ROBUST BAYESIAN ANALYSIS OF CAUSAL INFERENCE PROBLEMS

TATHAGATA BASU[1], MATTHIAS C. M. TROFFAES[2], AND JOCHEN EINBECK[2,3]

Abstract. Causal inference using observational data is an important aspect in many fields such as epidemiology, social science, economics, etc. In particular, our goal is to find the treatment effect on the subjects along with the causal links between the variables and the outcome. However, estimation for such problems are extremely difficult as the treatment effects may vary from subject to subject and modelling the underlying heterogeneity explicitly makes the problem practically unsolvable. Another issue we face is the dimensionality of the problem and we may wish to find a subset of explanatory variables. However, standard variable selection methods tend to maximise the predictive performance of the outcome model only and can also be sensitive with respect to the choice of priors, particularly in the case of limited information. So, in this paper, we suggest a robust Bayesian analysis of causal inferential methods for high-dimensional problems in a regressional framework. We consider a set of spike and slab priors to obtain robust estimates for both the treatment and outcome model. We are specifically interested in the importance of priors in the high dimensional causal inference as well as the identifying the confounder variables. However, indicator based confounder selection can be deceptive in some cases. Especially, when the predictor is strongly associated with either the treatment or the outcome. This increases the posterior expectation of the selection indicators. To avoid that we also apply a post-hoc selection scheme which successfully remove negligible non-zero effects from the model attaining a sparser model. Finally, we illustrate our result using synthetic and real dataset.

1. Introduction

In causal inference, we are interested in estimating the causal effect of independent variables on a dependent variable. Ideally, randomised trials are the most efficient way to perform this task. However, this is not very practical for several reasons; ethical concerns, design cost, population size, to name a few. This leaves us with observational studies which are usually obtained by means collecting data though surveys or record keeping. But this can be problematic in the presence of confounders. That is when the variables are associated with both the treatment and the outcome. In such cases, we need to be extra cautious as otherwise it will lead to unwanted bias in the treatment effect estimator [13]. Several works have been done in order to tackle the presence of confounder variables. One such work in the topic was by Robins [11] where the author used a graphical approach for the identification of the causal parameters. Rosenbaum and Rubin [14] suggested the use of a link model to estimate the propensity scores for all individuals. Later on several other methods have been proposed based on propensity score matching. A brief review on such methods can be found in [18, 16].

[1] UMR CNRS 7253 Heudiasyc, Université de Technologie de Compiègne

[2] Department of Mathematical Sciences, Durham University

[3] Durham Research Methods Centre

E-mail addresses: tathagatabasumaths@gmail.com, matthias.troffaes@durham.ac.uk, jochen.einbeck@durham.ac.uk.

Key words and phrases. high dimensional data; variable selection; Bayesian analysis; imprecise probability.

Bayesian approach in causal effect estimation has become a popular topic in recent days. However, one of the earlier works on this can be found in [15]. Lately with the notion of high dimensional problems, Bayesian methodologies have become more appealing. Crainiceanu et al. [3] proposed a bi-level Bayesian model averaging based method for estimating the causal effect. Wang et al. [17] suggested BAC (or, Bayesian adjustment for confounding) where they use an informative prior obtained from the treatment model and apply them on the outcome model for estimating causal effect. Several other methods were also proposed to tackle confounders from the point of view of Bayesian variable selection such as: Zigler and Dominici [20], Hahn et al. [5] etc.

In this paper we take inspiration from the approach of Koch et al. [9], where they proposed a bilevel spike and slab prior for causal effect estimation. They considered a data-driven adaptive approach to propose their prior which reduce the variance of the causal estimate. In our approach, we suggest a sensitivity analysis based approach where instead of using a single prior, we consider a set of priors [2]. This is particularly interesting as in many cases, causal effect estimation can be performed through a meta analysis and hence robust Bayesian analysis can be beneficial [10] under severe uncertainty. Moreover, for some problems we have to rely on very limited data to perform our Bayesian analysis and choosing a data-driven prior can lead to overfitting. Therefore, as to propose our robust Bayesian framework, we consider a set of continuous spike and slab priors [7] for confounder selection and construct a Bayesian group LASSO [19]. To perform the sensitivity analysis, we consider a set of beta priors on the covariate selection probability of the spike and slab priors. This sensitivity analysis allows us to investigate the regression coefficients with respect to a different level of sparsity of the model. Following that, we consider a post-hoc variable selection method as suggested by Hahn and Carvalho [4]. This ensures that we can efficiently separate the confounder variables with others and reduce the bias of treatment effect estimation.

The rest of the paper is organised as follows. In Section 2 we give a formal description of 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. In Section 4, we provide our result of simulation studies under different scenarios and in data analysis with real data in Section 5. Finally we discuss our findings and conclude this paper in Section 6.

