Bayesian Model Averaging for Causality Estimation and its Approximation based on Gaussian Scale Mixture Distributions

Shunsuke Horii
Waseda University

Abstract

In the estimation of the causal effect under linear Structural Causal Models (SCMs), it is common practice to first identify the causal structure, estimate the probability distributions, and then calculate the causal effect. However, if the goal is to estimate the causal effect, it is not necessary to fix a single causal structure or probability distributions. In this paper, we first show from a Bayesian perspective that it is Bayes optimal to weight (average) the causal effects estimated under each model rather than estimating the causal effect under a fixed single model. This idea is also known as Bayesian model averaging. Although the Bayesian model averaging is optimal, as the number of candidate models increases, the weighting calculations become computationally hard. We develop an approximation to the Bayes optimal estimator by using Gaussian scale mixture distributions.

1 Introduction

Research on causal inference to examine the magnitude of the effect of the treatment variable on the outcome variable is one of the important tasks in data science. Fisher's randomized controlled trial is one of the most important methods to examine the causal effect and is often considered as the gold standard for causal inference [Fisher, 1951]. However, complete randomization is often impracticable due to cost or ethical reasons, demanding methods for identifying causal effects for non-experimental observational studies.

To identify causal effects in observational studies, it is necessary to make some assumptions on the data generating process. There are various interrelated approaches to statistically estimate causal effects, including those based on propensity scores [Guo and Fraser, 2014], those based on instrumental variables [Angrist et al., 1996], and those based on Structural Causal Model (SCM) [Pearl, 2000]. In this study, we focus on the causality estimation based on SCM. In SCM, we are interested in examining the causal effect of a treatment variable X on an outcome variable Y, which is written as an experimental distribution $p(y|\mathrm{do}(X=x))$, where $\mathrm{do}(X=x)$ means that X is fixed to x by an intervention. The causal effect $p(y|\mathrm{do}(X=x))$ is calculated based on the knowledge on a causal graph G, which describes a coarse relationship among the observable variables, and a distribution $P(X,Y,Z_1,\ldots,Z_c)$, where $Z_1,\ldots,Z_c$ are observable variables (covariates) other than X and Y.

In general, we do not know the true causal graph G, so there are various methods for estimating the causal graph from the data [Spirtes and Glymour, 1991, Cooper and Herskovits, 1992, Heckerman et al., 1995, Shimizu et al., 2006]. Furthermore, to calculate the causal effect, it has to estimate the conditional distributions among the variables. Therefore, a general way to estimate the causal effect consits of the following steps.

1. Estimate the causal graph from the data
2. Estimate the conditional distributions among the variables from the data
3. Calculate the causal effect

However, if the goal is to estimate the causal effect, it is not necessary to fix a single causal graph or distributions. We first derive the Bayesian optimal estimator when the problem is to estimate the mean causal effect under the squared error loss. The Bayes optimal solution turns out to be the estimator that weights the mean causal effects estimated under each model, which is also known as Bayesian model

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

averaging [Hoeting et al., 1999]. In the literature of causal inference, Bayesian model averaging is applied for propensity score analysis [Kaplan and Chen, 2014], however, our study is an attempt to apply it for the causal inference based on SCM. The basic idea behind our proposal is to consider the causal graph and the parameters of the distributions as parameters to be marginalized in estimating the causal effect. Although the model setting is different, a similar idea can be found in [Rubin, 1978], where the values of the unobserved potential outcomes are treated as parameters to be marginalized.

Although the Bayesian model averaging is optimal, as the number of candidate models increases, the weighting calculations become computationally hard. In this paper, we develop an approximation algorithm for the case where the following assumptions hold.

1. We know the set of possible directed edges, including information on the directions of the edges.
2. We do not know whether each element of the above set of possible directed edges exists. We only know the prior probability that each directed edge exists.

The above assumptions make sense in some practical applications. For example, consider the case where there are m variables $X_1, \ldots, X_m$ and only the antecedent relations, such as nodes with higher index numbers do not precede nodes with lower index numbers, to the causal relations among these variables are known. This is the situation dealt with in [Wermuth and Lauritzen, 1982]. In this case, the set of possible directed edges are given by $\{(i,j) : i < j, 1 \leq i, j \leq m\}$.

Even if we put the constraint above, the number of models grows exponentially with the number of possible directed edges. We develop an approximation to the Bayes optimal estimator by using sparse modeling techniques. The basic idea is to approximate a mixture distribution of Dirac delta function and Gaussian distribution with Gaussian scale mixture (GSM). In the prediction task, it is reported that performance similar to Bayesian model averaging can be achieved by using horseshoe prior, which is a kind of GSM [Carvalho et al., 2009]. This study shows that a similar approach is also effective in estimating causal effects.

