup>18</sup> for more on the Bayesian lasso). Mitchell and Beauchamp<sup>19</sup> used zero inflated mixture priors for variable selection in a Bayesian framework, and George and McCulloch<sup>20</sup> used a hierarchical formulation

containing zero inflated normal mixture priors to perform variable selection in a linear regression model. Zhang et al.<sup>21</sup> introduced a mixture prior with a point mass at zero for simultaneous selection of grouped variables and variables within a group.

Taking a similar approach, Xu and Ghosh<sup>22</sup> used spike and slab priors on the model coefficients to propose the *Bayesian Spike and Slab Lasso (BSSL)*, a hierarchical formulation of the Bayesian lasso that allows draws from the posterior of each covariate coefficient to be exactly zero. Particularly, the prior on  $\beta$  is

$$\beta_j | \sigma^2, \tau_j^2 \stackrel{ind.}{\sim} (1 - \pi_0)N(0, \sigma^2 \tau_j^2) + \pi_0 \delta_0(\beta_j), j = 1, \dots, p,$$

where  $\delta_0()$  is a point mass at zero and  $\pi_0$  is the prior probability that a covariate coefficient is zero; conjugate priors are used for  $\sigma^2$  (Inverse Gamma),  $\pi_0$  (Beta), and  $\tau_j^2$  (Gamma), which control the amount of shrinkage of the covariate coefficients.

BSSL can be used to select variables for  $\mu(\mathbf{X}, A; \beta)$  and simultaneously estimate  $\beta_A$ , but it ignores the relationship between covariates and treatment. Even though BSSL (or a similar procedure based solely on the outcome model) may asymptotically adjust for all covariates that are truly related to the outcome (and therefore include all confounders), in finite samples it may not select important confounding variables that are weakly related to the outcome, even if they are strongly associated with the treatment. The next section proposes a novel framework that aims to reduce bias in finite samples by encouraging selection of covariates that are associated with treatment assignment.

### 3 Spike and slab causal estimation methodology

#### 3.1 Simultaneous modeling of outcome and treatment

We begin by specifying a probit model for the conditional probability that subject  $i$  receives treatment ( $A_i = 1$ ),

$$P(A_i = 1 | \mathbf{X}_i) = \Phi(\gamma_0 + \gamma^T \mathbf{X}_i),$$

where  $\Phi()$  is the cumulative distribution function of a standard normal random variable. An equivalent formulation of the probit model states that there exists  $A_i^*$  such that  $A_i = 1$  if  $A_i^* > 0$  and  $A_i = 0$  if  $A_i^* \leq 0$ , where  $A_i^*$  is an unobserved (latent) variable that is normally distributed with mean  $\gamma_0 + \gamma^T \mathbf{X}_i$  and unit variance. This equivalent formulation is key, as it allows us to model treatment as a continuous variable and hence apply BSSL to both the treatment and outcome models.

To select covariates based on their associations with both outcome and treatment assignment, we use an approach similar to Koch et al.<sup>23</sup> We first define a new vector  $\mathbf{O} = (\mathbf{Y}, \mathbf{A}^*)^T$  by stacking the outcomes and latent treatment variables, and the corresponding design matrix

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X}_{out} & 0 \\ 0 & \mathbf{X}_{trt} \end{pmatrix},$$

consisting of the design matrices  $\mathbf{X}_{out} = \{\mathbf{X}_1, \dots, \mathbf{X}_p, \mathbf{1}_n, \mathbf{A}\}$  for the outcome model and  $\mathbf{X}_{trt} = \{\mathbf{X}_1, \dots, \mathbf{X}_p, \mathbf{1}_n\}$  for the treatment model. We note that  $\mathbf{Z}$  contains two columns that are associated with each covariate, a crucial fact for the joint selection technique we introduce in the next section. With this notation, the likelihood can be written as

$$\mathbf{O}|\mathbf{Z}, \alpha, \sigma^2 \sim N_{2n}(\mathbf{Z}\alpha, \Sigma_O), \quad (2)$$

with  $\alpha = (\beta_1, \dots, \beta_p, \beta_0, \beta_A, \gamma_1, \dots, \gamma_p, \gamma_0)^T$  and  $\Sigma_O$  a  $2n \times 2n$  diagonal matrix with  $\sigma^2$  as the first  $n$  elements of the diagonal and 1 as the last  $n$  elements of the diagonal.

The variable selection techniques which we propose use a modified Bayesian group lasso approach. The group lasso is an extension of the lasso to allow selection of predetermined groups of variables.<sup>24</sup> For example, if  $X_i$  and  $X_j$  were related in such a way that we would want to include or exclude both variables simultaneously in a model, we could use the group lasso and group together the regression coefficients that correspond to  $X_i$  and  $X_j$  ( $\beta_i$  and  $\beta_j$ ); then after grouping the other  $p - 2$  coefficients in the model so that all  $p$  coefficients belong to exactly one group, the group lasso will force all coefficients in a group to be either all zero or all nonzero, meaning  $\beta_i$  and  $\beta_j$  will simultaneously be zero or nonzero. If all groups are of size one, the group lasso and lasso estimators are equivalent.

For our problem, we have two models (one for outcome and one for treatment assignment) and want each covariate to be included or excluded simultaneously from both models (i.e., want  $\beta_j$  and  $\gamma_j$  to be both zero or both nonzero). That is, we want the coefficients in the model for  $\mathbf{O}|\mathbf{Z}, \alpha, \sigma^2$  corresponding to the covariates  $(\mathbf{X}_j^T, \mathbf{0})$  and  $(\mathbf{0}, \mathbf{X}_j^T)$  to be included or excluded simultaneously. We, therefore, use the idea of the group lasso to form  $p$  groups of size 2, with each group  $k$  containing the outcome and treatment model coefficients corresponding to covariate  $k$  (as in Koch, Vock, & Wolfson (2018)):

$$\text{Group 1} = \{\beta_1, \gamma_1\}, \text{Group 2} = \{\beta_2, \gamma_2\}, \dots, \text{Group } p = \{\beta_p, \gamma_p\}.$$

### 3.2 Spike and slab causal estimator (SSCE)

Using a similar idea to the Bayesian spike and slab group lasso proposed by Xu and Ghosh <sup>22</sup> (an extension of BSSL), we propose the following prior with the likelihood in (2) for  $j = 1, \dots, p$ :

$$\left( \begin{array}{c} \beta_j \\ \gamma_j \end{array} \right) \middle| \sigma^2, \tau_j^2 \sim (1 - \pi_0)N \left\{ \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} \tau_j^2 \sigma^2 & 0 \\ 0 & \tau_j^2 \end{array} \right) \right\} + \pi_0 \delta_0 \left\{ \left( \begin{array}{c} \beta_j \\ \gamma_j \end{array} \right) \right\}.$$

As in Xu and Ghosh <sup>22</sup>, a conjugate Inverse Gamma(0.1, 0.1) prior is used for  $\sigma^2$ , and a non-informative conjugate Beta(1, 1) prior is used for  $\pi_0$ , the prior probability that  $(\beta_j, \gamma_j)^T = (0, 0)^T$ . Flat priors are used for  $\beta_0$ ,  $\beta_A$ , and  $\gamma_0$  since we do not want to shrink these coefficients toward zero. Similar to Xu and Ghosh <sup>22</sup>, independent conjugate Gamma priors are used for  $\tau_j^2$ ,

$$\tau_j^2 \sim \text{Gamma} \left( \frac{3}{2}, \frac{\lambda^2}{2} \right), j = 1, \dots, p,$$

where  $\lambda^2$  is estimated with a Monte Carlo EM algorithm. <sup>17,25</sup>

To implement this model, which we call the Spike and Slab Causal Estimator (SSCE), a fast Gibbs sampler is used to generate samples from the posterior distribution of the treatment effect  $\beta_A$  and other parameters (after standardizing covariates); all priors are conjugate, so full conditionals are easily derived and implemented. The full conditionals used in the Gibbs samplers are provided in the Appendix, and R code for implementation of SSCE (and BSSCE, which is proposed in Section 3.5) is available at <https://github.com/drjkoch/SSCE>.

### 3.3 A motivating example for SSCE

A simplified example is presented here to illustrate the properties and behavior of SSCE. Suppose we have outcome and treatment design matrices  $\mathbf{M}_{out}$  and  $\mathbf{M}_{trt}$  such that  $\mathbf{M}_{out}^T \mathbf{M}_{out} = n\mathbf{I}_n$  and  $\mathbf{M}_{trt}^T \mathbf{M}_{trt} = n\mathbf{I}_n$ , and  $\mathbf{M}_{out_i}^T \mathbf{M}_{out_j} = 0$ , and  $\mathbf{M}_{trt_i}^T \mathbf{M}_{trt_j} = 0$  for  $i \neq j$ , where  $\mathbf{M}_{out_j}$  and  $\mathbf{M}_{trt_j}$  denote the  $j$ th columns in  $\mathbf{M}_{out}$  and  $\mathbf{M}_{trt}$ , respectively. Note that this means there are no columns for intercepts in  $\mathbf{M}_{out}$  and  $\mathbf{M}_{trt}$  and also no column for the treatment effect in  $\mathbf{M}_{out}$  (i.e.,  $\beta_A = 0$ ). When BSSL is used

on only  $\mathbf{M}_{out}$ , the posterior probability that  $\beta_j$  is zero is

$$P(\beta_j = 0 | \mathbf{Y}, \mathbf{M}_{out}) = \frac{\pi_0}{\pi_0 + \frac{1-\pi_0}{(1+n\tau_j^2)^{1/2}} \exp \left\{ \frac{n^2 \tau_j^2}{2\sigma^2(1+n\tau_j^2)} |\hat{\beta}_j^{ls}| \right\}}, \quad (3)$$

where  $\hat{\beta}_j^{ls}$  is the least squares estimator of  $\beta_j$  under a full model. Equation (3) is plotted in black in Figure 1 (for fixed  $\tau_j^2$ ); the colored lines in Figure 1 show the posterior probability that  $\beta_j$  is zero under SSCE (for identical values of the  $\pi_0$ ,  $n$ ,  $\sigma^2$ , and  $\tau_j^2$ ), which is

$$P(\beta_j = 0 | \mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt}) = \frac{\pi_0}{\pi_0 + \frac{1-\pi_0}{1+n\tau_j^2} \exp \left\{ \frac{n^2 \tau_j^2}{2\sigma^2(1+n\tau_j^2)} \sqrt{\hat{\beta}_j^{ls2} + \tilde{\gamma}_j^2} \right\}}, \quad (4)$$

where  $\tilde{\gamma}_j$  is the maximum likelihood estimator of  $\gamma_j$  under a full model.