2. Causal Estimation

Let an observational study give us the outcomes $Y = (Y_1, \ldots, Y_n)$ along with corresponding treatment indicators $T = (T_1, \ldots, T_n)$. Then the treatment effect in the population is given by the expectation of the difference in outcome between the treatment and controls.

$$\delta = \mathbb{E}(Y \mid T = 1) - E(Y \mid T = 0). \tag{1}$$

Similarly, individual causal effect of the treatment T_i on outcome Y_i is given by:

$$\delta_i \coloneqq (Y_i \mid T_i = 1) - (Y_i \mid T_i = 0). \tag{2}$$

That is, the difference between the outcome when i-th subject receives the treatment and when i-th subject remain as a control.

In theory, both of these quantities exist. However, we can not observe $(Y_i \mid T_i = 1)$ and $(Y_i \mid T_i = 0)$ simultaneously for the i-th individual. Instead, we can estimate average causal effect of the treatment T by calculating the averaged outcome of all the subjects those received the treatment and all the subjects those remained as control.

$$\hat{\delta} \coloneqq \frac{\sum_{i=1}^{n} Y_i \cdot \mathbb{I}(T_i = 1) - \sum_{i=1}^{n} Y_i \cdot \mathbb{I}(T_i = 0)}{n}. \tag{3}$$

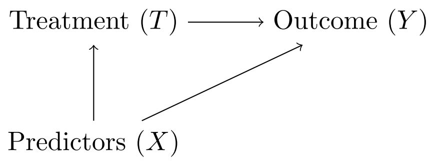

Figure 1. Confounding in causal models.

However, this relies on an important assumption that the treatment effect on the i-th subject given that they received the treatment is equal to the treatment effect when they remain as the control [18].

2.1. **Regression Model.** 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 [18, 6]. To do so, we rely on p different observed quantities or predictors denoted by $X := [X_1, \ldots, X_n]^T$. Now, let $\beta := (\beta_1, \ldots, \beta_p)$ denotes the vector of regression coefficients. Then we can define a linear model for the outcome so that

$$Y_i = \beta_T T_i + \beta_0 + X_i\beta + \epsilon_i \tag{4}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Clearly, when the underlying true outcome model is linear,

$$\delta = \beta_T \quad \text{and hence } \hat{\delta} - \delta = \hat{\beta}_T - \beta_T. \tag{5}$$

In the presence of confounders we also need to consider the association between the treatment indicators and the predictors. Koch et al. [9] suggested the use of a probit link function to construct the regression model. 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 $\gamma := (\gamma_1, \cdots, \gamma_p)$ we define

$$P(T_i = 1 \mid X_i) := \Phi(\gamma_0 + X_i\gamma) \tag{6}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution.

This allows us to define intermediate latent variables as suggested by Albert and Chib [1]

$$T_i^* = \gamma_0 + X_i\gamma + u_i \tag{7}$$

where, $u_i \sim \mathcal{N}(0, 1)$. Therefore, $T_i = 1$ if $T_i^* > 0$ and $T_i = 0$ if $T_i^* \leq 0$.

Now, following the approach of Koch et al. [8], we define an adjusted output vector $W := (Y, T^*)^T$ and corresponding $2n \times (2p + 3)$ dimensional design matrix

$$Z = \begin{bmatrix} X_O & 0 \\ 0 & X_T \end{bmatrix}, \tag{8}$$

where, $X_O = [T, 1_n, X]$ and $X_T = [1_n, X]$. Then, assuming a Gaussian error term, we have the following likelihood distribution

$$W \mid Z, \alpha, \beta, \gamma, \sigma^2 \sim \mathcal{N}\left(Z\nu, \Sigma\right), \tag{9}$$

where $\nu = (\beta_T, \beta_0, \beta, \gamma_0, \gamma)^T$ and

$$\Sigma = \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & I_n \end{bmatrix}. \tag{10}$$

3. Bayesian Causal Estimation

The likelihood formation given by Eq. (9) gives us a foundation for Bayesian group LASSO [19] type model and look into the posterior selection probability associated with the j-th predictor. There are several ways to construct spike and slab priors which achieve variable selection. In our case, we consider a continuous type [7] prior for faster posterior computation.