The paper is organized as follows. In Section 2, some preliminary materials about SCM and corresponding causality notions such as the intervention effect and the mean intervention effect are described. In Section 3, we formulate the problem to estimate the mean intervention effect as a statistical decision problem and derive the optimal decision function under the Bayes criterion. Section 4 describes the further assumptions for the causal graph in this study and derives an approximation algorithm for the Bayes optimal decision function under these assumptions. Section 5 presents some experimental results to evaluate our proposals. Finally, we give a summary in Section 6.

2 Preliminaries

As mentioned in the introduction, there are some approaches to statistically estimate causal effects. Since this study follows the framework of SCM, we borrow some basic notions of SCM.

Definition 1. *Let G be a directed acyclic graph (DAG) and $V = (X_1, X_2, \ldots, X_m)$ be a set of random variables that corresponds to the set of the vertices of G. We sometimes write E_G as the set of directed edges of G and G_E as the DAG whose set of edges is E. The DAG G is called a causal graph if it specifies the causal relationships among variables in the following form,*

$$X_i = g_i(\mathrm{pa}(X_i), \epsilon_i), \quad i = 1, \ldots, m, \tag{1}$$

where $\mathrm{pa}(X_i) \subset V$ is the set of variables that have a directed edge that heads to X_i and ϵ_i is an error term[1]. In this study, we assume that ϵ_i follows the Gaussian distribution $\mathcal{N}(0, s_\epsilon^{-1})$. The equations (1) are called structural equations for $X_1, X_2, \ldots, X_m$.

When the functions $g_i, i = 1, \ldots, m$ are linear, i.e.,

$$X_i = \sum_{X_j \in \mathrm{pa}(X_i)} \theta_{X_j X_i} X_j + \epsilon_i, \quad i = 1, \ldots, m, \tag{2}$$

the model is called linear SCM. If there is no confusion, $\theta_{X_i X_j}$ is written as θ_{ij} for short.

One may think that the parametric assumption among the variables is too strong, however, a theory built on a simple model will be the foundations for a theory in more complex models, and as we will show in later experiments, even such a simple model works for some real-world problems.

Structural equations and causal graphs express coarse causal relationships among the variables. Given these information, we want to know the causal effect of a treatment variable $X \in V$ on an outcome variable $Y \in V$ when X is fixed to a value x by an external intervention. It is mathematically defined as follows [Pearl, 2000].

Definition 2. *Let $V = \{X, Y, Z_1, \ldots, Z_c\}$ be the set of vertices of a causal graph G. The causal (intervention) effect on Y when X is fixed to x by an external*

[1] Unless it causes a confusion, $\mathrm{pa}(X)$ is also written as $\mathrm{pa}(x)$. They sometimes denote the indices of the nodes.

intervention is defined as

$$p(y|\mathrm{do}(X=x)) = \int \cdots \int \frac{p(x,y,z_1,\ldots,z_c)}{p(x|\mathrm{pa}(x))}\mathrm{d}z_1\ldots\mathrm{d}z_c, \quad (3)$$

where $\mathrm{do}(X=x)$ *means that* X *is fixed to* x *by an intervention.*

The intervention effect $p(y|\mathrm{do}(X=x))$ is defined as an experimental distribution. In this study, we focus on the estimation of their mean (expectation), the Mean Intervention Effect (MIE). The MIE $\bar{y}_x$ is defined as

$$\bar{y}_x = \int y \cdot p(y|\mathrm{do}(X=x))\mathrm{d}y. \quad (4)$$

When the model is linear SCM, it is known that the MIE is expressed as

$$\bar{y}_x = \left(\sum_{l\in\mathcal{P}}\prod_{(i,j)\in l}\theta_{ij}\right)x, \quad (5)$$

where $\mathcal{P}$ is the set of the directed paths from X to Y [Pearl, 2000]. This is an equivalent notion of the total effect in [Wright, 1921].

We note that the main interest of the existing studies in the literature is the identifiability of the total effect. For example, Pearl proved a valuable result that claims that we can identify the total effect if we can observe some covariates that satisfy the backdoor criterion [Pearl, 2000]. On the other hand, our focus is how to estimate the total effect. Although it would be interesting to combine our results with the existing results, we do not use those results in this paper.

3 Bayesian estimation of the mean intervention effect

In order to calculate the MIE (5), one has to know the underlying causal graph G and conditional distributions $p(x_i|\mathrm{pa}(x_i)), i=1,\ldots,m$. Therefore, the data analyst must either proceed with the belief that the assumed causal graph is correct, or estimate the causal graph from the data. Although there are various methods for estimating the causal graph from the data [Spirtes and Glymour, 1991, Cooper and Herskovits, 1992, Heckerman et al., 1995, Shimizu et al., 2006], there is a possibility that these methods output a wrong causal graph. Furthermore, previous studies are rarely concerned about how to estimate the probability distributions $p(x_i|\mathrm{pa}(x_i)), i=1,\ldots,m$ because their main concern is about the identifiability of the causal effects given the causal graph and probability distributions. However, it is uncommon to assume that the probability distribution is known without knowing the data generating causal graph. In our setting, we need to deal with the estimation of the probability distributions as well as the estimation of the causal graph.

