Highlights

Robust Bayesian causal estimation for causal inference in medical diagnosis

Tathagata Basu, Matthias C. M. Troffaes

- A robust variable selection method for high dimensional causal estimation
- Prior sensitivity analysis of spike and slab group-LASSO.
- We show importance of elicitation in variable selection problem
- We show empirical behaviour of our method and its usefulness in medical diagnosis.

Robust Bayesian causal estimation for causal inference in medical diagnosis

Tathagata Basu[a,*], Matthias C. M. Troffaes[b]

[a] *Civil and Environmental Engineering, University of Strathclyde, 16 Richmond St, Glasgow, G1 1XQ, United Kingdom*
[b] *Department of Mathematical Sciences, Durham University, South Road, Durham, DH13LE, United Kingdom*

Abstract

Causal effect estimation is a critical task in statistical learning that aims to find the causal effect on subjects by identifying causal links between a number of predictor (or, explanatory) variables and the outcome of a treatment. In a regressional framework, we assign a treatment and outcome model to estimate the average causal effect. Additionally, for high dimensional regression problems, variable selection methods are also used to find a subset of predictor variables that maximises the predictive performance of the underlying model for better estimation of the causal effect. In this paper, we propose a different approach. We focus on the variable selection aspects of high dimensional causal estimation problem. We suggest a cautious Bayesian group LASSO framework for variable selection using prior sensitivity analysis. We argue that in some cases, abstaining from selecting (or, rejecting) a predictor is beneficial and we should gather more information to obtain a more decisive result. We also show that for problems with very limited information, expert elicited variable selection can give us a more stable causal effect estimation as it avoids overfitting. Lastly, we carry a comparative study with synthetic dataset and show the applicability of our method in real-life situations.

Keywords: high dimensional regresssion, variable selection, Bayesian analysis, imprecise probability

1. Introduction

Causal inference using observational data is important in many fields, including epidemiology, social science, economics, and many more. Causal

*Corresponding author
Email addresses: tathagatabasumaths@gmail.com (Tathagata Basu), matthias.troffaes@durham.ac.uk (Matthias C. M. Troffaes)

inference concerns estimating the causal effect of predictor variables on an outcome variable, as well as identifying which predictors are causally linked with the outcome. Ideally, randomised trials are the most efficient way to perform this task. However, this is not always practical due to, for instance, ethical concerns, design cost, population size, to name a few. This leaves us with observational studies where data is collected though surveys or record keeping.

Unfortunately, without fully controlled randomised trials and full knowledge of confounders, it is well understood that statistical models are unable to infer causality, as correlation does not imply causation especially in the presence of confounders. Still, it is highly desirable to try to adjust for confounding in our statistical models to the best of our ability. This is termed (perhaps somewhat unfortunately) *causal inference* in the literature [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. This is also the approach that we will follow here, under the disclaimer that whether actual causality can be inferred remains a subject of interpretation and conjecture specific to the situation being studied. In this regard, we also refer to [11] where a detailed discussion on different interpretations of 'causality' in statistics and econometrics can be found

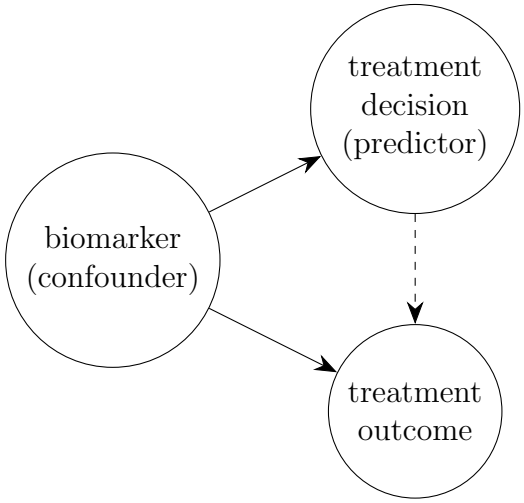

Figure 1: A biomarker influencing both the treatment decision and the treatment outcome, thereby acting as confounder. Solid arrows indicate causation, whilst the dashed arrow indicates correlation without causation.

A confounder is any variable which is causally linked with both a predictor and the outcome, giving the false impression that that predictor causes the outcome (see Fig. 1). Confounding happens commonly in observational treatment studies because many predictors are often causally linked with the treatment decision (which is also a predictor), whilst simultaneously affecting the outcome of the treatment. Any such predictors act as a confounders between treatment decision (as one of the predictor variables) and treatment outcome. In such cases, we must be extra cautious as we risk unwanted bias in the causal effect estimator [2], if we ignore such correlations. Several authors have tackled the presence of confounder variables. Robins [3] used a graphical approach to identify the causal parameters. Rosenbaum and Rubin [12] suggested a link model to estimate the propensity scores for all individuals. Subsequently, several other methods have been proposed based on propensity score matching; see [4, 5] for a brief review.

One of the earlier Bayesian approaches to causal inference can be found in [1]. More recently, with the rise of high dimensional data, Bayesian methodologies have grown in popularity. Crainiceanu et al. [13] proposed a bi-level Bayesian model averaging based method for estimating the causal effect. Wang et al. [7] suggested BAC (or, Bayesian adjustment for confounding), where an informative prior obtained from the treatment model is applied on the outcome model for estimating the causal effect. Several other methods were proposed to tackle confounders, see for instance [6, 9] among others for a Bayesian perspective and [14] for a survey of methods for addressing unmeasured confounding.

In this paper, we take inspiration from the approach of Koch et al. [10], who proposed a bi-level spike and slab prior for causal effect estimation in high dimensional problems (i.e. when number of predictors is larger than the number of observations). They considered a data-driven adaptive approach to propose their prior which reduces the variance of the causal estimate. Our approach however focuses on the other aspect of high dimensional causal inference problem, ie. variable selection. To achieve that we rely on prior sensitivity analysis, where instead of using a single prior, we consider a set of priors [15]. Prior sensitivity analysis for causal inference has been a topic of interest lately. Zaffalon et al. [16] used credal networks in structured causal models for causal inference; Raices Cruz et al. [17] performed a meta analysis in a robust Bayesian framework for causal effect estimation. However, variable selection in causal estimation problem has not been investigated in robust Bayesian framework. This motivates us to investigate the role and ap-

plicability of prior sensitivity analysis in high dimensional causal estimation problems. This is particularly beneficial, as in high dimensional problems, we have to rely on very limited observations to perform our Bayesian analysis and as a result variable selection with a single prior can be unreliable [18] in many cases, Moreover, in causal effect estimation, failing to correctly identify a relation between the treatment effect and predictor can lead to harmful side-effects. Therefore, it is extremely important to adopt a cautious approach in selecting or rejecting a variable. To achieve this cautious paradigm and to perform a prior sensitivity analysis, we rely on expert opinion to elicit a set of priors based on empirical evidence. This allows us to construct the problem of predictor selection in a framework where abstention has a relatively positive gain i.e. when the cost of further tests/data collection is lower than that of incorrectly treating a subject.

Our framework considers a set of continuous spike and slab priors [19] for predictor selection. We thereby construct a Bayesian group LASSO (least absolute shrinkage and selection operator) [20] type problem. To perform sensitivity analysis, we consider a set of beta priors on the covariate selection probability of the spike and slab priors. We use the posteriors of this covariate selection probability for identifying the active predictors. Finally, we consider a post-hoc coefficient adjustment method [21] to recover sparse estimates associated with either the outcome or the treatment model.

The rest of the paper is organised as follows. In Section 2 we give a formal description of the causal estimation problem in the context of linear regression. Section 3 is focused on the Bayesian analysis of causal inference problems, followed by the motivation of a robust Bayesian analysis along with our proposed decision theoretic framework for predictor selection. In Section 4, we provide results of simulation studies under different scenarios and show the possible applications in real life problems. Finally, we discuss our findings and conclude this paper in Section 5.

2. Causal Estimation

Let an observational study on n individuals give us *treatment outcomes* $Y \coloneqq (Y_1, \ldots, Y_n)$ with corresponding *treatment decisions* $T \coloneqq (T_1, \ldots, T_n)$. Here, we use an indicator to represent the treatment decision. That is, T_i is 1 if the ith patient was treated, and 0 otherwise. Similarly, Y_i is the treatment outcome of the ith patient, represented as some real-valued quantity.

Regression methods are widely used in causal effect estimation. The main idea behind these regression methods is to remove the correlation between the treatment indicator and the error term [4, 22]. To do so, we rely on p observed quantities, called *predictors*, denoted by $X \coloneqq [X_1^\top, \ldots, X_n^\top]^\top$ where each $X_i \in \mathbb{R}^p$. Each X_i is treated as a p-dimensional row vector, so X is a $n \times p$ matrix. Now, let $\beta \coloneqq (\beta_1, \ldots, \beta_p)^\top$ denote the vector of regression coefficients related to the predictors, and let β_T denote a regression coefficient related to the treatment decision. Following the usual approach in the literature (see for instance [4, 22]), we model the outcome using a linear model

$$Y_i = T_i\beta_T + X_i\beta + \beta_0 + \epsilon_i \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, independent of T_i and X_i. Note that both T_i and X_i are predictors for Y_i in the above model. However, when we talk about predictors in this paper, we usually mean just the components of X_i.