3.1. **Hierarchical model.**

Let, π_j denote the prior probability that the j-th predictor is not associated with either the outcome or the treatment. That is,

$$\pi_j = P\left((\beta_j, \gamma_j) = (0,0)\right). \tag{11}$$

Then we can define a spike and slab group LASSO so that: for $1 \leq j \leq p$,

$$(\beta_j, \gamma_j)^T \mid \pi_j, \sigma^2 \sim (1-\pi_j)\mathcal{N}\left((0,0)^T, \tau_1^2 \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix}\right) + \pi_j \mathcal{N}\left(0, \tau_0^2 \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{12}$$

$$\sigma^2 \sim \text{InvGamma}(a, b) \tag{13}$$

$$\pi_j \sim \text{Beta}\left(sq_j, s(1-q_j)\right). \tag{14}$$

We fix sufficiently small τ_0^2 $(1 \gg \tau_0^2 > 0)$ so that $(\beta_j, \gamma_j) = (0,0)$ has its probability mass concentrated around zero. Therefore, this represents the spike component of our prior specification. To construct the slab component, we consider τ_1^2 to be large so that $\tau_1^2 \geq 1$. This allows the prior for $(\beta_j, \gamma_j) \neq (0,0)$ to be flat. Inverse-Gamma is a natural choice for the variance of the Gaussian noise because of conjugacy. To make the prior flat, we consider $1 \gg a, b > 0$.

As indicated earlier, π_j is used as the rejection probability of the j-th predictor and we use a beta prior to specify these rejection probabilities where q_j represents our prior expectation of the rejection probability (π_j) and 's' acts as a concentration parameter. For the intercept terms of the outcome model and the causal effect, we consider a sufficiently flat prior so that $\beta_0.\beta_T \sim \mathcal{N}(0, \sigma^2)$. Similarly, for the intercept term in the treatment model, we consider $\gamma_0 \sim \mathcal{N}(0,1)$.

In Fig. 2, we show a probabilistic graphical representation of our hierarchical model. In the figure, grey circular nodes represent the prior hyper-parameters 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 we they are related through the latent variable T^*. This dashed edge also establish the notion of confounding as shown in Fig. 1.

3.2. **Variable selection and coefficient adjustment.**

For the co-variate selection, we look into the posterior expectation of π_j. We consider the j-th predictor to be removed from both the treatment and outcome model, if

$$\underline{\mathrm{E}}(\pi_j \mid W) \coloneqq \inf_{q_j \in \mathcal{P}_j} \mathrm{E}(\pi_j \mid W) > 1/2. \tag{15}$$

For the rest of the variables, some of them will be present in the model as confounders and some will only be associated with either the treatment. Let $\mathcal{S}$ denote the set of predictors such that,

$$\mathcal{S} \coloneqq \left\{j : \underline{\mathrm{E}}(\pi_j \mid W) < 1/2\right\}. \tag{16}$$

That is the set that contains all the variables which are not removed from both treatment and outcome model. Now, for each fixed value of q, let $\hat{\beta}_{\mathcal{S}}(q_{\mathcal{S}})$ be the posterior means of the regression coefficients of the outcome model with respect to the predictors that belong to $\mathcal{S}$. Similarly, $\hat{\gamma}_{\mathcal{S}}(q_{\mathcal{S}})$

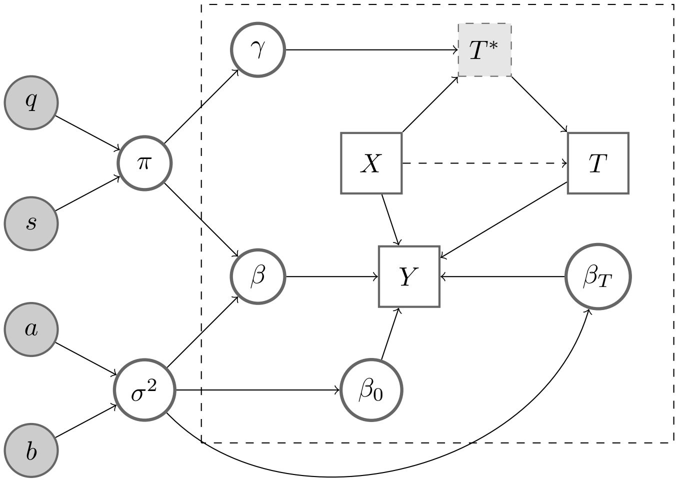

FIGURE 2. Probabilistic graphical representation for causal inference with Bayesian hierarchical model.

be the posterior means of the regression coefficients for the treatment effects. Since, we use a continuous type selection prior, these regression coefficients are non-zero in nature. Therefore, to adjust the sparsity, we apply the "decoupled shrinkage and selection" method proposed by Hahn and Carvalho [4]. To do so, we solve the following adaptive LASSO-type [21] problems

$$\hat{\beta}^*_{\mathcal{S}}(q) = \arg\min_{\beta_{\mathcal{S}}} \frac{1}{n}\|X_{\mathcal{S}}\hat{\beta}_{\mathcal{S}}(q) - X_{\mathcal{S}}\beta_{\mathcal{S}}\|_2^2 + \lambda \sum_{j\in\mathcal{S}} \frac{|\beta_j|}{|\hat{\beta}_j(q_j)|} \tag{17}$$

and

$$\hat{\gamma}^*_{\mathcal{S}}(q) = \arg\min_{\gamma_{\mathcal{S}}} \frac{1}{n}\|X_{\mathcal{S}}\hat{\gamma}_{\mathcal{S}}(q) - X_{\mathcal{S}}\gamma_{\mathcal{S}}\|_2^2 + \lambda \sum_{j\in\mathcal{S}} \frac{|\gamma_j|}{|\hat{\gamma}_j(q_j)|} \tag{18}$$

where $q_j \in \mathcal{P}_j$ for all $j \in \mathcal{S}$.