We assume that the causal diagram G is a random variable that takes its value in the set of DAGs $\mathcal{G}$ and whose prior distribution is $p(G)$. This prior distribution represents, for example, the data analyst's confidence in each causal graph. Since we deal with the linear SCM, given a causal graph G, conditional distributions $p(x_i|\mathrm{pa}(x_i))$ are parameterized by $\theta_{ij}, (i,j)\in E_G$ as follows.

$$p(x_i|\mathrm{pa}(x_i)) = \mathcal{N}\left(\sum_{j:(j,i)\in E_G}\theta_{ji}x_j, s_\epsilon^{-1}\right). \quad (6)$$

Throughout the paper, we assume that the precision parameter s_ϵ of error is known. We can extend our results to the unknown case by assuming a prior distribution for s_ϵ. Let $\boldsymbol{\theta}_G = (\theta_{ij} : (i,j)\in E_G)$. We assume that the parameter $\boldsymbol{\theta}_G$ is also a random vector whose conditional distribution under G is $p(\boldsymbol{\theta}_G|G)$. Since the MIE is the function of G and $\boldsymbol{\theta}_G$, we write it as $\bar{y}_x(G,\boldsymbol{\theta}_G)$.

Let $D^N = (x_n, y_n, z_{1,n},\ldots,z_{c,n})_{n=1,\ldots,N}$ be a sample of $V=(X,Y,Z_1,\ldots,Z_c)$ with sample size N. [2] We consider the problem of estimating the MIE (5) given D^N. Let $d: D^N \mapsto \mathbb{R}$ be a decision function that outputs an estimate of the MIE. The loss function is the metric between the estimand and the decision function. In this study, the squared error loss is used, i.e.,

$$\ell(G,\boldsymbol{\theta}_G,d(D^N)) = \left(\bar{y}_x(G,\boldsymbol{\theta}_G)-d(D^N)\right)^2. \quad (7)$$

In the statistical decision theory framework [Berger, 2013], the risk function and the Bayes risk function are defined as follows, respectively.

$$R(G,\boldsymbol{\theta}_G,d) = \mathrm{E}_{D^N|G,\boldsymbol{\theta}_G}\left[\ell(G,\boldsymbol{\theta}_G,d(D^N))\right], \quad (8)$$

$$BR(d) = \mathrm{E}_G\left[\mathrm{E}_{\boldsymbol{\theta}_G|G}R(G,\boldsymbol{\theta}_G,d)\right]. \quad (9)$$

The Bayes optimal estimator, that minimizes the Bayes risk function, is given as follows.

Theorem 1. *When the loss function is the squared loss (7), the Bayes optimal estimator of the MIE is given by*

$$d^*(D^N) = \sum_{G\in\mathcal{G}} p(G|D^N)\int \bar{y}_x(G,\boldsymbol{\theta}_G)p(\boldsymbol{\theta}_G|G,D^N)\mathrm{d}\boldsymbol{\theta}_G, \quad (10)$$

[2] We sometimes do not distinguish X and Y from other covariates $Z_1,\ldots,Z_c$. In that case, we write the set of variables as $V=(X_1,X_2,\ldots,X_m)$.

where

$$p(G|D^N) = \frac{p(D^N|G)p(G)}{\sum_{G\in\mathcal{G}} p(D^N|G)p(G)}, \tag{11}$$

$$p(D^N|G) = \int p(D^N|G,\boldsymbol{\theta}_G)p(\boldsymbol{\theta}_G|G)\mathrm{d}\boldsymbol{\theta}_G, \tag{12}$$

$$p(\boldsymbol{\theta}_G|G,D^N) = \frac{p(D^N|G,\boldsymbol{\theta}_G)p(\boldsymbol{\theta}_G|G)}{\int p(D^N|G,\boldsymbol{\theta}_G)p(\boldsymbol{\theta}_G|G)\mathrm{d}\boldsymbol{\theta}_G}. \tag{13}$$

Proof. It is known that the decision function that minimizes the loss function weighted by the posterior distribution is Bayes optimal [Berger, 2013]. That is,

$$d^*(D^N) = \arg\min_d \int \left\{ \left(\bar{y}_x(G,\boldsymbol{\theta}_G) - d(D^N)\right)^2 \times \sum_{G\in\mathcal{G}} p(G|D^N)p(\boldsymbol{\theta}_G|G,D^N) \right\} \mathrm{d}\boldsymbol{\theta}_G. \tag{14}$$

The solution of this minimization problem is given by (10). □

Note that the Bayes optimal estimator of the intervention effect under the Kullback-Leibler loss is given in [Horii and Suko, 2019].