To decide whether or not to treat a new individual with given predictors, we are mainly interested in the effect of the treatment on the outcome. More precisely, the causal effect of a new individual, indexed as $n + 1$, whose outcome Y_{n+1} is not yet observed, and with observed predictors $X_{n+1} = x_{n+1}$, is defined by:

$$\begin{aligned}\delta(x_{n+1}) \coloneqq \mathbb{E}(Y_{n+1} \mid X_{n+1} = x_{n+1}, T_{n+1} = 1)\\ - \mathbb{E}(Y_{n+1} \mid X_{n+1} = x_{n+1}, T_{n+1} = 0)\end{aligned} \tag{2}$$

For our model, due to linearity of expectation, we have that

$$\begin{aligned}\delta(x_{n+1}) = \beta_T + x_{n+1}\beta + \mathbb{E}(\epsilon_{n+1} \mid X_{n+1} = x_{n+1}, T_{n+1} = 1)\\ - x_{n+1}\beta - \mathbb{E}(\epsilon_{n+1} \mid X_{n+1} = x_{n+1}, T_{n+1} = 0)\end{aligned} \tag{3}$$

and because ϵ_{n+1} is independent from X_{n+1} and T_{n+1},

$$= \beta_T. \tag{4}$$

Note that, for this model, the causal effect $\delta(x_{n+1})$ does not depend on the observed value x_{n+1} of X_{n+1}. So, to find the causal effect, we simply need to estimate β_T. Note that, if interaction terms between X_i and T_i were present in the model (for example a term of the form, say, $T_iX_i\eta$ for some parameter vector η), that would result in a dependence of the causal effect on x_{n+1}.

To estimate β_T from the data X, Y and T, especially in the presence of confounders, we also need to consider the association between the treatment

indicators T and the predictors X. A common choice in the literature is to use a probit link function [4], though other link functions, such as the logit, can also be used [22]. In this way, we can specify the conditional probability that subject i receives the treatment through a linear model. That is, for another vector of regression coefficients $\gamma \coloneqq (\gamma_1, \cdots, \gamma_p)^\top$ we assume

$$P(T_i = 1 \mid X_i) = \Phi(X_i\gamma + \gamma_0) \tag{5}$$

where Φ denotes the cumulative distribution function of a standard normal distribution. The key assumption made here is that there is a monotone relationship between the predictors and the probability of treatment. Here too, interaction terms between the X_i could be added to form more complex models if so desired.

To incorporate this probit link function, we model the T_i as follows [23]:

$$T_i^* = X_i\gamma + \gamma_0 + u_i \tag{6}$$

$$T_i = \mathbb{I}(T_i^* > 0) = \begin{cases} 1 & \text{if } T_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $u_i \sim \mathcal{N}(0, 1)$. With this model, indeed

$$P(T_i = 1 \mid X_i) = P(T_i^* > 0) = P(u_i > -X_i\gamma - \gamma_0) = 1 - P(u_i \leq -X_i\gamma - \gamma_0) \tag{8}$$

$$= 1 - \Phi(-X_i\gamma - \gamma_0) = \Phi(X_i\gamma + \gamma_0). \tag{9}$$

Now, to construct the joint likelihood function, we define an extended output $2n \times 1$ column vector $W \coloneqq \left(\begin{smallmatrix} Y \\ T^* \end{smallmatrix}\right)$ and corresponding $2n \times (2p + 3)$ dimensional design matrix

$$Z \coloneqq \begin{bmatrix} T_1 & X_1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & 0 & 0 \\ T_n & X_n & 1 & 0 & 0 \\ 0 & 0 & 0 & X_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & X_n & 1 \end{bmatrix} = \begin{bmatrix} X_O & 0 \\ 0 & X_T \end{bmatrix} \tag{10}$$

where, $X_O \coloneqq [T, X, \mathbf{1}_n]$ and $X_T \coloneqq [X, \mathbf{1}_n]$. Then, considering the assumption of Gaussian error terms, we have the following likelihood distribution

$$W \mid Z, \beta_T, \beta, \beta_0, \gamma, \gamma_0, \sigma^2 \sim \mathcal{N}\left(Z\nu, \Sigma\right), \tag{11}$$

where $\nu \coloneqq (\beta_T, \beta^\top, \beta_0, \gamma^\top, \gamma_0)^\top$ and

$$\Sigma \coloneqq \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & I_n \end{bmatrix}. \tag{12}$$

3. Robust Bayesian Causal Estimation

The likelihood given by Eq. (11) gives us a foundation for a Bayesian group LASSO [20] type model. In this way, we can look into the posterior selection probability of each predictor. In this section, we formally introduce our proposed methodology for causal estimation and we call it as 'robust Bayesian causal estimation' as we perform a robust Bayesian analysis [15] to achieve a cautious variable selection paradigm. There are several ways to construct spike and slab priors for variable selection. In our method, we consider a continuous type prior [18, 19] for faster posterior computation.

3.1. Hierarchical model

Let π_j denote the prior probability that the j-th predictor is associated with the outcome or the treatment. That is, conceptually,

$$\pi_j \coloneqq P\left((\beta_j, \gamma_j) \neq (0, 0)\right). \tag{13}$$

Practically, we model this by defining the following hierarchical model so that, for $1 \leq j \leq p$,

$$(\beta_j, \gamma_j)^\top \mid \pi_j, \sigma^2 \sim \pi_j \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \tau_1^2 \begin{bmatrix}\sigma^2 & 0\\0 & 1\end{bmatrix}\right) + (1-\pi_j)\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \tau_0^2 \begin{bmatrix}\sigma^2 & 0\\0 & 1\end{bmatrix}\right) \tag{14}$$

$$\beta_T \mid \sigma^2 \sim \mathcal{N}\left(0, \sigma^2\right) \tag{15}$$

$$\beta_0 \mid \sigma^2 \sim \mathcal{N}\left(0, \sigma^2\right) \tag{16}$$

$$\gamma_0 \sim \mathcal{N}\left(0, 1\right) \tag{17}$$

$$\frac{1}{\sigma^2} \sim \text{Gamma}(a, b) \tag{18}$$

$$\pi_j \sim \text{Beta}\left(sq_j, s(1-q_j)\right). \tag{19}$$

In the hierarchical model, we fix sufficiently small τ_0 ($1 \gg \tau_0 > 0$) so that (β_j, γ_j) has its probability mass concentrated around zero. Therefore, this represents the spike component of our prior specification. For the slab component, we consider τ_1 to be large so that $\tau_1 \geq 1$. This allows the prior for

(β_j, γ_j) to be flat beyond the spike component at the origin. We illustrate the components of a bivariate spike and slab prior in Fig. 2 (with fixed $\sigma = 1$). We generate the spike component with $\tau_0 = 0.1$ and the slab component with $\tau_1 = 5$.

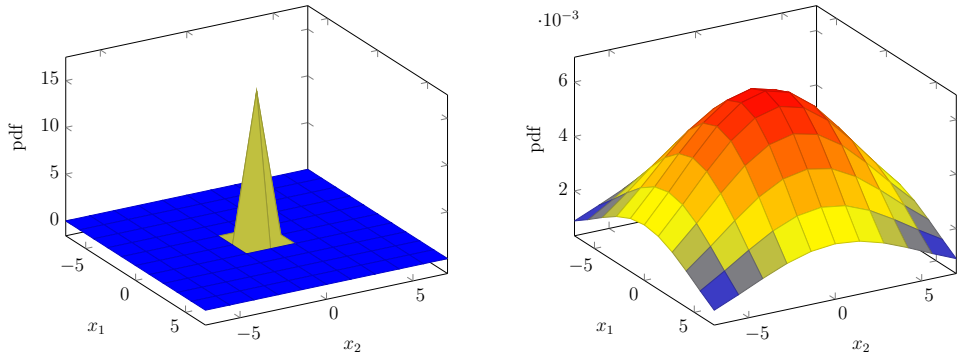

Figure 2: Spike (left) and slab (right) components of a bivariate distribution for $\tau_0 = 0.1$, $\tau_1 = 5$ and $\sigma = 1$.

For the precision term $1/\sigma^2$, a natural choice of prior is the gamma distribution as it allows the control of both the location and the scale of the precision. To ensure that the prior is able to represent the data, we consider $b = 1$ and fix a so that it represents the prior mean of the precision. Alternatively, when $b = 1$, we know that the interval $[0, 3a]$ contains the true value of the precision parameter with probability close to 0.95. So, we can also use a prior judgement on the 95% quantile to set a. We use a beta prior to model our uncertainty about the selection probabilities π_j where q_j represents our prior expectation of π_j and s acts as a concentration parameter. For the causal effect in Eq. (15) and intercept term of the outcome model in Eq. (16), we want to use a Gaussian distribution that matches the scale of the noise term. Therefore, we consider $\beta_T, \beta_0 \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2)$. Similarly, for the intercept of the treatment model we match the scale of the probit model and consider $\gamma_0 \sim \mathcal{N}(0, 1)$.

In Fig. 3, we show a probabilistic graphical representation of our hierarchical model. In the figure, grey circular nodes represent the prior hyperparameters which will be used for sensitivity analysis of the model. The transparent circular nodes are used to denote the modelling parameters which are our quantities of interest. The observed quantities are denoted with transparent rectangular nodes. We also use a grey rectangular node to denote the

intermediate latent variable T^*. We use directed edges to denote the relationship between different nodes. However, we use a dashed edge between X and T as they are related through the latent variable T^*.

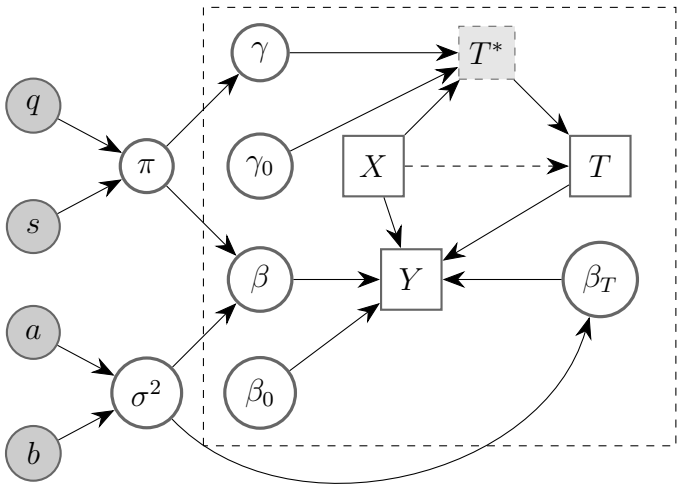

Figure 3: Probabilistic graphical representation for causal inference with our Bayesian hierarchical model.