As  $\tilde{\gamma}_j$  is increased, the posterior probability that  $\beta_j$  is zero decreases and eventually reaches zero, thereby increasing the inclusion probability of confounders in SSCE. Additionally, when covariate  $j$  is irrelevant (i.e.,  $\hat{\beta}_j^{ls}$  and  $\tilde{\gamma}_j$  are near zero), the posterior probability  $\beta_j$  is zero is larger with SSCE, meaning irrelevant covariates will be selected less frequently with SSCE compared to BSSL.

The posterior distribution of  $\beta_j$  for fixed  $\tau_j^2$  under design matrices  $\mathbf{M}_{out}$  and  $\mathbf{M}_{trt}$  is a spike and slab distribution,

$$\begin{aligned} \beta_j | \mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt} &\sim P(\beta_j = 0 | \mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt}) \delta_0(\beta_j) \\ &+ (1 - P(\beta_j = 0 | \mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt})) N \left( \hat{\beta}_j^{ls}, \frac{(1-B)\sigma^2}{n} \right), \end{aligned} \quad (5)$$

where  $P(\beta_j = 0 | \mathbf{O}, \mathbf{M}_{out}, \mathbf{M}_{trt})$  is given in (4) and  $B = \frac{1}{1+n\tau_j^2}$ .

### 3.4 Covariate inclusion probability for SSCE

Here we only assume that the design matrices for the outcome and treatment models,  $\mathbf{X}_{out}$  and  $\mathbf{X}_{trt}$ , have standardized covariates. The conditional probability that  $\beta_j$  (and  $\gamma_j$ ) is zero under SSCE is (with  $\mathbf{Z}$  and  $\mathbf{O}$  defined as in Section 3.1)

$$P((\beta_j, \gamma_j)^T = 0 | \mathbf{Z}, \mathbf{O}, A,$$

$$\sigma^2, \pi_0, \tau_j^2, \beta_0, \beta_A, \beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_p, \gamma_0, \gamma_1, \dots, \gamma_{j-1}, \gamma_{j+1}, \dots, \gamma_p) =$$

$$\frac{\pi_0}{\pi_0 + (1 - \pi_0)\sigma\tau_j^{-2} \left(n - 1 + \frac{1}{\tau_j^2}\right)^{-1} \exp\left\{\frac{\left(n - 1 + \frac{1}{\tau_j^2}\right)^{-1}}{2} \left(\frac{a^2}{\sigma^2} + b^2\right)\right\}},$$

where

$$a = \mathbf{X}_j^T \left( \mathbf{Y} - \beta_0 - \beta_A \mathbf{A} - \sum_{k \neq j}^p \beta_k \mathbf{X}_k \right) \text{ and } b = \mathbf{X}_j^T \left( \mathbf{A}^* - \gamma_0 - \sum_{k \neq j}^p \gamma_k \mathbf{X}_k \right),$$

with  $\mathbf{X}_j$  denoting the column in  $\mathbf{X}$  for the  $j$ th covariate. Note that  $a$  and  $b$  are proportional to the correlation between the  $j$ th covariate and the residual vectors without the  $j$ th covariate in the outcome and treatment models, respectively. Figure 2 plots this probability as a function of  $a$  and  $b$ . The probability  $\beta_j$  is zero decreases as a function of  $\frac{a^2}{\sigma^2} + b^2$ , meaning  $\beta_j$  is less likely to be zero if covariate  $j$  has stronger associations with the residuals in the outcome or treatment model. Thus, in finite samples SSCE is more likely than BSSL to include confounders with weak associations to outcome but strong associations to treatment, meaning it should have less bias than BSSL. Additionally, irrelevant covariates that have small  $\frac{a^2}{\sigma^2} + b^2$  will have small probability of having non-zero coefficients, meaning the variability in the estimator for  $\beta_A$  should be reduced using SSCE compared to a model that adjusts for such irrelevant covariates.

### 3.5 Bilevel spike and slab causal estimator (BSSCE)

**3.5.1 Motivation** Adjusting for covariates that are strongly related to treatment but unrelated to outcome may increase the variance of the treatment effect estimator without reducing its bias. The asymptotic variance of the least squares treatment effect estimator (i.e.,  $\hat{\beta}_A$ ) with covariates  $\mathbf{X}_{-j}$ , where  $\mathbf{X}_{-j}$  is the vector of covariates excluding the  $j$ th covariate, is

$$\text{Var}_{\hat{\beta}_A}(\beta_j, \eta_j; \mathbf{X}_{-j}) = \frac{1}{n} \frac{\sigma_{Y|A, \mathbf{X}_{-j}}^2}{\sigma_{A|\mathbf{X}_{-j}}^2},$$

and with covariates  $\mathbf{X}_{-j}$  and covariate  $X_j$ , the asymptotic variance is

$$\text{Var}_{\hat{\beta}_A}(\beta_j, \eta_j; \mathbf{X}_{-j}, X_j) = \frac{1}{n} \frac{\sigma_{Y|A, \mathbf{X}_{-j}}^2 - \sigma_{X_j|\mathbf{X}_{-j}, A}^2 \beta_j^2}{\sigma_{A|\mathbf{X}_{-j}}^2 - \sigma_{X_j|\mathbf{X}_{-j}}^2 \eta_j^2},$$

where  $\sigma_{L|\mathbf{R}}^2$  denotes the residual variance when fitting a linear model on  $L$  with covariates  $\mathbf{R}$ , and  $\beta_j$  is the true coefficient corresponding to covariate  $X_j$  in the outcome and  $\eta_j$  is

the coefficient corresponding to  $X_j$  from regressing  $A$  on  $\mathbf{X}$  using a linear model. The asymptotic bias of  $\hat{\beta}_A$  with covariates  $\mathbf{X}_{-j}$  (and  $X_j$  excluded) is

$$\text{Bias}_{\hat{\beta}_A}(\beta_j, \eta_j; \mathbf{X}_{-j}) = \beta_j \eta_j \frac{\sigma_{X_j|\mathbf{X}_{-j}}^2}{\sigma_{A|\mathbf{X}_{-j}}^2}.$$

Note that because we assume that we have measured all confounders, the bias with covariates  $\mathbf{X}_{-j}$  and  $X_j$  is 0. Adjusting for covariate  $X_j$  that is unrelated to outcome (i.e., with small  $|\beta_j|$ ) but strongly associated with treatment (i.e., with large  $|\eta_j|$ ) will have a small change in bias (no change if  $\beta_j = 0$ ) but potentially large increase in variance. The MSE of the treatment effect estimator that excludes  $X_j$  is less than the MSE of the estimator that includes  $X_j$  (i.e., the full model) when the square of the bias is less than the change in the variance, or when

$$\Delta\text{MSE}_j = \text{Bias}_{\hat{\beta}_A}(\beta_j, \eta_j; \mathbf{X}_{-j})^2 + \text{Var}_{\hat{\beta}_A}(\beta_j, \eta_j; \mathbf{X}_{-j}) - \text{Var}_{\hat{\beta}_A}(\beta_j, \eta_j; \mathbf{X}_{-j}, X_j)$$

$$= \left( \beta_j \eta_j \frac{\sigma_{X_j|\mathbf{X}_{-j}}^2}{\sigma_{A|\mathbf{X}_{-j}}^2} \right)^2 + \frac{1}{n} \left( \frac{\sigma_{Y|A, \mathbf{X}_{-j}}^2}{\sigma_{A|\mathbf{X}_{-j}}^2} - \frac{\sigma_{Y|A, \mathbf{X}_{-j}}^2 - \sigma_{X_j|\mathbf{X}_{-j}, A}^2 \beta_j^2}{\sigma_{A|\mathbf{X}_{-j}}^2 - \sigma_{X_j|\mathbf{X}_{-j}}^2 \eta_j^2} \right) < 0.$$

Values of  $\Delta\text{MSE}_j$  greater than zero indicate that inclusion of covariate  $X_j$  improves efficiency of the treatment effect estimator.