In general, numerical integration is required to calculate the integrals in (10) since $\bar{y}_x(G,\boldsymbol{\theta}_G)$ is a nonlinear function of $\boldsymbol{\theta}_G$. When the computational complexity of the numerical integration is large, we approximate (10) by

$$\tilde{d}^*(D^N) = \sum_{G\in\mathcal{G}} p(G|D^N)\bar{y}_x(G,\boldsymbol{\theta}_G^{MAP}), \tag{15}$$

$$\boldsymbol{\theta}_G^{MAP} = \arg\max_{\boldsymbol{\theta}_G} p(\boldsymbol{\theta}_G|G,D^N). \tag{16}$$

This approximation is based on the property that the posterior distribution $p(\boldsymbol{\theta}_G|G,D^N)$ is asymptotically concentrated around $\boldsymbol{\theta}_G^{MAP}$ under some appropriate conditions [Le Cam, 2012]. To keep the description concise, we call the estimator (15) Bayes quasi-optimal estimator.

If the prior distribution $p(\boldsymbol{\theta}_G|G)$ is a product of Gaussian, namely,

$$p(\boldsymbol{\theta}_G|G) = \prod_{(i,j)\in E_G} \mathcal{N}(\theta_{ij};0,\tau), \tag{17}$$

we can analytically calculate $p(G|D^N)$ and $p(\boldsymbol{\theta}_G|G,D^N)$. See the supplementary material for the derivation.

4 Approximate Bayes Optimal Estimator

The problem with calculating (10) or (15) in practical applications is that its computational complexity is proportional to the number of candidate models. The number of possible DAGs for a given set of variables $V = (X_1, X_2, \ldots, X_m)$ is $O\left(2^{m^2}\right)$. However, it would be uncommon to know nothing at all about the structure of a graph, and we might know the causal relationships among some variables and not the rest. Therefore, in this study, we classify the directed edges expressing the causal relationships among variables into three types.

- It is known that there is a causal relationship between variables, including information on which is the cause and which is the result. In other words, there is a directed edge between the corresponding variables with probability 1.
- It is known that there is no causal relationship between the corresponding variables. In other words, there is directed edge between the corresponding variables with probability 0.
- If there is a causal relationship, it is known that which is the cause and which is the result, but it is not clear whether the causal relationship exists. In other words, there is a directed edge between the corresponding variables with some probability.

For the sake of simplicity, we assume that the set of the edges of the first type is empty, since they can be considered as the edges of the third type with edge existence probability set to 1.

Let E_{full} be a set of possible directed edges connecting nodes that may or may not have a causal relationship between the corresponding variables. We assume that $(j,i) \notin E_{\text{full}}$ if $(i,j) \in E_{\text{full}}$. The set $\mathcal{G}$ of the candidate causal graphs is defined as

$$\mathcal{G} = \left\{ G \ : \ E_G \subseteq E_{\text{full}} \right\}. \tag{18}$$

Let G_{full} be the abbreviation for $G_{E_{\text{full}}}$. Figure 1 depicts an example of G_{full} and $\mathcal{G}$.

We assume that there is a causal relationship between X_i and X_j for $(i,j) \in E_{\text{full}}$ with probability p, that is, a directed edge (i,j) exists with probability p. Then, the prior distribution $p(G)$ is given by

$$p(G) = p^{|E_G|}(1-p)^{|E_{\text{full}} \setminus E_G|}, \quad \forall G \in \mathcal{G}. \tag{19}$$

Even if the set of candidate models is restricted to (18), the number of candidate models $|\mathcal{G}|$ is $2^{|E_{\text{full}}|}$, and

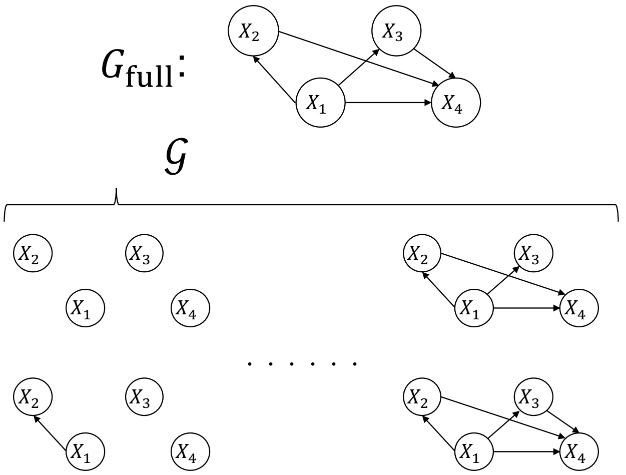

Figure 1: An example of G_{full} and $\mathcal{G}$. In this case, the number of candidate causal graphs $|\mathcal{G}|$ is $2^5 = 32$.

when $|E_{\mathrm{full}}|$ is large, the calculation of (10) becomes infeasible.