3.2. Robust Bayesian Analysis

The hierarchical model presented above is a standard spike and slab model for variable selection and performs well when we have sufficient data. However, especially in many situations, we do not always have sufficient data. Moreover, we also must be cautious about the side effects of a treatment. Therefore, we are particularly interested in constructing a robust Bayesian framework for variable selection. In this way, when we are preparing guidelines for treatment, we can have the option to ask for more data before reaching any conclusion. To achieve this, we consider a utility based framework with three possible outcomes.

In particular, through predictor selection, we also want to check if a certain bio-marker should be considered for the treatment decision. For example, say we can observe blood pressure with some other bio-markers and want to decide whether our treatment guideline should also consider the blood pressure of the subject before treating them. This is useful as an unnecessary treatment of a subject can have severe consequences because of the

medicinal side effects. In general, it is hard to associate such consequences with a suitable loss function. Instead, we assume that we can always revert any initial incorrect treatment by further treatments, and we can associate a loss function with the cost of further treatments. So, we will associate two constant loss values ℓ_1 and ℓ_2 with false positives (falsely selected predictor) and false negatives (falsely rejected predictor) respectively. Clearly, false positives may lead to unwanted side effects and false negatives may lead to incorrect treatment of the patient. Finally, we associate a loss value ℓ_3 for abstention from selecting a variable which can be interpreted as the cost of further tests to determine whether that bio-marker is important for constructing the treatment guideline. Ideally, in most cases, $\ell_3 \ll \ell_1, \ell_2$. However, in certain scenarios, this might not be the case, especially when the condition of a subject deteriorates rapidly over time.

Now, based on this notion of abstaining from selecting a predictor, we can perform a sensitivity analysis over a set of priors on the prior selection probability. That is, we can consider a set of possible values for q such that $q \in \mathcal{P}$, where $\mathcal{P} \subseteq (0,1)^p$. Here, the equality occurs for the near vacuous case. However, in real-life situations, performing a robust Bayesian analysis for the near vacuous case is not practical. Instead, we incorporate expert elicitation to define our model.

For instance, assume $\underline{k}$ and $\overline{k}$ represent the expert's bounds of the prior expectation on the total number of variables present in either of the models. We can then consider $\mathcal{P} = \left[\underline{k}/p, \overline{k}/p\right]^p$. Using an interval for the prior expectation on the total number of active variables gives us a more cautious approach to specifying the prior distribution on variable selection, and thus more robust inferences.

Alternatively, we may also use the empirically observed correlations from the data directly. This is particularly common in ultra high dimensional problems (when $p \gg n$) for reducing the dimensionality of the problem [24]. We can also use this approach to have a better prior judgement since any predictors that are correlated with the outcome are good candidates to be active. When doing so, we need a prior judgement on what is a reasonable correlation between active predictors and the outcome. Say the expert judges that an active predictor has a correlation with the outcome that lies typically in $[-1, -c] \cup [c, 1]$, i.e. an absolute correlation larger than c. Let k_c be the number of predictors with absolute marginal correlation greater than c. We could then consider $q = k_c/p$ for the prior, as it gives a prior estimate on the selection probability that is consistent with a prior predictive expectation of

k_c active variables. Now, it is in general quite difficult to specify an exact value for c a priori. Therefore, we consider an interval $[\underline{c}, \overline{c}]$ for c, leading to $\mathcal{P} = [k_{\overline{c}}/p, k_{\underline{c}}/p]^p$ (note that k_c is monotonically non-increasing in c).

Variable selection. Ideally, we should check the joint posterior probability of π_j's to select the most probable model. However, this means we have to search a space of dimension 2^p, which is practically impossible when p is very large. Instead, we can use the posterior of individual π_j as Barbieri and Berger [25] showed that median probability model gives the optimal model. That is we can set a threshold of $1/2$ to select a variable. Therefore, we consider the j-th predictor to be removed from both the treatment and outcome model, if

$$\overline{\mathbb{E}}(\pi_j \mid W) \coloneqq \sup_{q \in \mathcal{P}} \mathbb{E}_q(\pi_j \mid W) < 1/2. \tag{20}$$

Similarly, we consider the j-th predictor to be present in at least one of the models, if

$$\underline{\mathbb{E}}(\pi_j \mid W) \coloneqq \inf_{q \in \mathcal{P}} \mathbb{E}_q(\pi_j \mid W) \geq 1/2. \tag{21}$$

Otherwise, we consider the variable to be indeterminate, in which case we abstain from putting it in any of the models but instead just report a lack of information.

3.3. Coefficient Adjustment and Refit

In general, our framework is intended for robust variable selection in causal effect estimation problem. However, one might also be interested in model fitting and prediction, for that we need to evaluate the values of the regression coefficients. To do so, we first need to find the set of active predictors with respect to our prior expectation of the selection probability q. For any fixed q, we define the set $S(q)$ as the set of all variables which are active in the treatment model or in the outcome model:

$$S(q) \coloneqq \{j \colon \mathbb{E}_q(\pi_j \mid W) \geq 1/2\}. \tag{22}$$

For sensitivity analysis, the intersection of $S(q)$ over all q gives us the set of active variables obtained through Eq. (21). Similarly, the union gives us the

set of variables that are not removed through Eq. (20). That is:

$$\mathcal{S}_* \coloneqq \{j : \underline{\mathbb{E}}(\pi_j \mid W) \geq 1/2\} = \bigcap_{q \in \mathcal{P}} S(q), \tag{23}$$

$$\mathcal{S}^* \coloneqq \left\{j : \overline{\mathbb{E}}(\pi_j \mid W) \geq 1/2\right\} = \bigcup_{q \in \mathcal{P}} S(q). \tag{24}$$

Clearly, $\mathcal{S}_* \subseteq \mathcal{S}^*$. $\mathcal{S}_*$ represents the set of variables that are sure to be selected, $\{1, \ldots, p\} \setminus \mathcal{S}^*$ represents the set of variables that are sure to be removed, and $\mathcal{S}^* \setminus \mathcal{S}_*$ represents the set of variables about which we are undecided. In this way, through sensitivity analysis, our approach incorporates robustness.

We can derive bounds on the posterior means of the parameters as follows:

$$\underline{\beta}_j \coloneqq \underline{\mathbb{E}}(\beta_j \mid W) = \inf_{q \in \mathcal{P}} \mathbb{E}_q(\beta_j \mid W) \tag{25}$$

$$\overline{\beta}_j \coloneqq \overline{\mathbb{E}}(\beta_j \mid W) = \sup_{q \in \mathcal{P}} \mathbb{E}_q(\beta_j \mid W) \tag{26}$$

with similar expressions for $\underline{\beta}_T$, $\overline{\beta}_T$, $\underline{\gamma}_j$ and $\overline{\gamma}_j$. If we take the posterior expectation interval $[0, 0] = \{0\}$ on a regression coefficient to represent absence of a variable, then our bounds on the regression coefficients are generally not sparse, because we use continuous spike and slab priors.

Moreover, with our variable selection we only determine whether the variable is included in at least one of the models. To determine which predictors influence the outcome ($\beta_j \neq 0$), the treatment ($\gamma_j \neq 0$), or both, and to understand the degree of assocation (i.e. the magnitude of β_j and/or γ_j), we apply the "decoupled shrinkage and selection" (DSS) method proposed by [21]. For that, we solve the following adaptive LASSO-type [26] problems:

$$\hat{\beta}^D_{S(q)} = \arg\min_{\beta_{S(q)}} \frac{1}{n} \|X_{S(q)}\hat{\beta}_{S(q)} - X_{S(q)}\beta_{S(q)}\|_2^2 + \lambda \sum_{j \in S(q)} \frac{|\beta_{j,S(q)}|}{|\hat{\beta}_{j,S(q)}|} \tag{27}$$

and

$$\hat{\gamma}^D_{S(q)} = \arg\min_{\gamma_{S(q)}} \frac{1}{n} \|X_{S(q)}\hat{\gamma}_{S(q)} - X_{S(q)}\gamma_{S(q)}\|_2^2 + \lambda \sum_{j \in S(q)} \frac{|\gamma_{j,S(q)}|}{|\hat{\gamma}_{j,S(q)}|} \tag{28}$$

where $q \in \mathcal{P}$, where $\hat{\beta}_{S(q)}$ and $\hat{\gamma}_{S(q)}$ are the posterior means of the regression coefficients with respect to the predictors that belong to $S(q)$. By varying

q, this gives us a set of point estimates for the model parameters β and γ, along with a more detailed selection of individual β_j and γ_j.

To compute the posterior bounds (as in Eqs. (20), (21), (25) and (26)), unfortunately, we usually have to resort to brute force optimisation, due to the lack of tractable expressions for the posterior expectations. This is obviously a major drawback of this approach.

Refit. In our setting, the DSS method only gives us a set of point estimates for the final selection of variables: some coefficients may be always selected, some never, and some will be indeterminate. For the final inference model, the modeller will need to make a judgement about which of the indeterminate coefficients β_j and γ_j to include in the final model or not. Once done so, the model can be refitted to account for the effect of variable selection on the estimation of the model parameters.

To do so, we can again use our Bayesian model without π_j (as there is no selection anymore), and with priors

$$\beta_j \mid \sigma^2 \sim \mathcal{N}(0, \tau_1^2 \sigma^2) \tag{29}$$
$$\gamma_j \mid \sigma^2 \sim \mathcal{N}(0, \tau_1^2) \tag{30}$$

for those β_j and γ_j that are selected in the model, with the remaining β_j and γ_j set to zero. This is similar to the spike and slab prior from Eq. (14) but without the spike component.