3.3. Robust Bayesian Analysis.

We perform our robust Bayesian analysis on $q \coloneqq (q_1, \ldots, q_p) \in \mathcal{P}$, where

$$\mathcal{P} \coloneqq \mathcal{P}_1 \times \cdots \times \mathcal{P}_p \subseteq (0,1)^p \, . \tag{19}$$

4. SIMULATION STUDIES

5. DATA ANALYSIS

6. CONCLUSION

REFERENCES

[1] Albert, J. H. and Chib, S. [1993], 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association* **88**(422), 669–679.

URL: *http://www.jstor.org/stable/2290350*

[2] Berger, J. O. [1990], 'Robust bayesian analysis: sensitivity to the prior', *Journal of Statistical Planning and Inference* **25**(3), 303 – 328.

[3] Crainiceanu, C. M., Dominici, F. and Parmigiani, G. [2008], 'Adjustment uncertainty in effect estimation', *Biometrika* **95**(3), 635–651.
URL: *http://www.jstor.org/stable/20441491*

[4] Hahn, P. R. and Carvalho, C. M. [2015], 'Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective', *Journal of the American Statistical Association* **110**(509), 435–448.

[5] Hahn, P. R., Carvalho, C. M., Puelz, D. and He, J. [2018], 'Regularization and Confounding in Linear Regression for Treatment Effect Estimation', *Bayesian Analysis* **13**(1), 163 – 182.
URL: *https://doi.org/10.1214/16-BA1044*

[6] Heckman, J. J. and Robb, R. [1985], 'Alternative methods for evaluating the impact of interventions: An overview', *Journal of Econometrics* **30**(1), 239–267.
URL: *https://www.sciencedirect.com/science/article/pii/0304407685901393*

[7] Ishwaran, H. and Rao, J. S. [2005], 'Spike and slab variable selection: Frequentist and bayesian strategies', *Ann. Statist.* **33**(2), 730–773.

[8] Koch, B., Vock, D. M. and Wolfson, J. [2018], 'Covariate selection with group lasso and doubly robust estimation of causal effects', *Biometrics* **74**(1), 8–17.

[9] Koch, B., Vock, D. M., Wolfson, J. and Vock, L. B. [2020], 'Variable selection and estimation in causal inference using bayesian spike and slab priors', *Statistical Methods in Medical Research* **29**(9), 2445–2469.

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

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

[12] Robinson, P. M. [1988], 'Root-n-consistent semiparametric regression', *Econometrica* **56**(4), 931–954.
URL: *http://www.jstor.org/stable/1912705*

[13] ROSENBAUM, P. R. and RUBIN, D. B. [1983], 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.
URL: *https://doi.org/10.1093/biomet/70.1.41*

[14] Rosenbaum, P. R. and Rubin, D. B. [1985], 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score', *The American Statistician* **39**(1), 33–38.
URL: *http://www.jstor.org/stable/2683903*

[15] Rubin, D. B. [1978], 'Bayesian Inference for Causal Effects: The Role of Randomization', *The Annals of Statistics* **6**(1), 34 – 58.
URL: *https://doi.org/10.1214/aos/1176344064*

[16] Stuart, E. A. [2010], 'Matching Methods for Causal Inference: A Review and a Look Forward', *Statistical Science* **25**(1), 1 – 21.
URL: *https://doi.org/10.1214/09-STS313*

[17] Wang, C., Dominici, F., Parmigiani, G. and Zigler, C. M. [2015], 'Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models', *Biometrics* **71**(3), 654–665.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12315*

[18] Winship, C. and Morgan, S. L. [1999], 'The estimation of causal effects from observational data', *Annual Review of Sociology* **25**(1), 659–706.

[19] Xu, X. and Ghosh, M. [2015], 'Bayesian Variable Selection and Estimation for Group Lasso', *Bayesian Analysis* **10**(4), 909 – 936.
URL: *https://doi.org/10.1214/14-BA929*

[20] Zigler, C. M. and Dominici, F. [2014], 'Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects', *Journal of the American Statistical Association* **109**(505), 95–107.
URL: *http://www.jstor.org/stable/24247140*

[21] Zou, H. [2006], 'The Adaptive Lasso and Its Oracle Properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.