**3.5.2 Model** To reduce MSE of the treatment effect estimator, we propose BSSCE, which uses an additional layer of sparsity to set covariate coefficients in the outcome model to zero if including such a covariate increases MSE of the treatment effect estimator. Particularly, the following prior is used for  $(\beta_j, \gamma_j)^T$  for  $j = 1, \dots, p$ :

$$\begin{aligned} & \left( \begin{array}{c} \beta_j \\ \gamma_j \end{array} \right) \Bigg| \sigma^2, \tau_j^2, \eta_j, \sigma_{Y|A, \mathbf{X}_{-j}}^2, \sigma_{X_j|\mathbf{X}_{-j}, A}^2, \sigma_{A|\mathbf{X}_{-j}}^2, \sigma_{X_j|\mathbf{X}_{-j}}^2 \sim \\ & (1 - \pi_0)N \left\{ \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} \tau_j^2 \sigma^2 & 0 \\ 0 & \tau_j^2 \end{array} \right) \right\} I(\Delta\text{MSE}_j \geq 0) + \\ & \pi_0 \delta_0 \left\{ \left( \begin{array}{c} \beta_j \\ \gamma_j \end{array} \right) \right\} \{1 - I(\Delta\text{MSE}_j \geq 0)\}, \end{aligned}$$

where

$$\Delta\text{MSE}_j = \left( \beta_j \eta_j \frac{\sigma_{X_j|\mathbf{X}_{-j}}^2}{\sigma_{A|\mathbf{X}_{-j}}^2} \right)^2 + \frac{1}{n} \left( \frac{\sigma_{Y|A,\mathbf{X}_{-j}}^2}{\sigma_{A|\mathbf{X}_{-j}}^2} - \frac{\sigma_{Y|A,\mathbf{X}_{-j}}^2 - \sigma_{X_j|\mathbf{X}_{-j},A}^2 \beta_j^2}{\sigma_{A|\mathbf{X}_{-j}}^2 - \sigma_{X_j|\mathbf{X}_{-j}}^2 \eta_j^2} \right).$$

We use an empirical Bayes approach to estimate the five parameters in  $\Delta\text{MSE}_j$  that do not appear in SSCE ( $\eta_j, \sigma_{Y|A,\mathbf{X}_{-j}}^2, \sigma_{X_j|\mathbf{X}_{-j},A}^2, \sigma_{A|\mathbf{X}_{-j}}^2$ , and  $\sigma_{X_j|\mathbf{X}_{-j}}^2$ ); see Section 7.3 in the Appendix for details. The priors for  $\sigma^2$ ,  $\tau_j^2$ , and  $\pi_0$  remain the same as in SSCE (they are conjugate; see Section 3.2). This prior resembles the prior in SSCE because both priors encourage posterior distributions of  $(\beta_j, \gamma_j)^T$  with large mass at zero if covariate  $j$  is unrelated to both outcome and treatment (i.e., if covariate  $j$  is irrelevant). However, unlike SSCE, this prior encourages posterior distributions of  $(\beta_j, \gamma_j)^T$  with large mass at zero if inclusion of covariate  $j$  would increase the MSE of the treatment effect estimator.

Figure 3 plots four examples of  $N \{(0, 0)^T, \mathbf{I}_2\} I(\Delta\text{MSE}_j \geq 0)$  from the prior for  $(\beta_j, \gamma_j)^T$ . The prior is zeroed out (white area of Figure 3) in places where  $\Delta\text{MSE}_j < 0$ , which occurs when covariate  $j$  has a small outcome but relatively large treatment coefficient magnitude. We call this a bi-level model because it is reminiscent of the bi-level group lasso proposed by Xu and Ghosh<sup>22</sup>, in which selection is first done between and then within groups. At the first level, the spike in the prior ( $\delta_0$ ) removes covariates irrelevant to both treatment and outcome as in SSCE. However, in BSSCE the second “level” of selection specifically reduces MSE. The posterior mass at zero for  $(\beta_j, \gamma_j)^T$  is guaranteed to be larger for BSSCE compared to SSCE.

## 4 Simulations

To evaluate the performance of SSCE and BSSCE, we considered three different data-generating scenarios described in Table 1. In Scenario 1, covariates are independent, and there is a single confounder that is weakly associated with the outcome and strongly related to the treatment ( $X_2$ ). In Scenarios 2 and 3, the covariates are correlated, and there are multiple confounders weakly associated with outcome. There are no covariates related to treatment but unrelated to outcome in Scenario 2, while in Scenario 3, some covariates ( $X_{13}, \dots, X_{16}$ ) are related to treatment but not outcome. We vary the sample size in all scenarios and vary the number of irrelevant covariates in Scenario 1.

We compare SSCE and BSSCE to BSSL and to BAC with  $\omega = \infty$ . Standard R programming was used to implement SSCE, BSSCE, and BSSL, while the R package `bacr` was used to implement BAC. Note that treatment indicators are generated

according to a logistic regression model as assumed in the analysis by BAC but not by the other methods, which instead posit a probit model for treatment assignment.

Simulations are replicated over 500 Monte Carlo (MC) samples, and 5,000 Markov Chain MC samples were used per chain for each method, where the MC standard error for these chain lengths was estimated (using the R package `mcmcse`) to range from approximately 0.005-0.01. To compare variable selection performance, the probability of inclusion for each covariate is calculated as the proportion of posterior draws for which its outcome regression coefficient is non-zero. The MC bias, standard error, and MSE of the treatment effect estimator are calculated for each method using the posterior mean of  $\beta_A$ . 95% credible intervals for the treatment effect are estimated for each method using the 2.5% and 97.5% quantiles of the posterior distribution of  $\beta_A$ . Code to reproduce the simulation is available at <https://github.com/drjkoch/SSCE>.

Table 2 shows the average inclusion probabilities for the first ten covariates in Scenario 1 for different sample sizes ( $n$ ) and total number of covariates ( $p$ ). Using the Bayesian lasso on the outcome (BSSL), the average inclusion probability for the confounder weakly related to outcome ( $X_2$ ) is near zero for all considered combinations of  $n$  and  $p$ , except with the largest ratio of  $n$  to  $p$  ( $n = 500, p = 50$ ), where the inclusion probability for  $X_2$  is 0.19. Using SSCE, however, yields an average inclusion probability close or equal to one for the weak confounder for all considered  $n$  and  $p$ . Average inclusion probabilities of the weak confounder are similar using BAC compared to SSCE. However, the average inclusion probabilities of irrelevant covariates are approximately zero using SSCE, even with twice as many covariates as subjects ( $n = 250, p = 500$ ), and much smaller compared to those of BAC.

By including irrelevant covariates less frequently, SSCE decreases the variability and MSE of the treatment effect estimator compared to BAC for all considered combinations of  $n$  and  $p$  in Scenario 1 except the largest considered ratio of  $n$  to  $p$  ( $n = 500$  and  $p = 50$ ), where the variability and MSE of the treatment effect estimator are similar between BAC and SSCE (see Table 4). SSCE achieves similar or less bias than BAC in all simulations considered, and SSCE also has credible intervals containing the true treatment effect at or above the nominal level in all simulations considered; BAC, for example, has 95% credible intervals covering the true treatment effect in only 90.4% of the simulated datasets in Scenario 1 with  $n = 100$  and  $p = 50$ , while 96.4% of the credible intervals using SSCE cover the true treatment effect.

BSSCE, which aims to reduce MSE of the treatment effect estimator, leads to much smaller inclusion probabilities for covariates related only to treatment in Scenario 1

compared to SSCE and BAC, and slightly smaller inclusion probabilities for irrelevant covariates as SSCE. Though the weak confounder is selected less often with BSSCE than with SSCE and BAC in Scenario 1, BSSCE still achieves similar 95% CI coverage of the treatment effect as SSCE and BAC, and yields a significantly smaller MSE of the treatment effect than all considered methods for all considered ratios of  $n$  to  $p$ .

In Scenario 2, where covariates are correlated and there are numerous weak confounders, the BSSL estimator is biased, as expected, and yields the largest MSE of all methods where  $n = 500$ . BAC displays much larger inclusion probabilities for irrelevant variables and much larger variability in the treatment effect estimator compared to the other methods, and consequently has the largest MSE of all methods with  $n = 100$ . As in Scenario 1, using SSCE yields treatment effect estimates with small bias for all sample sizes in Scenario 2, and the treatment effect estimates have less variability using SSCE compared to BAC with  $n = 100$  and  $n = 250$ . BSSCE obtains the smallest MSE of all considered methods when  $n = 100$  and  $n = 250$ . With  $n = 500$ , the MSEs using SSCE, BSSCE, and BAC are similar, which is expected since there are no covariates strongly related to treatment but unrelated (or very weakly related) to outcome. That is, SSCE, BSSCE, and BAC all target the same set of covariates in Scenario 2, so we would expect that with moderate sample size their performance would be similar.

In Scenario 3, which is similar to Scenario 2 except that there are covariates directly related to treatment but not directly related to outcome, BSSCE yields a significantly smaller MSE than BAC and SSCE with all considered sample sizes. BSSCE excludes those covariates which are only related to treatment (which lead to increased variance of the treatment effect estimator but no difference in bias) with much higher probability compared to SSCE and BAC. Even though BSSCE reduces the MSE by over 30% compared to SSCE and BAC (for example, Scenario 3 with  $n = 250$ ), it still has credible intervals containing the true treatment effect at or above the nominal level in all simulations considered.