We expect this to have only a small effect on the mean and variance of the estimated parameters. This refit is useful to validate the variable selection and to improve the estimating of the model parameters, including the causal effect β_T. Indeed, since there are fewer parameters for the same data, the estimates are expected to have less uncertainty.

Note that here, we described a precise Bayesian refit model, but obviously this could be extended to robust Bayesian refit models too.

4. Simulation Studies

For the simulation studies, we consider 2 different cases each with 2 sub-cases, amounting to 4 studies in total. In each of these 4 studies, we generate the design matrix X such that $X_i \sim \mathcal{N}(0, \Sigma)$ for $1 \leq i \leq n$ where $\Sigma_{ij} = 0.3^{|i-j|}$. In this way, we generate predictors for our model with mild

correlations between them. We then use the following distributions to generate the outcome and the treatment indicator:

$$T_i \sim \text{Bernoulli}\left(1/(1+\exp(-X_i\gamma))\right) \quad \text{and} \quad Y_i = 4T_i + X_i\beta + \epsilon_i. \tag{31}$$

where $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$. Note that the simulated causal effect β_T is equal to 4.

In case 1, we consider an increasing number of observations. We have two sub-cases: in case 1a we consider all active variables to be confounders and in case 1b we consider some active variables which are only related to the outcome model.

Case 1a — $|\gamma_j|, |\beta_j| > 0$ for $j \leq 10$

Case 1b — $|\gamma_j| > 0$ for $j \leq 10$ and $|\beta_j| > 0$ for $j \leq 15$

For both case 1a and 1b, we consider different numbers of observations n where $n = 20 + 5k$ for $k = 1, 2, \ldots, 11$ and $p = 50$ predictors. This way, we check the efficiency of our method with varying level of information.

For case 2, we check our method for varying number of predictors (and hence sparsity level, i.e. the percentage of active variables present in the model). Similar to case 1 we also have two sub-cases:

Case 2a — $|\gamma_j|, |\beta_j| > 0$ for $j \leq 10$

Case 2b — $|\gamma_j| > 0$ for $j \leq 10$ and $|\beta_j| > 0$ for $j \leq 15$

For both case 2a and 2b, we consider different numbers of predictors p where $p = 20 + 5k$ for $k = 1, 2, \ldots, 11$ and $n = 40$ subjects.

For all four cases, we consider 20 replicates for an empirical statistical analysis to check the consistency and robustness of our approach.

We use these studies to compare our method with three other approaches. From now on, for the sake of illustration, we use the following acronyms: RBCE for robust Bayesian causal estimation (our method); SSCE for spike and slab causal estimation [10]; BSSCE for bi-level spike and slab causal estimation [10]; and BSSL for Bayesian spike and slab LASSO [20].

Metrics. As mentioned earlier, to perform our statistical analyses we use 20 replications. To evaluate the accuracy of estimation, we consider mean and median values obtained from these 20 samples. Similarly, to check the dispersion, we use standard deviation (denoted by sd), mean squared error

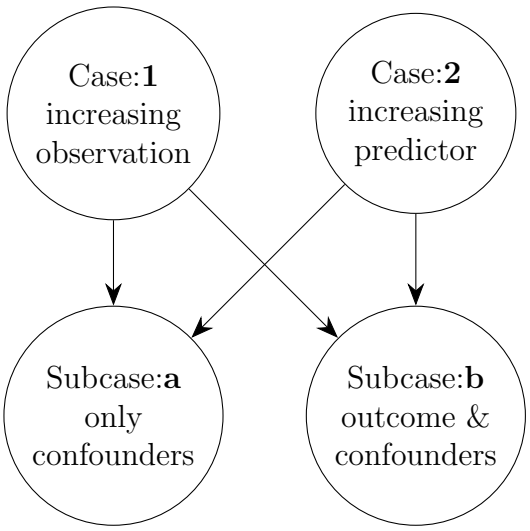

Figure 4: Diagram of the simulation setup for comparative study.

with respect to the true value (denoted by MSE) and coverage (percentage) of the true value within the 95% posterior credible interval (denoted by CI%). Finally, to check the accuracy of variable selection, we evaluate false average positive numbers (denoted by FP), average false negative numbers (denoted by FN) and average number of indeterminate variables (denoted by ID). Clearly for other classical methods ID is equal to zero and therefore is not presented in the tables. We also define a misspecification loss in the following way for illustration :

$$\text{Misspecification Loss} = \frac{FP}{TN} + \frac{FN}{TP} + 0.2 * \frac{ID}{\text{Total no of predictors}}. \quad (32)$$

Note that since RBCE gives interval estimates, CI% is calculated using the minimum of the lower bounds of credible intervals and the maximum of the upper bounds of credible intervals.

Elicitation. To elicit $\mathcal{P}$, as discussed earlier in Section 3.2, we use the empirically observed correlations from the data directly. For expert elicitation, we follow the correlation guidelines mentioned in [27] where the authors provide Table 1.

From Table 1, we notice that the number of labelled relations is different for different columns. As a result it is difficult to obtain a single value for c as

Table 1: Interpretation of the Pearson's and Spearman's correlation coefficients in absolute values

Absolute Correlation	Dancey & Reidy[28] (Psychology)	Chan YH[29] (Medicine)	Quinnipiac University (Politics)
1.0	Perfect	Perfect	Perfect
0.9	Strong	Very strong	Very strong
0.8	Strong	Very Strong	Very strong
0.7	Strong	Moderate	Very strong
0.6	Moderate	Moderate	Strong
0.5	Moderate	Fair	Strong
0.4	Moderate	Fair	Strong
0.3	Weak	Fair	Moderate
0.2	Weak	Poor	Weak
0.1	Weak	Poor	Negligible
0.0	Zero	Zero	Zero

mentioned in Section 3.2. Instead, we can disregard the last labelled relation (other than zero) and assume that c is typically larger than a value lying in the interval $[0.15, 0.35]$ Let $\overline{k}$ be the number of predictors with absolute marginal correlation greater than 0.15 and let $\underline{k}$ be number of predictors with absolute marginal correlation greater than 0.35. Then $\mathcal{P} = [\underline{k}/p, \overline{k}/p]^p$ gives us a prior bound on the selection probability of each predictor, reflecting our prior expert judgement.

Initialisation. To implement our method, we use `rjags` [30] and for the other three methods we use the code provided in the appendix of [10]. However, we modify to accommodate analysis with 'high dimensional' data. For our method, we set $\tau_0 = 10^{-6}$ and $\tau_1 = 1$ to construct the spike and slab prior. For the noise term, we set $a = 50$ and $b = 1$. To perform our Bayesian analysis with `rjags`, we discard 500 burn in samples and consider 2500 MCMC samples to compute the posterior estimates. For the other methods we use the in-built settings to initiate the analyses. We also transform the data so that the data is centred around 0 for the outcome model to avoid having an intercept term.

Results. Table 2 shows the results of estimating the causal effect β_T for case 1a. For reference, recall the true value is $\beta_T = 4$. As we perform a sensitivity analysis, our method gives an interval estimate for the causal effect. So we

Table 2: Comparison of different methods for varying number of observations where all the active variables are confounders.

(a) Accuracy in estimation of causal effect

	RBCE				SCCE		BSSCE		BSSL	
Obs	Mean		Median		Mean	Median	Mean	Median	Mean	Median
25	2.6	3.5	2.5	3.6	20.4	20.4	20.9	20.7	15.6	19.0
30	3.0	3.7	2.9	3.7	18.4	19.9	20.9	21.2	10.3	4.1
35	3.3	3.8	3.3	3.8	15.8	18.5	19.6	19.8	7.2	4.1
40	3.4	3.8	3.4	3.8	11.8	10.4	16.4	18.3	4.2	4.0
45	3.6	3.8	3.6	3.8	8.1	4.1	11.0	11.3	4.1	4.0
50	3.7	3.8	3.6	3.8	7.5	4.1	7.8	4.2	4.0	4.0
55	3.7	3.9	3.7	3.9	4.5	4.0	4.4	4.0	4.0	4.0
60	3.8	3.9	3.8	3.9	4.2	4.0	4.0	4.0	4.0	4.0
65	3.8	3.9	3.8	3.9	4.1	4.0	4.0	4.0	4.0	4.0
70	3.8	3.9	3.8	3.9	4.0	4.0	4.0	4.0	4.0	4.0
75	3.8	3.9	3.8	3.9	4.0	4.0	4.0	4.0	4.0	4.0

(b) Dispersion of estimated causal effect: values less than 0.05 are replaced with *

	RBCE					SCCE			BSSCE			BSSL		
Obs	sd		MSE		CI%	sd	MSE	CI%	sd	MSE	CI%	sd	MSE	CI%
25	0.4	0.5	0.5	2.1	100	3.1	277.4	0	3.0	295.1	0	9.3	217.6	20
30	0.4	0.5	0.2	1.3	100	5.9	239.9	15	2.7	291.9	0	8.4	107.2	60
35	0.3	0.4	0.1	0.7	100	7.2	187.0	25	4.3	261.4	10	6.7	53.7	80
40	0.2	0.3	0.1	0.4	100	7.8	118.7	45	6.4	191.7	20	0.9	0.7	95
45	0.2	0.2	0.1	0.2	100	6.0	50.7	60	6.6	90.5	45	0.5	0.3	95
50	0.1	0.2	*	0.1	100	5.3	39.6	65	5.5	43.0	65	0.1	*	100
55	0.1	0.1	*	0.1	100	1.6	2.6	95	1.6	2.5	95	0.1	*	95
60	0.1	0.1	*	0.1	100	0.9	0.9	90	0.1	*	100	*	*	95
65	0.1	0.1	*	*	100	0.5	0.3	95	0.1	*	95	*	*	95
70	0.1	0.1	*	*	100	*	*	95	*	*	95	*	*	95
75	0.1	0.1	*	*	100	*	*	95	*	*	95	*	*	95