For G_{full}, consider the following experimental prior distribution.

$$p_{ss}(\boldsymbol{\theta}_{G_{\mathrm{full}}}) = \prod_{(i,j)\in E_{\mathrm{full}}} p_{ss}(\theta_{ij}), \tag{20}$$

$$p_{ss}(\theta_{ij}) = (1-p)\delta_0(\theta_{ij}) + p\mathcal{N}(\theta_{ij}; 0, \tau)), \tag{21}$$

where $\delta_0(\cdot)$ is the Dirac delta function. Such priors are often called spike-and-slab priors [Ishwaran et al., 2005]. Then, the following proposition holds.

Proposition 1. *The prior distribution of $\theta_{ij}, (i,j) \in E_{\mathrm{full}}$ is the same when (20) and (21) are assumed for $p(\boldsymbol{\theta}_{G_{\mathrm{full}}})$ and (17) and (19) are assumed for $p(\boldsymbol{\theta}_G|G)$ and $p(G)$. That is,*

$$p_{ss}(\boldsymbol{\theta}_{G_{\mathrm{full}}}) = \sum_{G\in\mathcal{G}} p(\boldsymbol{\theta}_G|G)p(G). \tag{22}$$

Proof. For $(i,j) \in E_{\mathrm{full}}$,

$$\sum_{G\in\mathcal{G}} p(\theta_{ij}|G)p(G) = \\ \sum_{G:(i,j)\notin E_G} \delta_0(\theta_{ij})p(G) + \sum_{G:(i,j)\in E_G} \mathcal{N}(\theta_{ij}; 0, \tau)p(G) \tag{23}$$

Then, $\sum_{G:(i,j)\in E_G} p(G)$ is the probability that a graph G has the edge (i,j) and it is given by p. Similarly, $\sum_{G:(i,j)\notin E_G} p(G) = 1 - p$. □

When (20) and (21) are assumed, the Bayes optimal estimator can be rewritten as

$$\int \bar{y}_x(G_{\mathrm{full}}, \boldsymbol{\theta}_{G_{\mathrm{full}}}) p_{ss}(\boldsymbol{\theta}_{G_{\mathrm{full}}}|D^N)\mathrm{d}\boldsymbol{\theta}_{G_{\mathrm{full}}}, \tag{24}$$

where

$$p_{ss}(\boldsymbol{\theta}_{G_{\mathrm{full}}}|D^N) = \frac{p(D^N|\boldsymbol{\theta}_{G_{\mathrm{full}}})p_{ss}(\boldsymbol{\theta}_{G_{\mathrm{full}}})}{\int p(D^N|\boldsymbol{\theta}_{G_{\mathrm{full}}})p_{ss}(\boldsymbol{\theta}_{G_{\mathrm{full}}})\mathrm{d}\boldsymbol{\theta}_{G_{\mathrm{full}}}}. \tag{25}$$

With the introduction of the spike-and-slab prior, the problem of calculating the summation over the candidate models has apparently disappeared, but the calculation of (24) is still difficult because the calculation of the posterior distribution (25) is difficult.

We approximate $p_{ss}(\theta_{ij})$ by Gaussian scale mixture (GSM) [Andrews and Mallows, 1974, Boris Choy and Chan, 2008]. The probability density function (pdf) of GSM is given by

$$p_{gs}(\theta_{ij}) = \int \mathcal{N}(\theta_{ij}; 0, \tau_{ij})p(\tau_{ij}; \boldsymbol{\alpha})\mathrm{d}\tau_{\mathrm{ij}}, \tag{26}$$

where $\boldsymbol{\alpha}$ is the parameter of the distribution $p(\boldsymbol{\tau}; \boldsymbol{\alpha})$. Note that the values of τ_{ij} are different for different $(i,j) \in E_{\mathrm{full}}$. If τ_{ij} is small, it means that the probability that θ_{ij} takes a value close to 0 is large, and we want to estimate them from the data. The distribution $p(\tau_{ij}; \boldsymbol{\alpha})$ is a distribution of the variance of Gaussian and it is often called mixing distribution. It is known that GSM can express various distributions, such as Laplace distribution, Student-t distribution and horseshoe distribution, by changing the mixing distribution. GSM is often used in the Bayesian sparse modeling literature. See [Ji et al., 2008], for example.

Even if we use GSM for the prior distribution $p(\theta_{ij})$, it is still difficult to calculate the exact posterior, however, there are various approximation algorithms to efficiently calculate the approximate posterior which are based on Expectation-Maximization (EM) algorithm [Figueiredo, 2003], Markov-Chain Monte-Carlo (MCMC) algorithm [Park and Casella, 2008], and Variational Bayes (VB) algorithm [Babacan et al., 2014]. As an example, we give an estimation algorithm based on the variational Bayes algorithm. See the supplementary material for the derivation of the algorithm.