## 5 Application

In critical care resolving hypotensive episodes (HEs) in a timely manner is crucial to minimizing end organ damage.<sup>26</sup> The procedures for treating HEs vary, and there is evidence that certain treatments could be associated with a shorter HE duration for patients in critical care.<sup>26</sup> Data on patients treated in intensive care units (ICUs) were obtained from the publicly available Multi-parameter Intelligent Monitoring in Intensive

Care III (MIMIC-III) database to infer the average causal effect of fluid resuscitation compared to vasoactive therapy on HE duration. MIMIC-III contains descriptive de-identified clinical data (demographics, vital signs, laboratory tests, medications, etc.) from approximately 50,000 adult patients admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012.<sup>27</sup>

Patients from the MetaVision data management system were included in the analyses if they experienced a HE, which was defined based on mean arterial pressure (MAP) measurements generally recorded every 10-15 minutes, and received vasoactive therapy or fluid resuscitation. Following Lee et al.<sup>26</sup> the beginning of a HE was defined as the time of two consecutive MAP measurements  $\leq 60$  mm Hg, preceded by two consecutive MAP measurements  $> 60$  mm Hg. The end of a HE was then defined as the first time that two consecutive MAP measurements  $> 60$  mm Hg. Only data on the first HE for each subject was collected so that observations are independent. Vasoactive therapy was defined as an initiation or a dosage increase of dobutamine, dopamine, epinephrine, norepinephrine, phenylephrine or vasopressin during the HE. Fluid resuscitation was defined as at least one infusion of either any volume of colloids or a bolus of isotonic crystalloid.

The potential confounders are described in Table 5. After removing cases with missing covariate values,  $n = 898$  observations remained in the analyses with  $p = 38$  potential confounders. The outcome of interest was time from treatment to the end of the HE (in minutes and log-transformed). Because these data involve human subjects, the data use agreement does not permit the authors to directly release or archive the data but it may be requested from <https://mimic.physionet.org/>; code to reproduce these analyses is available from <https://github.com/drjkoch/SSCE>.

Figure 4 shows the estimated posterior distribution of  $\beta_A$  ( $A = 1$  if vasoactive therapy is used and  $A = 0$  if fluid resuscitation is used) using SSCE, BSSCE, BSSL, and BAC. All methods yield credible intervals (using the 0.025 and 0.975 quantiles of the posterior distribution of  $\beta_A$ ) that contain zero, indicating no significant difference in HE duration for patients receiving vasoactive therapy compared to fluid resuscitation. Using SSCE, the estimated posterior mean of  $\beta_A$  is -0.10 (95% credible interval = (-0.28, 0.07)), while the estimated posterior mean of  $\beta_A$  using BSSCE is also -0.10 (95% credible interval = (-0.28, 0.07)). The estimated posterior mean of  $\beta_A$  using BSSL is -0.14 (95% credible interval = (-0.39, 0.11)), while BAC is similar to SSCE and BSSCE and estimates the posterior mean of  $\beta_A$  to be -0.09 (95% credible interval = (-0.27, 0.10)).

We can see from Table 5 that the potential confounders have larger inclusion probability when using BAC compared to the spike and slab methods, which is consistent with our simulation results. The SAPSII score, which estimates severity of disease, has its outcome coefficient equal to zero in nearly all posterior samples when using BSSL, but has a non-zero coefficient in all of the samples when using SSCE and BAC and in 91% of the samples when using BSSCE (see Table 5). The differences in adjustment for the SAPSII score most likely explain why BSSL suggests a more favorable treatment effect than the other methods.

## 6 Discussion

We have proposed two novel Bayesian methods for variable selection and estimation in causal inference that simultaneously model the outcome and treatment assignment using spike and slab priors on the model coefficients. By simultaneously modeling the outcome and treatment assignment, the proposed methods can identify confounding variables with weak associations in one model that may otherwise be ignored by a procedure using only that model, such as the Bayesian lasso on the outcome model. Furthermore, both approaches show substantial improvements over BAC with  $\omega = \infty$ . SSCE aims to only adjust for covariates related to outcome or treatment assignment in order to minimize bias. Additionally, our proposed method, SSCE, very infrequently includes irrelevant covariates that are unrelated to outcome and treatment assignment, which can greatly reduce variability and the MSE of the treatment effect estimator in finite samples compared to competing approaches such as BAC. On the other hand, BSSCE aims to reduce MSE of the treatment effect estimator by setting coefficients to zero for those covariates which if included in the outcome model would otherwise increase MSE of the treatment effect estimator. BSSCE provides an effective means of reducing MSE without having to data-adaptively choose a tuning parameter like  $\omega$  in BAC. Furthermore, unlike BAC or other model averaging techniques, the proposed approaches performed well even when the number of covariates exceeded the sample size.

Another advantage of our approach is that it allows the variability of all parameters – including both the treatment effect and covariate inclusion probabilities – to be summarized from their posterior distributions. When using non-Bayesian methods such as the standard lasso or simultaneous variable selection technique proposed by Koch et al.<sup>23</sup>, the variance of parameter estimators is often difficult to estimate; with many covariates or subjects, resampling methods can be computationally intensive as

parameters must be re-estimated many times, and in some cases the bootstrap is not guaranteed to consistently estimate the relevant limiting distributions.<sup>28</sup>

A linear model was assumed throughout the paper, but the proposed method could be extended to more complicated models with covariate-covariate interactions, polynomial transformations of covariates, covariate-treatment interactions, or smoothers of various kinds (e.g., penalized splines expressed as mixed linear models). However, additional work needs to be done for such extensions as the ideal grouping structure for interactions and transformations is unclear.

## References

1. Rosenbaum P and Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
2. Funk MJ, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects. *Am J Epidemiol* 2011; **173**: 761–767.
3. Rubin D. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; **127**: 757–763.
4. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modeling* 1986; **7**: 1393–1512.
5. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol* 2008; **167**: 523–531.
6. Schisterman E, Cole S and Platt R. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 2009; **20**: 488–495.
7. Vansteelandt S, Bekaert M and Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res* 2012; **21**: 7–30.
8. Ertefaie A, Asgharian M and Stephens DA. Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference* 2018; **6**.
9. Raftery AE, Madigan D and Hoeting JA. Bayesian model averaging for linear regression models. *J Am Stat Assoc* 1997; **92**: 179–191.
10. Crainiceanu CM, Dominici F and Parmigiani G. Adjustment uncertainty in effect estimation. *Biometrika* 2008; **95**: 635–651.
11. Wang C, Parmigiani G and Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* 2012; **68**: 661–671.

12. Wang C, Dominici F, Parmigiani G, et al. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* 2015; **71**: 654–665.
13. Zigler CM and Dominici F. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *J Am Stat Assoc* 2014; **109**: 95–107.
14. Hahn PR, Carvalho CC, Puelz D, et al. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal* 2018; **13**: 163–182.
15. Lefebvre G, Atherton J and Talbot D. The effect of the prior distribution in the Bayesian Adjustment for Confounding algorithm. *Comput Stat Data Anal* 2014; **70**: 227–240.
16. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc: Ser B (Stat Methodol)* 1996; **58**: 267–288.
17. Park T and Casella G. The Bayesian lasso. *J Am Stat Assoc* 2008; **103**: 681–686.
18. Kyung M, Gill J, Ghosh M, et al. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal* 2010; **5**: 369–411.
19. Mitchell TJ and Beauchamp JJ. Bayesian variable selection in linear regression. *J Am Stat Assoc* 1988; **83**: 1023–1032.
20. George E and McCulloch R. Approaches for Bayesian variable selection. *Stat Sin* 1997; **7**: 339–374.
21. Zhang L, Baladandayuthapani V, Mallick BK, et al. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J Roy Stat Soc Ser C (Appl Stat)* 2014; **63**: 595–620.
22. Xu X and Ghosh M. Bayesian variable selection and estimation for group lasso. *Bayesian Anal* 2015; **10**: 909–936.
23. Koch B, Vock DM and Wolfson, J. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics* 2018; **74**: 8–17.
24. Yuan M and Lin Y. Model selection and estimation in regression with grouped variables. *J Roy Stat Soc: Ser B (Stat Methodol)* 2006; **68**: 49–67.
25. Casella G. Empirical Bayes Gibbs sampling. *Biostatistics* 2001; **2**: 485–500.
26. Lee J, Kothari R, Ladapo JA, et al. Interrogating a clinical database to study treatment of hypotension in the critically ill. *BMJ Open* 2012; **2**: e000916.
27. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; **3**: Article number 160035.
28. Chatterjee A and Lahiri SN. Bootstrapping lasso estimators. *J Am Stat Assoc* 2011; **106**: 608–625.

29. Hoerl AE and Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**: 55–67.

## 7 Appendix

Here we provide the full Gibbs samplers for SSCE and BSSCE. R code for implementation is available at <https://github.com/drkbkoch/SSCE>.

### 7.1 Gibbs sampler for SSCE

To implement SSCE, we use a Gibbs sampler that draws each parameter from the following conditional distributions:

- $\sigma^2 | \beta, \tau^2 \sim \text{IG} \left( \frac{n + \sum_{j=1}^p I(\beta_j \neq 0)}{2} + 0.1, \frac{(\mathbf{Y} - \mathbf{X}_{out}\beta)^T (\mathbf{Y} - \mathbf{X}_{out}\beta) + \sum_{j=1}^p \frac{\beta_j^2}{\tau_j^2}}{2} + 0.1 \right)$ ,

where  $\text{IG}(a,b)$  denotes the Inverse Gamma distribution with shape a and scale b.

- $\pi_0 | \beta \sim \text{Beta}(1 + p - \sum_{j=1}^p I(\beta_j \neq 0), 1 + \sum_{j=1}^p I(\beta_j \neq 0))$ .