(c) Accuracy of variable selection: all the values are averaged over 20 replications

	RBCE			SCCE		BSSCE		BSSL	
Obs	FP	FN	ID	FP	FN	FP	FN	FP	FN
25	0.7	0.2	30.4	0	9.8	0	10	1.4	7.1
30	0	0	19.2	0	8.8	0	10	0	3.8
35	0	0	8.1	0	7.3	0	9.4	0	2.0
40	0	0	3.0	0	5.1	0	8.2	0	0.2
45	0	0	0.6	0	2.9	0	5.0	0	0
50	0	0	0.4	0	2.9	0	3.0	0	0
55	0	0	0	0	0.6	0	0.4	0	0
60	0	0	0	0	0.2	0	0	0	0
65	0	0	0	0	0.2	0	0	0	0
70	0	0	0	0	0	0	0	0	0
75	0	0	0	0	0	0	0	0	0

present mean, median, sd and MSE for RBCE using two columns where the left columns give the lower bounds and the right columns give the upper bounds. We notice that as we increase the number of observation our approach provides more precise estimates which shows that our set of priors is able to learn from the data. We also observe that our method tends to under estimate the causal effect. However, our approach does not produce any extreme values and is more consistent in terms of estimating the causal effect, especially for fewer number of observations which is not the case for SSCE and BSSCE. For BSSL we notice that median value is close to the true value for fewer number of observations but mean value is higher which shows that for some experiments BSSL tend to produce extreme values. This can

also be from the table for dispersion in estimation where BSSL tends to have a high MSE and lower CI% for fewer number of observations.

We also provide the performance in variable selection in Table 2. We notice that for fewer number of observations our method tends to give many indeterminate variables but this number gradually decreases as we increase the number of observations. However, our elicitation based approach ensures that we have very few false negative and false positive variables which is not the case for other approaches.

Table 3: Comparison of different methods in estimating the causal effect for varying number of observations where some variables are only related to the outcome model.

(a) Accuracy in estimation of causal effect

	RBCE				SCCE		BSSCE		BSSL	
Obs	Mean		Median		Mean	Median	Mean	Median	Mean	Median
25	2.8	3.7	2.8	3.8	25.5	25.0	26.2	25.5	20.5	24.1
30	3.2	4.1	3.1	4.1	25.4	25.1	26.4	26.2	15.8	10.8
35	3.7	4.3	3.7	4.3	22.3	25.1	26.2	26.0	10.9	4.0
40	3.9	4.3	3.9	4.4	21.4	24.6	26.0	25.4	4.7	4.0
45	3.9	4.2	3.9	4.2	18.7	23.9	22.5	25.2	4.0	4.0
50	3.9	4.2	3.9	4.2	9.6	4.0	13.2	6.2	4.0	4.0
55	4.0	4.1	3.9	4.2	7.1	4.0	9.3	4.1	4.0	4.0
60	4.0	4.1	4.0	4.1	4.0	4.0	4.0	4.0	4.0	4.0
65	4.0	4.1	4.0	4.1	4.6	4.0	4.5	4.0	4.0	4.0
70	4.0	4.1	4.0	4.1	4.0	4.0	4.0	4.0	4.0	4.0
75	4.0	4.1	4.0	4.1	4.0	4.0	4.0	4.0	4.0	4.0

(b) Dispersion of estimated causal effect: values less than 0.05 are replaced with *

	RBCE					SCCE			BSSCE			BSSL		
Obs	sd		MSE		CI%	sd	MSE	CI%	sd	MSE	CI%	sd	MSE	CI%
25	0.4	0.6	0.4	1.6	100	5.6	492.2	10	5.2	517.1	0	10.4	375.8	10
30	0.6	0.6	0.3	0.9	100	5.8	488.8	5	5.0	525.3	0	12.2	280.9	50
35	0.6	0.6	0.3	0.5	100	8.5	405.3	20	4.7	514.7	0	9.4	130.9	60
40	0.4	0.5	0.2	0.3	100	9.6	390.6	20	4.4	500.1	0	3.4	11.3	95
45	0.3	0.3	0.1	0.1	100	11.0	330.5	30	8.5	413.2	15	*	*	100
50	0.3	0.3	0.1	0.1	100	9.8	123.0	70	11.0	199.1	50	*	*	100
55	0.2	0.3	*	0.1	100	7.5	63.4	85	9.4	110.9	75	*	*	100
60	0.2	0.2	*	0.1	100	*	*	100	*	*	100	*	*	100
65	0.1	0.2	*	*	100	2.4	6.0	95	2.1	4.3	95	*	*	100
70	0.1	0.1	*	*	100	*	*	100	*	*	100	*	*	100
75	0.1	0.1	*	*	100	*	*	100	*	*	100	*	*	100

(c) Accuracy of variable selection: all the values are averaged over 20 replications

	RBCE			SCCE		BSSCE		BSSL	
Obs	FP	FN	ID	FP	FN	FP	FN	FP	FN
25	1.8	0.6	31.8	0	14.7	0	14.9	2.0	11.6
30	0.6	0.4	27.2	0	14.8	0	14.9	1.0	7.7
35	0.2	0.5	16.9	0	13.1	0	15.0	0	5.2
40	0	0.2	10	0	12.8	0	15.0	0	0.6
45	0	0	2.9	0	10.2	0	12.9	0	0
50	0	0	0.9	0	3.9	0	6.6	0	0
55	0	0	0.4	0	2.0	0	3.5	0	0
60	0	0	0.3	0	0	0	0	0	0
65	0	0	0.3	0	0.6	0	0.6	0	0
70	0	0	0	0	0	0	0	0	0
75	0	0	0	0	0	0	0	0	0

We present our analysis for case 1b in Table 3. Similar to our analyses for case 1a, we notice that as we obtain more observations the imprecision in the estimation reduces. However, unlike case 1a, our approach tends to

over estimate the causal effect for higher number of observations. This also shows an overall increasing trend of the estimated causal effect similar to case 1a. We also notice that for case 1b number of indeterminate variables is higher than that of case 1a. This happens as some of the variables are only related to the outcome model. This also contributes to higher number of false negative variables in for other methods. We also observe that similar to the previous case, other methods often produces extreme values for the causal effect increasing the sd and MSE of the estimated causal effect.

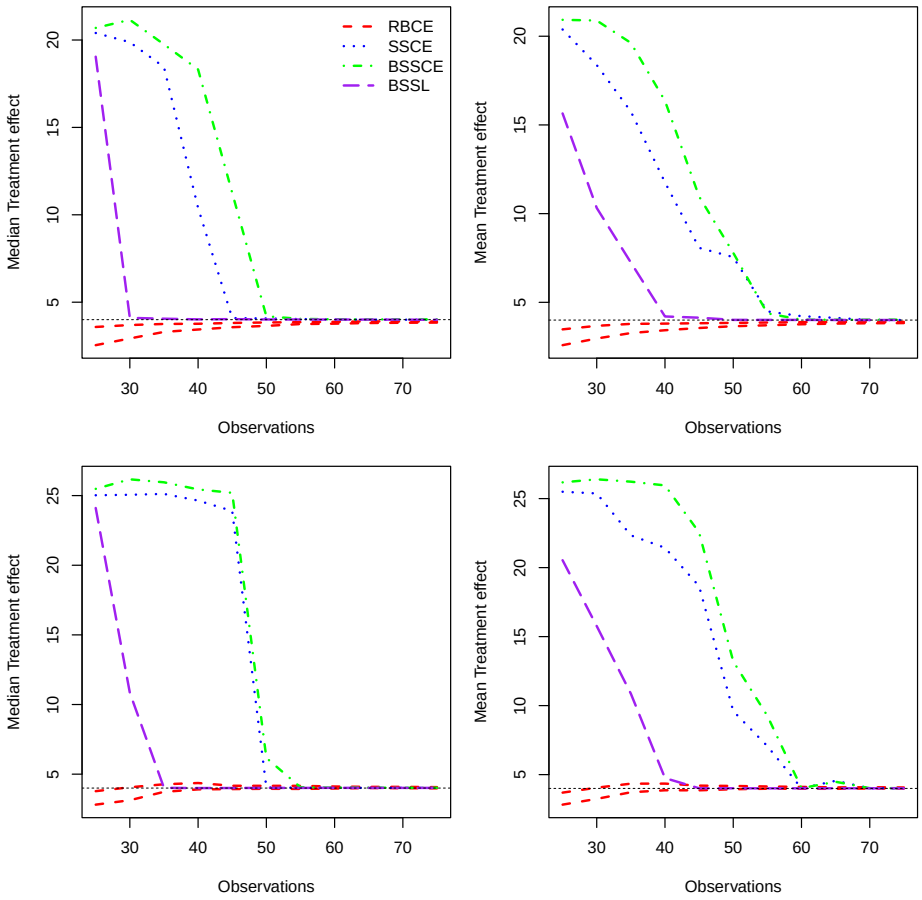

Figure 5: Comparison of different methods in estimating the causal effect for varying number of observations. The top (bottom) row represents case 1a (case 1b). The left(right) images show the average(median) causal effects obtained from 20 replications.

We illustrate the estimated causal effect in Fig. 5 as well. In the figure, the

top row illustrates the case 1a where the left image shows the average value of the estimated causal effect with respect to observations and the right image shows the median value. Similarly, the bottom row represents the same for case 1b. In the figure, RBCE bounds are given by red lines; SSCE estimates by blue lines; BSSCE estimates by green lines; BSSL estimates by purple line; and true value by black lines. In the figure, we can also notice the increase trend of the estimated causal effect as we obtain more observations and also the estimation becomes more precise.