Let $\boldsymbol{\theta}_j = (\theta_{ij} : (i,j) \in E_{\mathrm{full}})$ and we describe an estimation algorithm for $\boldsymbol{\theta}_j$ since we can estimate $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ independently. We use the exponential distribution as the mixing distribution, that is,

$$p(\tau_{ij}|\alpha_{ij}) = \begin{cases} \alpha_{ij}e^{-\alpha_{ij}\tau_{ij}} & \tau_{ij} \geq 0, \\ 0 & \tau_{ij} < 0. \end{cases} \tag{27}$$

We have to determine or estimate the values of α_{ij}. We take a Bayesian hierarchical modeling approach, that is, we further assume gamma distribution for α_{ij} which is the conjugate distribution of the exponential distribution.

$$p(\alpha_{ij};\kappa,\nu) = \frac{\nu^{\kappa}}{\Gamma(\kappa)}\alpha_{ij}^{\kappa-1}e^{-\nu\alpha_{ij}}, \tag{28}$$

where $\Gamma(\cdot)$ is the gamma function and κ, ν are hyperparameters. By setting κ, ν so that $p(\alpha_{ij};\kappa,\nu)$ is flat, the algorithm can estimate α_{ij} as well as other parameters. The update equations for the variational Bayes algorithm are summarized as follows.

- Update equation for $\boldsymbol{\theta}_j$

  $$\bar{\boldsymbol{\theta}}_j^{(t+1)} = s_\epsilon \boldsymbol{\Sigma}_j^{(t+1)} \boldsymbol{X}_j^T \boldsymbol{x}_j, \tag{29}$$

  $$\boldsymbol{\Sigma}_j^{(t+1)} = \left(s_\epsilon \boldsymbol{X}_j^T \boldsymbol{X}_j + \bar{\boldsymbol{S}}_j^{(t)}\right)^{-1}, \tag{30}$$

  where

  $$\bar{\boldsymbol{S}}_j^{(t)} = \mathrm{diag}\left(\bar{s}_{j,1}^{(t)}, \ldots, \bar{s}_{j,m_j}^{(t)}\right), \tag{31}$$

  and $\mathrm{diag}(\boldsymbol{a})$ is the diagonal matrix whose diagonal elements are $\boldsymbol{a}$.

- Update equation for $\{\tau_{j,i}\}$

  $$\bar{\tau}_{j,i}^{(t+1)} = \frac{1 + \sqrt{\bar{\alpha}_{j,i}^{(t)}\left((\bar{\theta}_{j,i}^{(t+1)})^2 + \Sigma_{j,ii}^{(t+1)}\right)}}{\bar{\alpha}_{j,i}^{(t)}}, \tag{32}$$

  $$\bar{s}_{j,i}^{(t+1)} = \sqrt{\frac{\bar{\alpha}_{j,i}^{(t)}}{(\bar{\theta}_{j,i}^{(t+1)})^2 + \Sigma_{j,ii}^{(t+1)}}}, \tag{33}$$

  where $\bar{\theta}_{j,i}^{(t)}$ and $\Sigma_{j,ii}^{(t+1)}$ are the i-th element of $\bar{\boldsymbol{\theta}}_j^{(t)}$ and (i,i)-element of $\boldsymbol{\Sigma}_j^{(t+1)}$, respectively[3].

- Update equation for $\{\alpha_{j,i}\}$

  $$\bar{\alpha}_{j,i}^{(t+1)} = (\kappa+1)\left(\nu + \frac{\bar{\tau}_{j,i}^{(t+1)}}{2}\right). \tag{34}$$

Starting from some initial values $\left\{\bar{s}_{j,i}^{(0)}\right\}, \left\{\bar{\alpha}_{j,i}^{(0)}\right\}$ and iterating the above algorithm until it converges, we obtain an approximation posterior distribution $q(\boldsymbol{\theta}_j|D^N) = \mathcal{N}(\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\Sigma}}_j), j = 1, \ldots, m$, where $\hat{\boldsymbol{\theta}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ are the convergence values of $\bar{\boldsymbol{\theta}}_j^{(t)}$ and $\boldsymbol{\Sigma}_j^{(t)}$, respectively. Using these, (24) and (15) are approximated as

$$d_{VB}(D^N) = \int \bar{y}_x(G_{\mathrm{full}}, \boldsymbol{\theta}_{G_{\mathrm{full}}}) \prod_{j=1}^{m} q(\boldsymbol{\theta}_j|D^N)\mathrm{d}\boldsymbol{\theta}_{G_{\mathrm{full}}}, \tag{35}$$

$$\tilde{d}_{VB}(D^N) = \bar{y}_x\left(G_{\mathrm{full}}, \left\{\hat{\boldsymbol{\theta}}_j\right\}_{j=1,\ldots,m}\right), \tag{36}$$

respectively.