- If  $A_i = 0 : A_i^* | \gamma, \mathbf{X}_i \sim TN(\mathbf{X}_i\gamma, 1)^-$ ; if  $A_i = 1 : A_i^* | \gamma, \mathbf{X}_i \sim TN(\mathbf{X}_i\gamma, 1)^+$ ,

where  $TN(\mathbf{X}_i\gamma, 1)^-$  denotes the Truncated Normal distribution with support  $(-\infty, 0)$  and  $TN(\mathbf{X}_i\gamma, 1)^+$  denotes the Truncated Normal distribution with support  $(0, \infty)$ .

- $\beta_0 | \sigma^2, \beta_1, \dots, \beta_p, \beta_A \sim N \left( \frac{1}{n} \mathbf{1}^T (\mathbf{Y} - \mathbf{X}_{out,1}\beta_1 - \dots - \mathbf{X}_{out,p}\beta_p - \mathbf{A}\beta_A), \frac{\sigma^2}{n} \right)$

- $\gamma_0 | \gamma_1, \dots, \gamma_p \sim N \left( \frac{1}{n} \mathbf{1}^T (\mathbf{A}^* - \mathbf{X}_{trt,1}\gamma_1 - \dots - \mathbf{X}_{trt,p}\gamma_p), \frac{1}{n} \right)$

- $\beta_A | \sigma^2, \beta_0, \beta_1, \dots, \beta_p \sim N \left( \frac{\mathbf{A}^T (\mathbf{Y} - \beta_0 - \mathbf{X}_{out,1}\beta_1 - \dots - \mathbf{X}_{out,p}\beta_p)}{\sum_{i=1}^n A_i}, \frac{\sigma^2}{\sum_{i=1}^n A_i} \right)$

- $\beta_j | \sigma^2, \beta_{-j}, \tau_j^2, \pi_0 \sim (1 - l_j)N \left( \mu_{1j}, \frac{\sigma^2}{(n-1+1/\tau_j^2)} \right) + l_j \delta_0(\beta_j), j=1,\dots,p,$

- $\gamma_j | \gamma_{-j}, \sigma^2, \tau^2, \pi_0 \sim (1 - l_j)N \left( \mu_{2j}, \frac{1}{(n-1+1/\tau_j^2)} \right) + l_j \delta_0(\gamma_j), j=1,\dots,p$ , where

$$\mu_{1j} = \frac{1}{(n-1+1/\tau_j^2)} \mathbf{X}_{out,-j}^T (\mathbf{Y} - \beta_0 - \sum_{k \neq j} \mathbf{X}_{out,k}\beta_k - A\beta_A),$$

$$\mu_{2j} = \frac{1}{(n-1+1/\tau_j^2)} \mathbf{X}_{trt,-j}^T (\mathbf{A}^* - \gamma_0 - \sum_{k \neq j} \mathbf{X}_{trt,k}\gamma_k), \text{ and}$$

$$l_j = \frac{\pi_0}{\pi_0 + (1 - \pi_0) \left( \frac{\tau - 2}{n-1+1/\tau^2} \right) \exp \left( \frac{a_j^2}{2\sigma^2(n-1+1/\tau^2)^{1/2}} + \frac{b_j^2}{2(n-1+1/\tau^2)^{1/2}} \right)}, \text{ where}$$

$$a_j = \mathbf{X}_{out,-j}^T (\mathbf{Y} - \mathbf{X}_{out,-j}\beta_{-j}),$$

$$b_j = \mathbf{X}_{trt,-j}^T (\mathbf{A}^* - \mathbf{X}_{trt,-j}\gamma_{-j})$$

- $1/\tau_j^2 | \beta_j, \gamma_j, \lambda^2 \sim \text{IG} \left( \frac{3}{2}, \frac{\lambda^2}{2} \right)$  if  $\beta_j = 0$ ;

- $1/\tau_j^2 | \beta_j, \gamma_j, \lambda^2 \sim \text{Inverse Gaussian} \left( \text{mean} = \sqrt{\frac{\lambda^2 \sigma^2}{\beta_j^2 + \sigma^2 \gamma_j^2}}, \text{scale} = \frac{1}{\lambda^2} \right)$  if  $\beta_j \neq 0$ .

The kth update in the Monte Carlo EM algorithm for  $\lambda^2$  is

$$\sqrt{\frac{3p}{\sum_{k=1}^p E_{\lambda(k-1)}(\tau_k^2 | \mathbf{O})}}.$$

## 7.2 Gibbs sampler for BSSCE

To implement BSSCE, we use a Gibbs sampler that draws each parameter from its conditional distribution. The conditional distributions of  $\sigma^2$ ,  $\pi_0$ ,  $A^*$ ,  $\beta_0$ ,  $\gamma_0$ ,  $\beta_A$ , and  $\tau_j$ ,  $j = 1, \dots, p$ , remain the same as SSCE (see Section 7.1 above for details). However, the conditional distribution of  $(\beta_j, \gamma_j)^T$ ,  $j = 1, \dots, p$  is different using BSSCE than it is when using SSCE. The conditional distribution of  $(\beta_j, \gamma_j)^T$ ,  $j = 1, \dots, p$  using BSSCE is:

$$\begin{aligned} & \bullet \beta_j | \sigma^2, \beta_{-j}, \tau_j^2, \pi_0 \sim (1 - l_j) \xi_j N \left( \mu_{1j}, \frac{\sigma^2}{(n-1+1/\tau_j^2)} \right) + l_j \delta_0(\beta_j), j=1,\dots,p, \\ & \bullet \gamma_j | \gamma_{-j}, \sigma^2, \tau^2, \pi_0 \sim (1 - l_j) \xi_j N \left( \mu_{2j}, \frac{1}{(n-1+1/\tau^2)} \right) + l_j \delta_0(\gamma_j), j=1,\dots,p, \text{ where} \\ & \mu_{1j} = \frac{1}{(n-1+1/\tau_j^2)} \mathbf{X}_{out,-j}^T (\mathbf{Y} - \beta_0 - \sum_{k \neq j} \mathbf{X}_{out,k} \beta_k - A \beta_A), \\ & \mu_{2j} = \frac{1}{(n-1+1/\tau^2)} \mathbf{X}_{trt,-j}^T (\mathbf{A}^* - \gamma_0 - \sum_{k \neq j} \mathbf{X}_{trt,k} \gamma_k), \text{ and} \\ & l_j = \frac{\pi_0}{\pi_0 + C_j (1 - \pi_0) \left( \frac{\tau^{-2}}{(n-1+1/\tau^2)} \right) \exp \left( \frac{a_j^2}{2\sigma^2(n-1+1/\tau^2)^{1/2}} + \frac{b_j^2}{2(n-1+1/\tau^2)^{1/2}} \right)}, \\ & C_j = 1 - \int_{\beta_j: I(\Delta \text{MSE}_j < 0 | \hat{\eta}_j, \hat{\sigma}_{Y|A, \mathbf{X}_{-j}}, \hat{\sigma}_{X_j|\mathbf{X}_{-j}, A}^2, \hat{\sigma}_{A|\mathbf{X}_{-j}}^2, \hat{\sigma}_{X_j|\mathbf{X}_{-j}}^2)} N \left( \beta_j; \mu_{1j}, \frac{\sigma^2}{n-1+1/\tau_j^2} \right) d\beta_j, \\ & a_j = \mathbf{X}_{out,-j}^T (\mathbf{Y} - \mathbf{X}_{out,-j} \beta_{-j}), \\ & b_j = \mathbf{X}_{trt,-j}^T (\mathbf{A}^* - \mathbf{X}_{trt,-j} \gamma_{-j}), \text{ and} \\ & \xi_j = I \left\{ \left( \beta_j \hat{\eta}_j \frac{\hat{\sigma}_{X_j|\mathbf{X}_{-j}}^2}{\hat{\sigma}_{A|\mathbf{X}_{-j}}^2} \right)^2 + \frac{1}{n} \left( \frac{\hat{\sigma}_{Y|A, \mathbf{X}_{-j}}^2}{\hat{\sigma}_{A|\mathbf{X}_{-j}}^2} - \frac{\hat{\sigma}_{Y|A, \mathbf{X}_{-j}}^2 - \hat{\sigma}_{X_j|\mathbf{X}_{-j}, A}^2 \beta_j^2}{\hat{\sigma}_{A|\mathbf{X}_{-j}}^2 - \hat{\sigma}_{X_j|\mathbf{X}_{-j}}^2 \hat{\eta}_j^2} \right) < 0 \right\}, \end{aligned}$$

where estimation of  $\hat{\eta}_j$ ,  $\hat{\sigma}_{Y|A, \mathbf{X}_{-j}}^2$ ,  $\hat{\sigma}_{X_j|\mathbf{X}_{-j}, A}^2$ ,  $\hat{\sigma}_{A|\mathbf{X}_{-j}}^2$ , and  $\hat{\sigma}_{X_j|\mathbf{X}_{-j}}^2$  is detailed in the next subsection (Section 7.3).

## 7.3 Estimating the prior in BSSCE

An empirical Bayes approach is used to estimate  $\eta_j$ ,  $\sigma_{Y|A, \mathbf{X}_{-j}}^2$ ,  $\sigma_{X_j|\mathbf{X}_{-j}, A}^2$ ,  $\sigma_{A|\mathbf{X}_{-j}}^2$ , and  $\sigma_{X_j|\mathbf{X}_{-j}}^2$ , which are the parameters that appear in the prior for BSSCE but not SSCE (note that  $\beta_j$  is  $\Delta \text{MSE}_j$  in the prior for BSSCE, but  $\beta_j$  is drawn in each iteration of the Gibbs sampler).