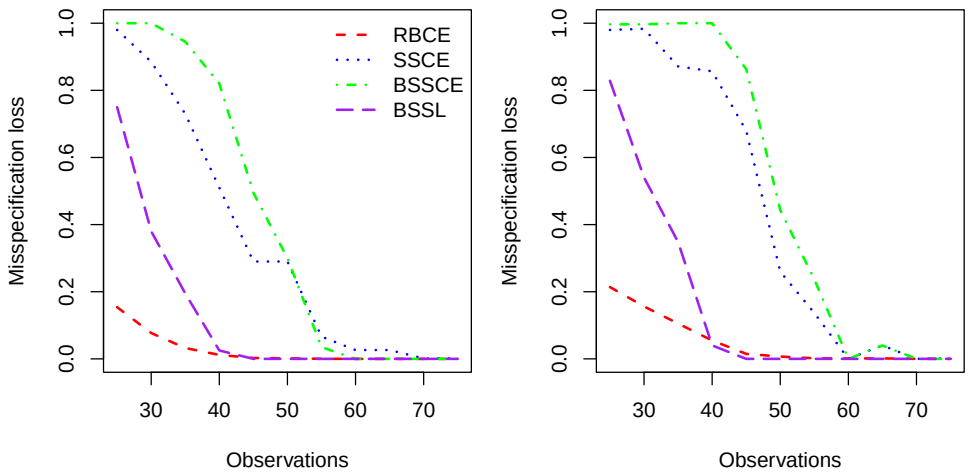

Figure 6: Comparison of different methods in identifying the confounders for varying number of observations. One the left (right) we present case 1a (1b). The red line presents RBCE; blue line presents SSCE; green line presents BSSCE; and purple line presents BSSL

We also illustrate the performance of variable selection in Fig. 6. For variable selection, we use a loss function as described earlier. Here, we consider $\ell_1 = \ell_2 = 1$ and $\ell_3 = 0.2$, i.e. we associate a loss of 1 with false positive and false negative selections, and a loss of 0.2 with indeterminate selections. Note that we could also choose more sophisticated loss functions based on [31]. We evaluate the misspecification loss using the equation that we described before. From the figure it can be seen that for case 1, our method abstains from identifying some variables for $n \leq 40$. However, later on our method gives more precise results in terms of variable selection. We also notice that the SSCE, BSSCE and BSSL tend to perform poorly in terms of variable selection. However, BSSL performs better than the rest for higher number of observations

Table 4: Comparison of different methods in estimating the causal effect for varying number of predictors where all the active variables are confounders.

(a) Accuracy in estimation of causal effect

	RBCE				SCCE		BSSCE		BSSL	
Pred	Mean		Median		Mean	Median	Mean	Median	Mean	Median
1	2	3	4	5	6	7	8	9	10	11
25	3.6	3.8	3.7	3.8	9.5	4.6	11.8	14.0	4.0	4.0
30	3.6	3.8	3.6	3.8	8.8	4.1	9.7	4.7	4.5	4.0
35	3.6	3.8	3.6	3.8	12.8	17.4	12.8	14.9	4.9	4.0
40	3.5	3.8	3.5	3.8	12.3	15.2	14.9	17.6	4.0	4.0
45	3.5	3.8	3.4	3.8	10.7	4.1	16.6	18.3	4.0	4.0
50	3.4	3.8	3.5	3.8	11.8	10.4	16.4	18.3	4.2	4.0
55	3.4	3.8	3.4	3.8	12.4	15.6	16.6	18.8	4.5	4.0
60	3.3	3.8	3.3	3.8	14.2	18.7	17.3	19.1	4.1	4.0
65	3.2	3.8	3.3	3.8	12.6	15.5	19.4	19.2	5.0	4.0
70	3.1	3.8	3.2	3.8	11.1	4.1	16.7	19.1	4.0	4.0
75	3.1	3.8	3.1	3.8	11.7	9.3	18.4	19.2	4.5	4.0

(b) Dispersion of estimated causal effect: values less than 0.05 are replaced with *

	RBCE					SCCE			BSSCE			BSSL		
Pred	sd		MSE		CI%	sd	MSE	CI%	sd	MSE	CI%	sd	MSE	CI%
25	0.2	0.2	0.1	0.2	100	6.6	72.1	50	7.1	109.3	45	0.1	*	100
30	0.2	0.2	0.1	0.2	100	6.6	64.4	65	6.8	76.7	60	2.0	3.9	95
35	0.2	0.2	0.1	0.2	100	7.4	129.2	35	7.4	130.1	40	4.1	16.6	90
40	0.2	0.2	0.1	0.2	100	7.9	129.0	45	6.7	162.0	30	0.1	*	95
45	0.2	0.2	0.1	0.3	100	7.8	102.5	55	5.7	190.4	20	0.1	*	100
50	0.2	0.3	0.1	0.4	100	7.8	118.7	45	6.4	191.7	20	0.9	0.7	95
55	0.2	0.3	0.1	0.5	100	8.0	132.4	45	6.8	201.9	25	2.3	5.2	95
60	0.2	0.3	0.1	0.6	100	8.1	165.8	30	6.2	214.7	15	0.6	0.4	100
65	0.2	0.4	0.1	0.8	100	8.3	139.2	40	2.8	244.0	0	4.4	19.1	95
70	0.2	0.4	0.1	0.9	100	8.2	114.9	55	6.6	202.1	25	0.1	*	95
75	0.2	0.4	0.1	1.0	100	8.0	120.2	50	4.8	228.7	10	3.5	11.9	85

(c) Accuracy of variable selection: all the values are averaged over 20 replications

	RBCE			SCCE		BSSCE		BSSL	
Pred	FP	FN	ID	FP	FN	FP	FN	FP	FN
25	0	0	5.2	0	4.0	0	5.5	0	0
30	0	0	2.4	0	3.5	0	4.1	0	0.4
35	0	0	2.0	0	6.0	0	6.0	0	0.5
40	0	0	2.1	0	5.4	0	7.3	0	0
45	0	0	2.6	0	4.4	0	8.4	0	0
50	0	0	3.0	0	5.1	0	8.2	0	0.2
55	0	0	3.8	0	5.4	0	7.9	0	0.4
60	0	0	5.0	0	6.4	0	8.4	0	0
65	0	0	4.4	0	5.4	0	9.9	0	0.5
70	0	0	4.8	0	4.4	0	8.0	0	0
75	0	0	6.2	0	4.9	0	9.2	2.4	0.6

We show the result of our analyses case 2a in Table 4. Similar to our analyses with increasing number of observations, we notice that our method is overall in agreement with BSSL. However, similar to case 1a our method tends to underestimate the treatment effect (approximately 5%) for case 2b. We also notice that the imprecision in estimation increases as we increase the number of predictors. This happens as observation per predictor reduces. We also notice that BSSL outperforms RBCE in terms of median value of estimated causal effect over 20 replications. However, in very few cases BSSL provides extreme values which can be understood from mean and CI% as well as MSE. Moreover, for 75 predictors BSSL gives higher number of false positives which is not the case for RBCE. Unlike the case 1a and 1b, SSCE

and BSSCE performs poorly for every value of predictors.

Table 5: Comparison of different methods in estimating the causal effect for varying number of predictors where some variables are only to the outcome model.

(a) Accuracy in estimation of causal effect

	RBCE				SCCE		BSSCE		BSSL	
Pred	Mean		Median		Mean	Median	Mean	Median	Mean	Median
1	2	3	4	5	6	7	8	9	10	11
25	3.8	4.1	3.8	4.1	19.0	22.3	24.0	25.3	4.0	4.0
30	3.9	4.2	3.9	4.1	20.6	24.6	20.9	24.5	7.4	4.0
35	3.9	4.3	3.9	4.2	21.9	24.7	23.7	24.8	7.3	4.0
40	3.9	4.3	3.9	4.3	20.9	24.4	24.7	25.5	5.1	4.0
45	3.8	4.3	3.9	4.3	20.2	23.9	26.1	25.4	4.8	4.0
50	3.9	4.3	3.9	4.3	21.4	24.6	26.0	25.4	4.7	4.0
55	3.8	4.4	3.9	4.4	21.8	24.8	26.0	25.1	9.5	4.0
60	3.8	4.4	3.9	4.4	16.6	19.7	25.1	25.5	8.4	4.0
65	3.7	4.4	3.8	4.4	18.7	22.8	26.2	25.6	5.4	4.0
70	3.7	4.5	3.8	4.4	20.4	23.9	25.5	25.2	8.5	4.0
75	3.7	4.5	3.7	4.4	17.7	22.8	26.0	25.2	5.0	4.0

(b) Dispersion of estimated causal effect: values less than 0.05 are replaced with *

	RBCE					SCCE			BSSCE			BSSL		
Pred	sd		MSE		CI%	sd	MSE	CI%	sd	MSE	CI%	sd	MSE	CI%
25	0.2	0.3	0.1	0.1	100	11.2	343.7	35	8.0	459.5	10	0.1	*	100
30	0.3	0.4	0.1	0.1	100	10.4	377.0	25	10.8	397.7	25	8.3	77.9	85
35	0.3	0.4	0.1	0.2	100	9.9	412.9	20	7.4	438.6	10	8.0	71.7	85
40	0.4	0.4	0.1	0.3	100	9.3	366.8	15	7.0	475.8	10	4.8	23.1	95
45	0.4	0.4	0.2	0.3	100	10.4	365.5	25	4.4	504.8	0	3.4	11.3	95
50	0.4	0.5	0.2	0.3	100	9.6	390.6	20	4.4	500.1	0	3.4	11.3	95
55	0.4	0.5	0.2	0.3	100	9.9	410.6	20	4.5	503.9	0	9.9	123.1	75
60	0.5	0.5	0.2	0.4	100	11.7	288.5	40	6.4	484.0	5	9.3	102.4	80
65	0.4	0.5	0.2	0.4	100	11.2	334.4	30	4.5	510.0	0	6.4	40.5	95
70	0.5	0.5	0.2	0.5	100	11.1	385.3	25	5.2	488.8	5	9.3	102.1	80
75	0.5	0.5	0.2	0.5	100	11.8	318.1	40	4.5	502.1	0	4.0	16.3	95