The computational complexity of an iteration of the algorithm is $O(m_j^3)$, which comes from the inversion of the $m_j \times m_j$ matrix. If the dimension of the problem is too high to explicitly calculate $\boldsymbol{\Sigma}_j^{(t+1)}$, we have to use an approximate matrix inversion algorithm with low complexity. In this paper, we do not deal with such high-dimensional problems.

5 Experiments

5.1 Experiments on synthetic data

To verify the effectiveness of the proposed method, we have to compare it with a conventional method. A general method of calculating the MIE is to first estimate the graph structure G, estimate the parameters $\boldsymbol{\theta}_G$ of the conditional probability distributions, and finally calculate the MIE by (5). The K2 algorithm [Cooper and Herskovits, 1992] is used for the learning of the graph structure and the posterior mean is used for the estimator of $\boldsymbol{\theta}_G$. The K2 algorithm requires a metric to compare two models. The posterior probabilities of the models are used as the metric for the K2 algorithm. It is a greedy algorithm and it adds a directed edge if the posterior of the model increases. Another method to be compared is to estimate the MIE under the graph G_{full}, which assumes that all potential direct edges exist. It calculates $\boldsymbol{\theta}_{G_{\mathrm{full}}}^{MAP}$ according to (16) and then computes the MIE.

Graphs in the form of Figure 2 are used as G_{full}, where $W_1, \ldots, W_{n_1}$ and $Z_1, \ldots, Z_{n_2}$ are the covariates. This graph has following edges:

- Edges (W_i, X) and (W_i, Y) for all $i = 1, \ldots, n_1$
- Edges (X, Z_i) for all $i = 1, \ldots, n_2$
- Edges (W_i, Z_j) for all $i = 1, \ldots, n_1, j = 1, \ldots, n_2$

A covariate W_i causes a pseudo correlation between X and Y if the edges (W_i, X) and (W_i, Y) exist, while X has an indirect effect on Y through Z_i if the edges (X, Z_i) and (Z_i, Y) exist.

First, we compare the proposed method with a conventional method based on the K2 algorithm and the

[3] To make the description concise, indices of the variables are replaced so that $\theta_{j,i} = \theta_{j_i j}$.

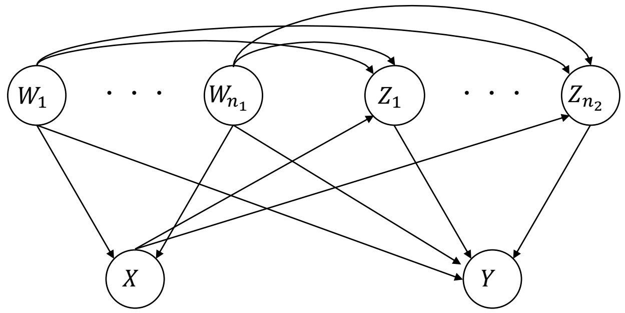

Figure 2: An example of G_{full} used for the experiment. $W_1, \ldots, W_{n_1}$ would cause pseudo correlations between X and Y and X would have indirect effects on Y through $Z_1, \ldots, Z_{n_2}$.

Bayes quasi-optimal estimator for a small G_{full}. The proposed method is the estimator (36) and the Bayes quasi-optimal estimator is the estimator (15). The hyper-parameters κ, ν of the proposed method is set to $\kappa = \nu = 10^{-6}$. The number of the covariates is set to $n_1 = n_2 = 2$ and the edge appearance probability is set to $p = 0.5$. We consider the problem of estimating the MIE (5) when $x = 1$. Figure 3 shows the squared error curves for the MIE as the functions of the sample size. We can see that the proposed method outperforms the estimator based on the K2 algorithm and that based on the full model. We can also see that its performance is close to that of the Bayes quasi-optimal estimator.

Then, we compare the proposed method with the conventional estimators for a large G_{full}. The number of the covariates is set to $n_1 = n_2 = 30$ and the edge appearance probability is set to $p = 0.3, 0.5, 0.7$. The other settings are the same as the previous experiment. In this case, since the number of candidate graphs $|\mathcal{G}|$ is 2^{960}, we can not compute the Bayes quasi-optimal estimator. Figure 4 shows the mean squared error curves for the MIE as the functions of the sample size. We can see that the proposed method outperforms the K2 based estimator and full model based estimator for the all values of p. We think that these experiments demonstrate that the Bayesian model averaging can be approximated with high accuracy by using GSM for the total effect inference problem.

5.2 Experiments on semi-synthetic data

In real-world applications, ground truth causal effects are rarely available. Thus, we evaluate our proposed method through some experiments on semi-synthetic data. For the evaluation, we use two pre-established benchmarking data for causal inference.