Recall from Section 3.5.1 that  $\eta_j$  is the coefficient corresponding to  $X_j$  from regressing  $A$  on  $\mathbf{X}$  using a linear model. Our empirical Bayes estimator of  $\eta_j$  is therefore the  $j$ th element of

$$(\mathbf{X}_{trt}^T \mathbf{X}_{trt} + \lambda_\eta \mathbf{I}_{p+1})^{-1} \mathbf{X}_{trt}^T A,$$

where  $\lambda_\eta = 0$  if  $p+1 \leq n$  (i.e.,  $\hat{\eta}_j$  is the least squares estimator of  $\eta_j$ ) and  $\lambda_\eta$  is small and positive if  $p+1 > n$  (i.e.,  $\hat{\eta}_j$  is the ridge regression<sup>29</sup> estimator of  $\eta_j$ ).

To estimate  $\sigma_{Y|A,\mathbf{X}_{-j}}^2$ , we first obtain the standard lasso<sup>16</sup> estimator in the regression of  $Y$  on  $\mathbf{X}$  and  $A$  (with intercept); call this estimator  $\hat{\xi}^{lasso}$ . Next, for all  $j$  such that  $\hat{\xi}_j^{lasso} = 0$ , we estimate  $\sigma_{Y|A,\mathbf{X}_{-j}}^2$  with

$$\hat{\sigma}_{Y|A,\mathbf{X}_{-j}}^2 =$$

$$\frac{\left( Y - \mathbf{X}\hat{\xi}^{lasso} - \hat{\xi}_0^{lasso} - \hat{\xi}_A^{lasso} A \right)^T \left( Y - \mathbf{X}\hat{\xi}^{lasso} - \hat{\xi}_0^{lasso} - \hat{\xi}_A^{lasso} A \right)}{n - \sum_k I(\hat{\xi}_k^{lasso} \neq 0)}.$$

Then, for all  $j$  such that  $\hat{\xi}_j^{lasso} \neq 0$ , we obtain the lasso estimator (call it  $\hat{\xi}_{-j}^{lasso}$ ) that regresses  $Y$  on  $\mathbf{X}_{-j}$  and  $A$  (with intercept), and estimate  $\sigma_{Y|A,\mathbf{X}_{-j}}^2$  with

$$\hat{\sigma}_{Y|A,\mathbf{X}_{-j}}^2 =$$

$$\frac{\left( Y - \mathbf{X}_{-j}\hat{\xi}_{-j}^{lasso} - \hat{\xi}_{-j}^{lasso} 0 - \hat{\xi}_{-j}^{lasso} A A \right)^T \left( Y - \mathbf{X}_{-j}\hat{\xi}_{-j}^{lasso} - \hat{\xi}_{-j}^{lasso} 0 - \hat{\xi}_{-j}^{lasso} A A \right)}{n - \sum_k I(\hat{\xi}_{-j}^{lasso} k \neq 0)}.$$

A similar technique is used to estimate  $\sigma_{A|\mathbf{X}_{-j}}^2$ , replacing  $Y$  with  $A$  (and excluding  $A$  as a covariate).

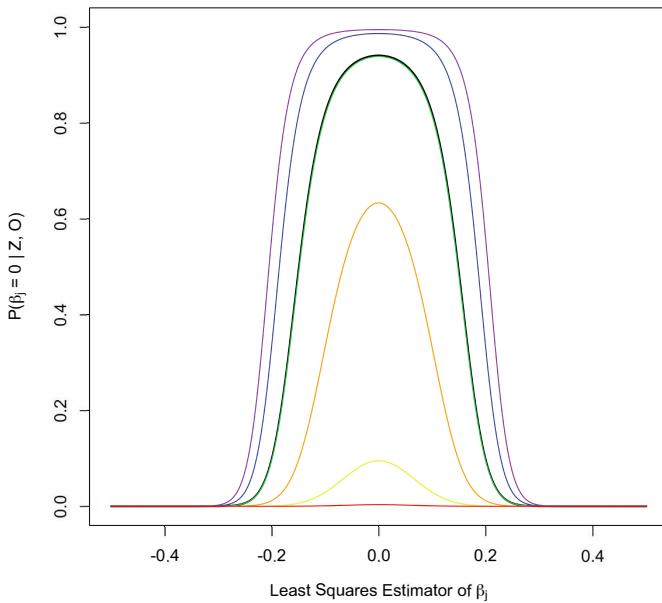
Our estimate of  $\sigma_{X_j|\mathbf{X}_{-j}}^2$  is

$$\hat{\sigma}_{X_j|\mathbf{X}_{-j}}^2 = \frac{\left( X_j - \mathbf{X}_{-j}\hat{\xi}_{x_{-j}}^{lasso} \right)^T \left( X_j - \mathbf{X}_{-j}\hat{\xi}_{x_{-j}}^{lasso} \right)}{n - \sum_k I(\hat{\xi}_{x_{-j}}^{lasso} k \neq 0)},$$

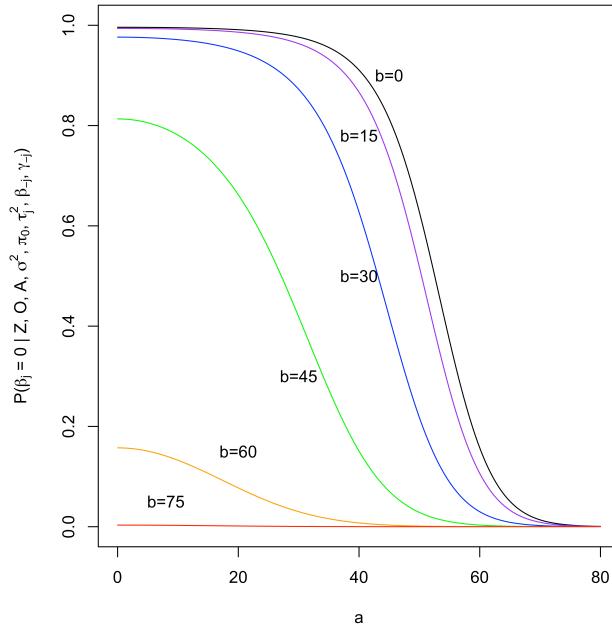
where  $\hat{\xi}_{x_{-j}}^{lasso} k$  is the lasso estimator that is obtained when regressing  $X_j$  on  $\mathbf{X}_{-j}$  (without intercept, since covariates are standardized to have mean zero). We estimate  $\sigma_{X_j|A,\mathbf{X}_{-j}}^2$  with

$$\hat{\sigma}_{X_j|A,\mathbf{X}_{-j}}^2 = \frac{\left( X_j - \mathbf{X}_{-j}\hat{\xi}_{x_{-j},A}^{lasso} - \xi_{x_{-j},A}^{lasso} A \right)^T \left( X_j - \mathbf{X}_{-j}\hat{\xi}_{x_{-j},A}^{lasso} - \xi_{x_{-j},A}^{lasso} A \right)}{n - \sum_k I(\hat{\xi}_{x_{-j},A}^{lasso} k \neq 0)},$$

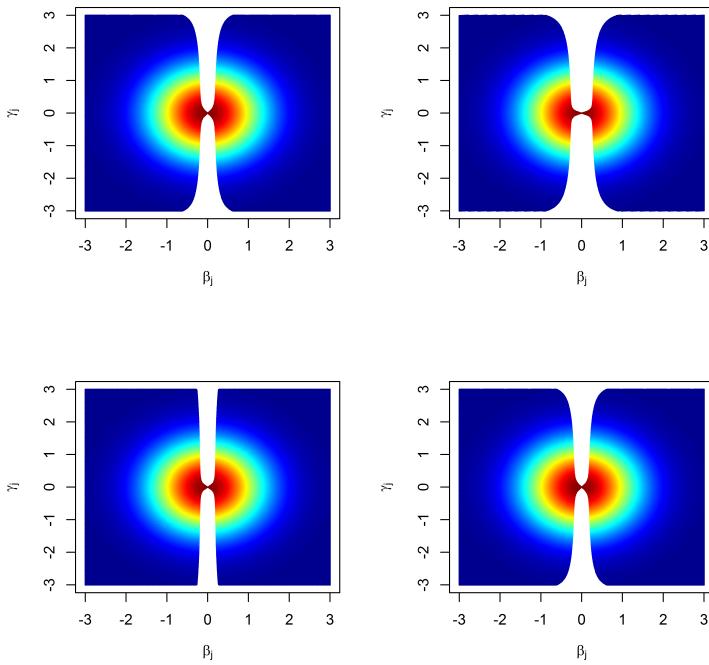
where  $\hat{\xi}_{x_{-j},A}^{lasso} k$  is the lasso estimator that is obtained when regressing  $X_j$  on  $\mathbf{X}_{-j}$  and  $A$ .