(c) Accuracy of variable selection: all the values are averaged over 20 replications

	RBCE			SCCE		BSSCE		BSSL	
Pred	FP	FN	ID	FP	FN	FP	FN	FP	FN
25	0	0	7.8	0	10.2	0	13.5	0	0
30	0	0	8.2	0	11.6	0	11.2	0	2.0
35	0	0	9.2	0	12.6	0	14.0	0	2.2
40	0	0	10.1	0	11.9	0	13.8	0	0.8
45	0	0.1	10.8	0	11.8	0	15.0	0	0.6
50	0	0.2	10.1	0	12.8	0	15.0	0	0.6
55	0	0.3	8.9	0	12.0	0	15.0	0	3.5
60	0	0.4	11.1	0	9.2	0	14.4	0.5	2.9
65	0	0.8	11.9	0	10.6	0	15.0	0.4	0.8
70	0	0.8	12.8	0	10.9	0	14.8	0	2.8
75	0	1.0	12.3	0.1	9.6	0	15.0	0	0.9

The result for case 2b is presented in Table 5. We notice that for this case the true causal effect is always contained within the estimated bounds unlike the previous cases. For this case, the imprecision in the estimated causal effect increases with respect to predictors similar to case 2a as the observation per predictor reduces. We also see that SSCE and BSSCE performs poorly similar to case 2a and BSSL is mostly consistent in estimation but produces extreme values for some experiments giving a significant differences between mean and median of the estimated causal treatments.

We also show the causal effect estimation and performance in variable selection in Figs. 7 and 8. From Fig. 7 we can see the increase in imprecision

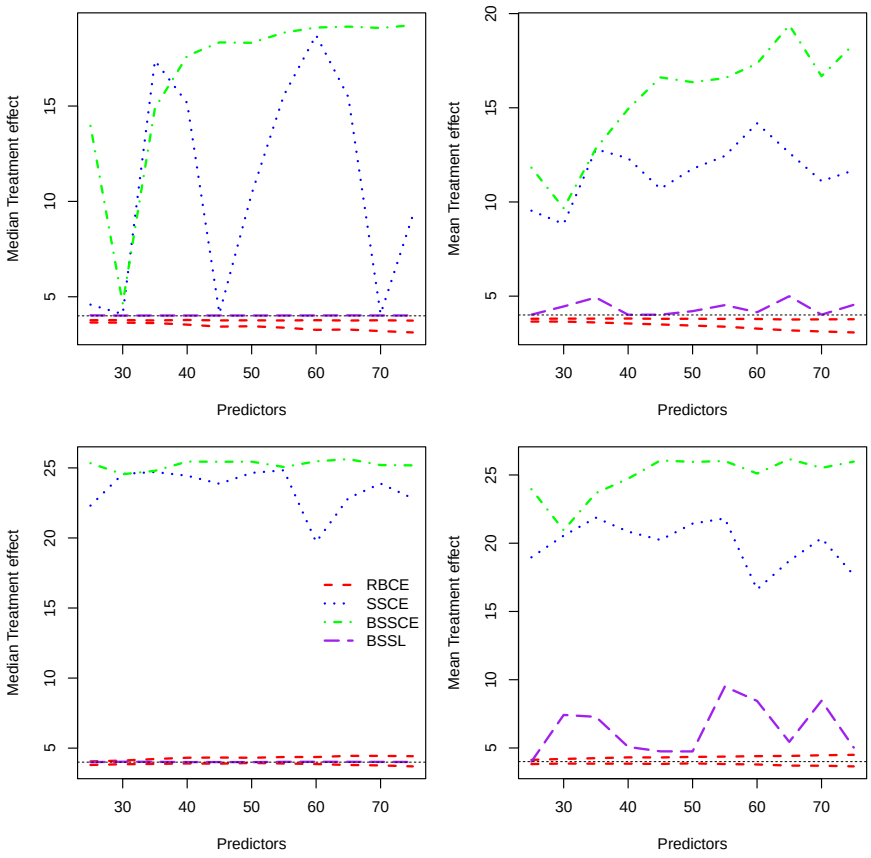

Figure 7: Comparison of different methods in estimating the causal effect for varying number of predictors. The top (bottom) row represents case 2a (case 2b). The left(right) images show the average(median) causal effects obtained from 20 replications.

as we increase then number of predictors. We also see that RBCE performs more consistently than other methods in terms of estimating the causal effect. We also notice that BSSL outperforms RBCE in terms of estimating the causal effect and performs at per in terms of variable selection for case 2a. However, for case 2b, BSSL appears to be less consistent in terms of variable selection. From these two figures we can also see that SSCE and BSSCE performs poorly as we increase the number of predictors and is particularly unstable for case 2a.

Importance of prior elicitation. Our method relies on expert elicitation and prior sensitivity analysis. So we also explore the effect of prior elicitation

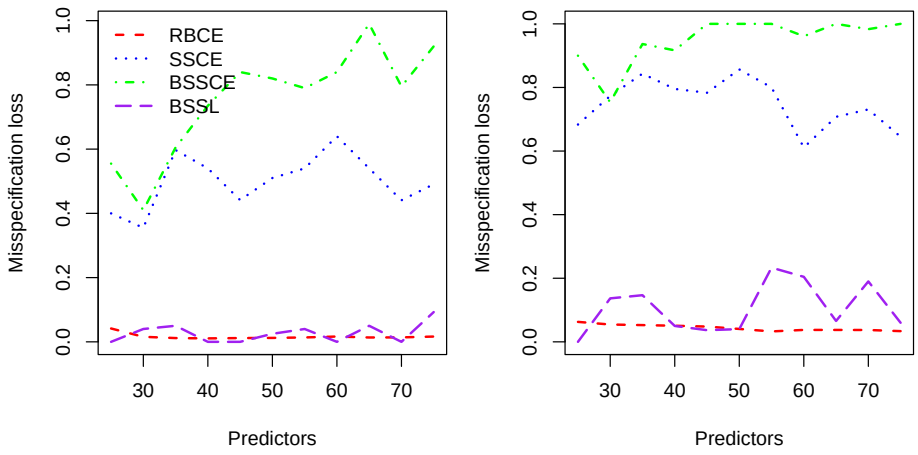

Figure 8: Comparison of different methods in identifying the confounders for varying number of predictors. One the left (right) we present case 1a (1b). The red line presents RBCE; blue line presents SSCE; green line presents BSSCE; and purple line presents BSSL

in identifying the active variables in our model. As mentioned earlier, we consider c is expected to be lying in the interval $[0.15, 0.35]$ based on Table 1. However, we might want to choose a different value for c. To compare the effect of having different value for c, we use the case 2b and set $c \in [0.2, 0.4]$. That is we set a higher threshold for the correlation so that $\overline{k}$ becomes smaller and hence the prior expectation of the inclusion probability. We show our results in Table 6. In the left hand side we elicit the expected number of active variables by setting the marginal correlation threshold $c \in [0.15, 0.35]$ and the right we set $c \in [0.2, 0.4]$. On the left hand side, we see our method tends to give higher number of indeterminate variables for fewer observations than the right hand side. As we increase the predictors the higher threshold of marginal correlation plays an important role and we see more cases of false negative variables on the right hand side. This also results to over estimation of the causal effect as many true active variables are shrinked towards zero. As a result the lower bound of the averaged causal effect is more than four on the right hand side.

The analyses and simulation studies can be investigated using the code from `https://github.com/tathagatabasu/Causal-Inference`.

Table 6: Effect of elicitation of the inclusion probability of the variables where FP stands for false positive, FN stands for false negative and ID stands for indeterminate.

$[\underline{c}, \overline{c}]$	$[0.15, 0.35]$					$[0.2, 0.4]$				
Pred	Mean		FP	FN	IDR	Mean		FP	FN	IDR
25	3.8	4.1	0	0	7.8	3.9	4.2	0	0	4.9
30	3.9	4.2	0	0	8.2	3.9	4.3	0	0	4.2
35	3.9	4.3	0	0	9.2	4.0	4.3	0	0.2	3.5
40	3.9	4.3	0	0	10.1	4.0	4.4	0	0.3	3.4
45	3.8	4.3	0	0.1	10.8	4.0	4.4	0	0.4	3.5
50	3.9	4.3	0	0.2	10.1	4.0	4.4	0	0.6	3.0
55	3.8	4.4	0	0.3	8.9	4.0	4.4	0	0.6	3.0
60	3.8	4.4	0	0.4	11.1	4.0	4.5	0	0.8	3.4
65	3.7	4.4	0	0.8	11.9	4.0	4.5	0	1.0	3.8
70	3.7	4.5	0	0.8	12.8	4.0	4.5	0	1.3	3.0
75	3.7	4.5	0	1.0	12.3	4.0	4.5	0	1.4	3.1

5. Conclusion

Causal effect estimation is an important tool in statistical learning. Especially in risk-sensitive situations, such as medicine, it needs to be performed with the utmost care as in many cases poor estimation can have severe adverse consequences. In this paper, we tackle this issue by proposing a robust Bayesian analysis of the causal effect estimation problem for high dimensional data. Our framework is focused on the effect of prior elicitation on predictor selection as well as causal effect estimation. We consider a spike and slab type prior for predictor selection and discuss the possible sources of uncertainty that need to be tackled carefully. We were particularly focused on the uncertainty associated with prior selection probabilities for which we consider a set of beta priors to perform sensitivity analysis. We showed that the sensitivity analysis on the prior selection probability gives us a robust predictor selection scheme. In this way, we can abstain from selecting a predictor when the available data is not sufficient. We also propose a more relaxed utility based framework, where we associate a loss for abstaining which can be interpreted as the cost of further data collection. We illustrate our method with synthetic dataset and compare with other state of the art Bayesian methods. We could see that our elicitation based approach helps to have a more consistent causal effect estimation for very limited number

of observations and avoids producing extreme values for the causal effect. Moreover, we also notice correct elicitation of the inclusion probability plays a crucial role in identifying the active variables and therefore can be extremely useful in cases where we need to design a treatment guideline with multiple bio-markers.