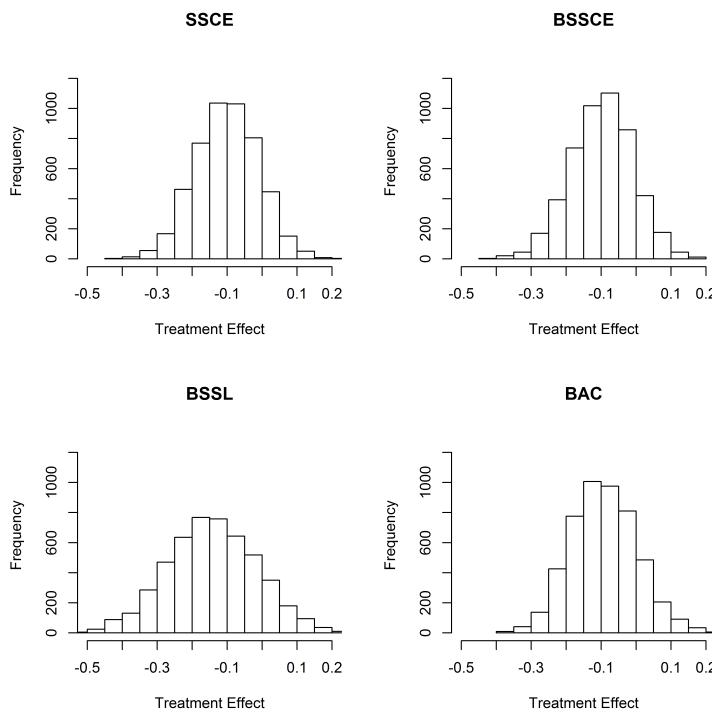
**Figure 1.**  $P(\beta_j = 0 | \mathbf{O}, \mathbf{Z})$  as a function of the least squares estimate of  $\beta_j$  under orthogonal outcome and treatment design matrices. The colors denote the posterior probability  $\beta_j$  is zero using the proposed method for different values of  $\hat{\gamma}_j$ ; purple, blue, green, orange, yellow, and red respectively represent  $\hat{\gamma}_j$  equal to 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30. The black line denotes the posterior probability  $\beta_j$  is zero under BSSL. For this figure,  $n = 250$ ,  $\pi_0 = 0.5$ ,  $\sigma^2 = 1$ , and  $\tau_j^2 = 1$ .



**Figure 2.**  $P(\beta_j = 0 | \mathbf{Z}, \mathbf{O}, A, \sigma^2, \pi_0, \tau_j^2, \beta_{-j}, \gamma_{-j})$  as a function of  $a$  and  $b$  using SSCE, where  $a$  and  $b$  are proportional to the correlation between the  $j$ th covariate and the residual vectors without the  $j$ th covariate in the outcome and treatment models, respectively. For this figure,  $n = 250$ ,  $\pi_0 = 0.5$ ,  $\sigma^2 = 1$ , and  $\tau_j^2 = 1$ .



**Figure 3.** Examples of the slab prior on  $(\beta_j, \gamma_j)^T$  using BSSCE (recall the prior has a spike at zero) as a function of  $\beta_j$  and  $\gamma_j$ . The density of the prior is given by the color: white indicates a prior density of zero, and blue colors indicate a smaller prior density than red colors. The top left plot has  $\sigma_{Y|A,\mathbf{x}_{-j}}^2 = 10$ ,  $\sigma_{X_j|\mathbf{x}_{-j},A}^2 = 0.25$ ,  $\sigma_{A|\mathbf{x}_{-j}}^2 = 10$ , and  $\sigma_{X_j|\mathbf{x}_{-j}}^2 = 1$ . The other three plots double one of these parameters while leaving the other parameters unchanged: the top right plot has  $\sigma_{Y|A,\mathbf{x}_{-j}}^2 = 20$ , the bottom left plot has  $\sigma_{A|\mathbf{x}_{-j}}^2 = 20$ , and the bottom right plot has  $\sigma_{X_j|\mathbf{x}_{-j},A}^2 = 1$ . For this figure,  $n = 250$ ,  $\sigma^2 = 1$ , and  $\tau_j^2 = 1$ .



**Figure 4.** Estimated posterior distributions of  $\beta_A$  (causal effect of fluid resuscitation vs. vasoactive therapy on HE duration in minutes, log-transformed) for each method.

**Table 1.** Covariates  $X_1, \dots, X_p$  are generated with mean  $\mu_x$  and variance  $V_x$ , where  $\text{Cor}(X_i, X_j) = \rho$  for  $i \neq j, i, j \leq 20$ , and  $\text{Cor}(X_i, X_j) = 0$  for  $i \neq j, i, j > 20$ , and treatment indicators and corresponding outcomes are generated from Bernoulli( $\text{expit}(\mu_A)$ ) and Normal( $\mu_Y, V_Y$ ), respectively, where  $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$ , for the following scenarios.

Scenario	$\mu_x$	$V_x$	$\rho$	$\mu_A$	$\mu_Y$	$V_Y$
1	1	4	0	$0.2X_1 - 2X_2 + X_5 - X_6 + X_7 - X_8$	$2X_1 + 0.2X_2 + 5X_3 + 5X_4$	$\frac{4}{4}$
2	0	1	0.5	$-2X_1 - 2X_2 - 2X_3 + X_4 + X_5$ $-0.2X_6 + 2X_7 + X_8 - 0.5X_9 + X_{10}$	$\sum_{j=1}^9 0.2X_j + 0.5X_{10} - 5X_{11} - 5X_{12}$	$1$
3	0	1	0.5	$-2X_1 - 2X_2 - 2X_3 + X_4 + X_5 - 0.2X_6 + 2X_7$ $+X_8 - 0.5X_9 + X_{10} - 2\sum_{j=13}^{16} (-1)^j X_j$	$\sum_{j=1}^9 0.2X_j + 0.5X_{10} - 5X_{11} - 5X_{12}$	$1$

**Table 2.** Covariate inclusion probabilities in Scenario 1 for the first 10 covariates under numerous combinations of  $n$  (sample size) and  $p$  (number of covariates).  $X_1$  is a confounder strongly associated with outcome and weakly associated with treatment;  $X_2$  is a confounder weakly associated with outcome,  $X_5 - X_8$  are only associated with treatment, and  $X_9$  and  $X_{10}$  are irrelevant.

	Confounder X1	Confounder X2	Weak Only Related to Outcome		Only Related to Treatment			Irrelevant		
			X3	X4	X5	X6	X7	X8	X9	X10
<i>n = 100, p = 50</i>										
BSSL	1	0.02	1	1	0.01	0.01	0.01	0.01	0.01	0.01
BAC	1	0.95	1	1	0.83	0.80	0.83	0.82	0.34	0.34
SSCE	1	1	1	1	0.78	0.76	0.78	0.81	0.02	0.02
BSSCE	1	0.53	1	1	0.04	0.04	0.04	0.05	0	0
<i>n = 150, p = 50</i>										
BSSL	1	0.03	1	1	0.01	0.01	0.01	0.01	0.01	0.01
BAC	1	1	1	1	0.89	0.89	0.90	0.91	0.40	0.41
SSCE	1	1	1	1	0.96	0.96	0.96	0.97	0.02	0.01
BSSCE	1	0.73	1	1	0.09	0.07	0.08	0.09	0	0
<i>n = 200, p = 50</i>										
BSSL	1	0.04	1	1	0.01	0.01	0.01	0.01	0	0.01
BAC	1	1	1	1	0.96	0.97	0.93	0.96	0.45	0.46
SSCE	1	1	1	1	1	1	0.99	1	0.01	0.01
BSSCE	1	0.81	1	1	0.13	0.12	0.13	0.12	0	0
<i>n = 250, p = 50</i>										
BSSL	1	0.05	1	1	0.01	0	0.01	0	0	0
BAC	1	1	1	1	1	1	0.99	1	0.51	0.50
SSCE	1	1	1	1	1	1	1	1	0.01	0.01
BSSCE	1	0.87	1	1	0.24	0.23	0.26	0.24	0	0
<i>n = 500, p = 50</i>										
BSSL	1	0.19	1	1	0	0	0	0	0	0
BAC	1	1	1	1	1	1	1	1	0.10	0.12
SSCE	1	1	1	1	1	1	1	1	0.01	0.01
BSSCE	1	0.96	1	1	0.39	0.37	0.38	0.37	0	0
<i>n = 250, p = 100</i>										
BSSL	1	0.03	1	1	0	0	0	0	0	0
BAC	1	1	1	1	1	1	1	1	0.10	0.12
SSCE	1	1	1	1	1	1	1	1	0.01	0.01
BSSCE	1	0.86	1	1	0.15	0.14	0.16	0.14	0	0
<i>n = 250, p = 250</i>										
BSSL	1	0.02	1	1	0	0	0	0	0	0
BAC	*	*	*	*	*	*	*	*	*	*
SSCE	1	1	1	1	1	1	1	1	0	0
BSSCE	1	0.88	1	1	0.34	0.35	0.36	0.37	0	0
<i>n = 250, p = 500</i>										
BSSL	1	0.01	1	1	0	0	0	0	0	0
BAC	*	*	*	*	*	*	*	*	*	*
SSCE	1	1	1	1	0.99	1	1	1	0	0
BSSCE	1	0.88	1	1	0.36	0.37	0.37	0.38	0	0

\*Too many covariates for BAC

**Table 3.** MC Bias, standard error (SE), MSE, and 95% credible interval (CI) coverage probability for the treatment effect estimators.

	Bias	SE	MSE	95% CI Coverage
<b>Scenario 1</b>				
<i>n</i> = 100, <i>p</i> = 50				
BSSL	-0.420	0.434	0.364	0.972
BAC	-0.059	0.741	0.551	0.904
CTMLE	-0.077	1.632	2.665	0.544
SSCE	-0.011	0.650	0.422	0.958
BSSCE	-0.090	0.536	0.295	0.954
<i>n</i> = 150, <i>p</i> = 50				
BSSL	-0.412	0.362	0.300	0.956
BAC	0.005	0.566	0.320	0.938
CTMLE	-0.211	1.240	1.579	0.568
SSCE	0.001	0.531	0.281	0.952
BSSCE	-0.025	0.439	0.193	0.950
<i>n</i> = 200, <i>p</i> = 50				
BSSL	-0.410	0.303	0.259	0.944
BAC	0.019	0.461	0.213	0.942
CTMLE	-0.085	0.869	0.760	0.608
SSCE	-0.003	0.439	0.192	0.972
BSSCE	-0.019	0.370	0.137	0.968
<i>n</i> = 250, <i>p</i> = 50				
BSSL	-0.398	0.277	0.235	0.916
BAC	0.023	0.421	0.177	0.928
CTMLE	-0.146	0.820	0.692	0.628
SSCE	0.010	0.397	0.157	0.966
BSSCE	-0.006	0.346	0.120	0.960
<i>n</i> = 500, <i>p</i> = 50				
BSSL	-0.325	0.238	0.163	0.864
BAC	0.022	0.292	0.085	0.956
CTMLE	-0.082	0.445	0.205	0.630
SSCE	0.015	0.291	0.085	0.944
BSSCE	0.011	0.269	0.072	0.960
<b>Scenario 2</b>				
<i>n</i> = 100, <i>p</i> = 30				
BSSL	-0.215	0.440	0.239	0.992
BAC	0.002	0.657	0.431	0.934
CTMLE	-0.119	1.018	1.048	0.684
SSCE	-0.002	0.580	0.335	0.942
BSSCE	-0.125	0.473	0.239	0.938
<i>n</i> = 250, <i>p</i> = 30				
BSSL	-0.225	0.285	0.132	0.982
BAC	-0.026	0.363	0.132	0.938
CTMLE	-0.098	0.418	0.184	0.764
SSCE	-0.011	0.354	0.125	0.952
BSSCE	-0.057	0.319	0.105	0.954
<i>n</i> = 500, <i>p</i> = 30				
BSSL	-0.227	0.221	0.100	0.954
BAC	-0.012	0.256	0.066	0.950
CTMLE	-0.089	0.287	0.090	0.774
SSCE	-0.014	0.248	0.061	0.950
BSSCE	0.001	0.250	0.062	0.950
<b>Scenario 3</b>				
<i>n</i> = 100, <i>p</i> = 30				
BSSL	-0.182	0.430	0.218	0.990
BAC	-0.024	0.666	0.443	0.944
CTMLE	-0.121	1.079	1.177	0.706
SSCE	-0.029	0.581	0.338	0.968
BSSCE	-0.120	0.441	0.208	0.944
<i>n</i> = 250, <i>p</i> = 30				
BSSL	-0.162	0.281	0.105	0.986
BAC	0.035	0.399	0.160	0.944
CTMLE	-0.088	0.491	0.248	0.768
SSCE	0.023	0.390	0.152	0.960
BSSCE	-0.049	0.317	0.103	0.940
<i>n</i> = 500, <i>p</i> = 30				
BSSL	-0.142	0.194	0.058	0.986
BAC	0.011	0.268	0.072	0.940
CTMLE	-0.061	0.305	0.097	0.790
SSCE	0.014	0.264	0.070	0.944
BSSCE	0.013	0.228	0.052	0.966

\*Too many covariates for BAC

**Table 4.** MC Bias, standard error (SE), MSE, and 95% credible interval (CI) coverage probability for the treatment effect estimators when the outcome model is misspecified.

	Bias	SE	MSE	95% CI Coverage
<b>Scenario 4</b>				
<i>n</i> = 100				
BSSL	-0.037	1.472	2.164	0.994
BAC	-0.221	1.512	2.331	0.932
CTMLE	-0.023	1.899	3.599	0.790
SSCE	-0.061	1.432	2.050	0.958
BSSCE	-0.053	1.445	2.085	0.950
<i>n</i> = 250				
BSSL	0.038	0.918	0.842	0.992
BAC	-0.125	0.910	0.842	0.936
CTMLE	0.079	1.061	1.130	0.896
SSCE	-0.007	0.899	0.806	0.958
BSSCE	0.008	0.898	0.805	0.952
<i>n</i> = 500				
BSSL	0.041	0.627	0.394	0.996
BAC	-0.043	0.627	0.394	0.944
CTMLE	0.062	0.714	0.513	0.930
SSCE	0.000	0.624	0.389	0.966
BSSCE	-0.005	0.624	0.388	0.965
<b>Scenario 5</b>				
<i>n</i> = 100				
BSSL	-0.030	0.762	0.580	0.992
BAC	-0.080	0.745	0.561	0.916
CTMLE	0.045	0.803	0.646	0.854
SSCE	-0.016	0.739	0.546	0.946
BSSCE	-0.020	0.736	0.540	0.946
<i>n</i> = 250				
BSSL	0.008	0.443	0.196	0.996
BAC	-0.058	0.440	0.197	0.932
CTMLE	0.037	0.457	0.210	0.924
SSCE	0.005	0.440	0.194	0.960
BSSCE	0.005	0.439	0.193	0.952
<i>n</i> = 500				
BSSL	0.004	0.307	0.094	0.998
BAC	-0.040	0.307	0.096	0.954
CTMLE	0.025	0.317	0.101	0.950
SSCE	0.002	0.307	0.094	0.968
BSSCE	0.006	0.308	0.094	0.963

\*Too many covariates for BAC

**Table 5.** Covariates considered in the application. For continuous covariates, the mean and standard deviation (SD) by treatment status is provided, and for categorical covariates, the number observed (N) and percentage (%) by treatment status is given. Inclusion probabilities for SSCE, BSSCE, BSSL, and BAC are also shown.

	Vasoactive therapy Mean/N (SD/%)	Fluid resuscitation Mean/N (SD/%)	Inclusion probabilities			
			SSCE	BSSCE	BAC	BSSL
Mean MAP 3hrs prior to treatment	68.0 (9.6)	66.1 (6.8)	1	1	1	0.99
SAPSII	72.1 (14.1)	64.7 (14.5)	1	0.91	1	0.02
Liver disease	66 (9.4%)	39 (20.0%)	0.83	0.02	0.45	0
MAP at treatment	56.4 (13.3)	58.1 (11.5)	0.81	0.44	0.90	0.05
Medical ICU	79 (11.2%)	7 (3.6%)	0.61	0.01	0.23	0.01
Surgical ICU	209 (29.7%)	68 (34.9%)	0.36	0	0.52	0.01
Coronary Care Unit	93 (13.2%)	28 (14.4%)	0.28	0.04	0.28	0.02
Age	66.7 (13.7)	65.5 (16.0)	0.15	0.01	0.20	0.01
Blood anemias	5 (0.7%)	5 (2.6%)	0.13	0.01	0.13	0
Creatinine value	1.6 (1.5)	1.4 (1.3)	0.12	0.04	0.31	0.01
Rheumatoid arthritis	24 (3.4%)	9 (4.6%)	0.07	0.01	0.18	0.01
Alcohol abuse	53 (7.5%)	24 (12.3%)	0.07	0.01	0.11	0.01
Chromic pulmonary disease	167 (23.8%)	36 (18.5%)	0.07	0	0.08	0.01
Fluid and electrolyte disorders	341 (48.5%)	80 (41.0%)	0.06	0.01	0.06	0
Hypothyroidism	90 (12.8%)	28 (14.4%)	0.06	0.01	0.12	0.01
Sex	428 (60.9%)	107 (54.9%)	0.06	0.01	0.07	0.01
Congestive heart failure	210 (29.9%)	46 (23.6%)	0.04	0.01	0.03	0
Renal failure	33 (4.7%)	7 (3.6%)	0.04	0	0.14	0.01
Other neurological disorders	74 (10.5%)	14 (7.2%)	0.04	0.00	0.03	0
Peripheral vascular disorders	75 (10.7%)	18 (9.2%)	0.03	0	0.07	0.01
Urine output 3hrs prior to treatment	39.6 (693.8)	17.5 (184.0)	0.03	0	0.03	0
Paralysis	11 (1.6%)	4 (2.1%)	0.02	0	0.05	0.01
AIDS	3 (0.4%)	1 (0.5%)	0.02	0	0.02	0.01
Valvular disease	158 (22.5%)	42 (21.5%)	0.02	0	0.01	0.01
Cardiac arrhythmias	284 (40.4%)	74 (37.9%)	0.02	0	0.03	0
Cardiac Surgery Recovery Unit	92 (13.1%)	33 (16.9%)	0.02	0	0.04	0
Solid tumor without metastasis	101 (14.4%)	22 (11.3%)	0.02	0	0.05	0
Pulmonary circulation disorders	71 (10.1%)	24 (12.3%)	0.02	0	0.09	0
Lymphoma	17 (2.4%)	4 (2.1%)	0.02	0	0.06	0.01
Diabetes, complicated	49 (7.0%)	11 (5.6%)	0.02	0	0.08	0.01
Coagulopathy	162 (23.0%)	49 (25.1%)	0.02	0	0.06	0.01
Hypertension, uncomplicated	327 (46.5%)	91 (46.7%)	0.02	0	0.02	0
Peptic ulcer disease excluding bleeding	12 (1.7%)	6 (3.1%)	0.01	0	0.06	0.01
Deficiency anemias	137 (19.5%)	35 (17.9%)	0.01	0	0.04	0
Weight loss	46 (6.5%)	16 (8.2%)	0.01	0	0.02	0
Diabetes, uncomplicated	182 (25.9%)	48 (24.6%)	0.01	0	0.04	0.01
Drug abuse	25 (3.6%)	10 (5.1%)	0.01	0	0.03	0
Psychoses	29 (4.1%)	9 (4.6%)	0.01	0	0.04	0