Currently, the paper proposes a robust Bayesian approach for causal effect estimation where we rely on sampling strategies to obtain the posterior bounds as well as performing variable selection. A weakness of our approach is simulation efficiency, as we resorted to brute force optimisation. However, there is ample opportunity to improve computational aspects. In the future, it will be interesting to derive inner approximation bounds for the posterior estimates to reduce the computational cost, or to find better ways than brute force optimisation, such as for instance iterative importance sampling [32].

To compare the different methods, we rely on simple loss functions associated with the predictor selection. However, loss functions could be used for a generalised decision theoretic framework as well. For instance, the selection problem itself could be formulated as a decision problem, potentially leading to different selection thresholds or even selection systems that are directly based on a loss function. Additionally, we could formulate the problem of whether or not to treat a subject as a decision support problem based on predictor selection.

Another topic of interest pertinent to medical diagnosis is missing data. It has been shown that using bounded probability is particularly suitable for dealing with instances where data cannot be assumed missing at random [33]. Incorporating robustness against missing data could lead to an interesting extension of the model in this paper.

We also notice with our simulation studies that our method tends to underestimate the causal effect when only confounders are present in the model. This suggests that we might want to use a correction formula for the causal effect. Moreover, in future, we would like to investigate different elicitation strategies for different prior parameters and their importance in causal effect estimation.

In general, we noticed that our method is in good agreement with other methods with an added level of robustness. This shows that our method has good potential for real-life problems, and we intend to apply it on a real dataset in future work.

Acknowledgements

We sincerely thank Jochen Einbeck for his contributions to an earlier version of this paper. We also thank all reviewers for their kind and constructive comments which helped improve the paper.

References

[1] D. B. Rubin, Bayesian inference for causal effects: The role of randomization, The Annals of Statistics 6 (1978) 34–58. URL: https://doi.org/10.1214/aos/1176344064. doi:doi:10.1214/aos/1176344064.

[2] P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1983) 41–55. URL: https://doi.org/10.1093/biomet/70.1.41. doi:doi:10.1093/biomet/70.1.41.

[3] J. M. Robins, 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 (1986) 1393–1512.

[4] C. Winship, S. L. Morgan, The estimation of causal effects from observational data, Annual Review of Sociology 25 (1999) 659–706. doi:doi:10.1146/annurev.soc.25.1.659.

[5] E. A. Stuart, Matching methods for causal inference: A review and a look forward, Statistical Science 25 (2010) 1–21. URL: https://doi.org/10.1214/09-STS313. doi:doi:10.1214/09-STS313.

[6] C. M. Zigler, F. Dominici, Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects, Journal of the American Statistical Association 109 (2014) 95–107. URL: http://www.jstor.org/stable/24247140.

[7] C. Wang, F. Dominici, G. Parmigiani, C. M. Zigler, Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models, Biometrics 71 (2015) 654–665. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12315. doi:doi:https://doi.org/10.1111/biom.12315.

[8] B. Koch, D. M. Vock, J. Wolfson, Covariate selection with group lasso and doubly robust estimation of causal effects, Biometrics 74 (2018) 8–17. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12736. doi:doi:10.1111/biom.12736.

[9] P. R. Hahn, C. M. Carvalho, D. Puelz, J. He, Regularization and confounding in linear regression for treatment effect estimation, Bayesian Analysis 13 (2018) 163–182. URL: https://doi.org/10.1214/16-BA1044. doi:doi:10.1214/16-BA1044.

[10] B. Koch, D. M. Vock, J. Wolfson, L. B. Vock, Variable selection and estimation in causal inference using Bayesian spike and slab priors, Statistical Methods in Medical Research 29 (2020) 2445–2469. doi:doi:10.1177/0962280219898497.

[11] G. W. Imbens, Causality in econometrics: Choice vs chance, Econometrica 90 (2022) 2541–2566. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA21204. doi:doi:https://doi.org/10.3982/ECTA21204.

[12] P. R. Rosenbaum, D. B. Rubin, Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, The American Statistician 39 (1985) 33–38. URL: http://www.jstor.org/stable/2683903.

[13] C. M. Crainiceanu, F. Dominici, G. Parmigiani, Adjustment uncertainty in effect estimation, Biometrika 95 (2008) 635–651. URL: http://www.jstor.org/stable/20441491.

[14] X. Zhang, D. E. Faries, H. Li, J. D. Stamey, G. W. Imbens, Addressing unmeasured confounding in comparative observational research, Pharmacoepidemiology and Drug Safety 27 (2018) 373–382. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.4394. doi:doi:https://doi.org/10.1002/pds.4394.

[15] J. O. Berger, Robust Bayesian analysis: sensitivity to the prior, Journal of Statistical Planning and Inference 25 (1990) 303 – 328. doi:doi:10.1016/0378-3758(90)90079-A.

[16] M. Zaffalon, A. Antonucci, R. Cabañas, Structural causal models are (solvable by) credal networks, in: M. Jaeger, T. D. Nielsen (Eds.), Proceedings of the 10th International Conference on Probabilistic Graphical Models, volume 138 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 581–592. URL: https://proceedings.mlr.press/v138/zaffalon20a.html.

[17] I. Raices Cruz, M. C. M. Troffaes, J. Lindström, U. Sahlin, A robust Bayesian bias-adjusted random effects model for consideration of uncertainty about bias terms in evidence synthesis, Statistics in Medicine 41 (2022) 3365–3379. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9422. doi:doi:https://doi.org/10.1002/sim.9422.

[18] T. Basu, M. C. Troffaes, J. Einbeck, A robust bayesian analysis of variable selection under prior ignorance, Sankhya A 85 (2023) 1014–1057.

[19] H. Ishwaran, J. S. Rao, Spike and slab variable selection: Frequentist and Bayesian strategies, Ann. Statist. 33 (2005) 730–773. doi:doi:10.1214/009053604000001147.

[20] X. Xu, M. Ghosh, Bayesian variable selection and estimation for group lasso, Bayesian Analysis 10 (2015) 909–936. URL: https://doi.org/10.1214/14-BA929. doi:doi:10.1214/14-BA929.

[21] P. R. Hahn, C. M. Carvalho, Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective, Journal of the American Statistical Association 110 (2015) 435–448. doi:doi:10.1080/01621459.2014.993077.

[22] J. J. Heckman, R. Robb, Alternative methods for evaluating the impact of interventions: An overview, Journal of Econometrics 30 (1985) 239–267. URL: https://www.sciencedirect.com/science/article/pii/0304407685901393. doi:doi:https://doi.org/10.1016/0304-4076(85)90139-3.

[23] J. H. Albert, S. Chib, Bayesian analysis of binary and polychotomous response data, Journal of the American Statistical Association 88 (1993) 669–679. URL: http://www.jstor.org/stable/2290350.

[24] J. Fan, J. Lv, Sure Independence Screening for Ultrahigh Dimensional Feature Space, Journal of the Royal Statistical Society Series B: Statistical Methodology 70 (2008) 849–911. URL: https://doi.org/10.1111/j.1467-9868.2008.00674.x. doi:doi:10.1111/j.1467-9868.2008.00674.x.

[25] M. M. Barbieri, J. O. Berger, Optimal predictive model selection, The Annals of Statistics 32 (2004) 870 – 897. URL: https://doi.org/10.1214/009053604000000238. doi:doi:10.1214/009053604000000238.

[26] H. Zou, The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association 101 (2006) 1418–1429. doi:doi:10.1198/016214506000000735.

[27] H. Akoglu, User's guide to correlation coefficients, Turkish Journal of Emergency Medicine 18 (2018) 91–93. URL: https://www.sciencedirect.com/science/article/pii/S2452247318302164. doi:doi:https://doi.org/10.1016/j.tjem.2018.08.001.

[28] C. P. Dancey, J. Reidy, Statistics without maths for psychology, Pearson education, 2007.

[29] Y. Chan, Biostatistics 104: correlational analysis, Singapore Med J 44 (2003) 614–619.

[30] M. Plummer, rjags: Bayesian Graphical Models using MCMC, 2023. URL: https://CRAN.R-project.org/package=rjags, R package version 4-15.

[31] M. Zaffalon, G. Corani, D. Mauá, Evaluating credal classifiers by utility-discounted predictive accuracy, International Journal of Approximate Reasoning 53 (2012) 1282–1301. URL: https://www.sciencedirect.com/science/article/pii/S0888613X12000989. doi:doi:https://doi.org/10.1016/j.ijar.2012.06.022, imprecise Probability: Theories and Applications (ISIPTA'11).

[32] I. R. Cruz, J. Lindström, M. C. M. Troffaes, U. Sahlin, Iterative importance sampling with Markov chain Monte Carlo sampling in robust Bayesian analysis, Computational Statistics & Data Analysis 176 (2022) 107558. doi:doi:10.1016/j.csda.2022.107558. arXiv:2206.08728.

[33] G. de Cooman, M. Zaffalon, Updating beliefs with incomplete observations, Artificial Intelligence 159 (2004) 75–125.