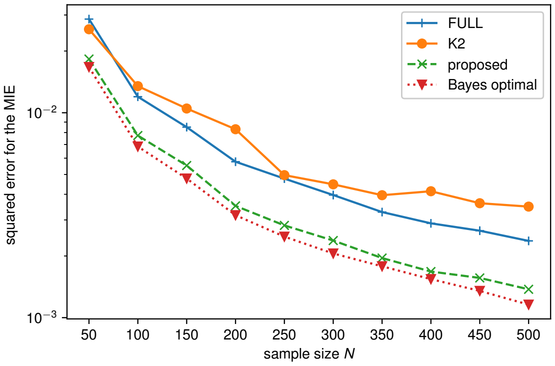

Figure 3: Curves of the squared errors for the MIE for a small G_{full} as the function of sample size. The means of the squared errors of 1000 experiments are depicted.

- **IHDP:** The dataset is constructed from the Infant Health and Development Program (IHDP) [Hill, 2011]. Each observation consists of 25 covariates, an indicator variable that indicates whether the infant received a special care, and an outcome variable that represents the final cognitive test score. The dataset has 747 observations.
- **LBIDD:** The dataset was developed for the 2018 Atlantic Causal Inference Conference competition [Shimoni et al., 2018]. It was derived from the Linked Birth and Infant Death Data (LBIDD). There are 63 distinct data which are generated from different distributions and we randomly pick two of them. Each observation consists of 177 covariates, an indicator variable for the treatment status, and an outcome variable. The dataset has 1000 observations.

Both IHDP and LBIDD data consists of some covariates, an indicator variable for the treatment status, and an outcome variable. We denote the covariates as $W_1, \ldots, W_n$, the indicator variable for the treatment status as X, and the outcome variable as Y. We are interested in estimating $\bar{y}_1 - \bar{y}_0$, namely, the average treatment effect (ATE) of X on Y. Since both datasets contain counterfactual data, we can calculate the treatment effect for each record. We assume that the average of them is the true value of the ATE and evaluate the estimators with the squared errors between the true ATE and estimates.

For both datasets, we assume that G_{full} is the form of Figure 5. That is, we assume that $W_1, \ldots, W_n$ are the possible common causes between X and Y. The hyperparameters for each estimation algorithm are set

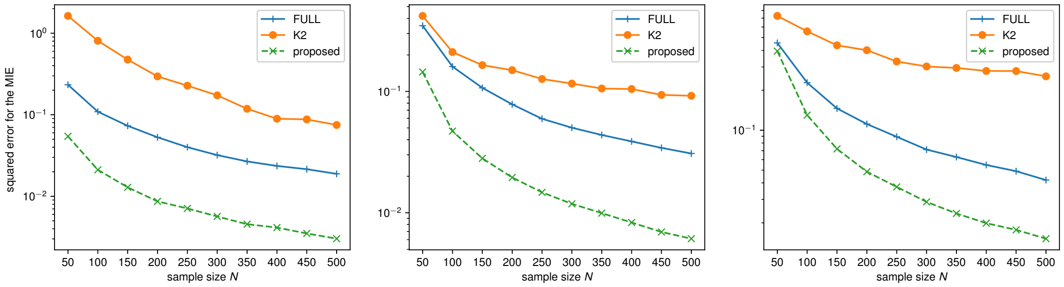

Figure 4: Curves of the squared errors for the MIE for G_{full} with $n_1 = n_2 = 30$ (left: $p = 0.3$, center: $p = 0.5$, right: $p = 0.7$). The means of the squared errors of 1000 experiments are depicted.

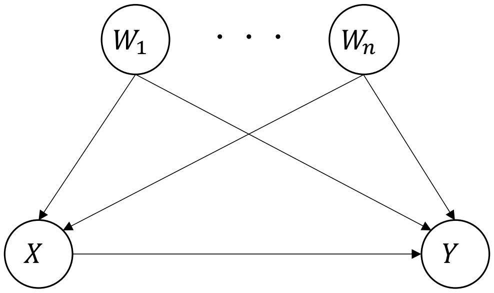

Figure 5: Assumed G_{full} for the experiments on semi-synthetic data.

to the same values as those for the experiments on synthetic data.

We also compare the methods with the IPW estimator, a well known non-Bayesian estimator of the ATE [Horvitz and Thompson, 1952]. IPW estimator requires an estimator of the conditional probability $p(x|w_1, \ldots, w_n)$. We modeled the conditional probability by logistic regression model and estimate its parameters by ℓ_1 penalized maximum likelihood estimation.

Figure 6 shows the mean squared curves for the ATE. We can see that the proposed method is superior to the conventional methods. Especially, for the first LBIDD data, the performance of the K2 algorithm based method is poor and it seems to fail to capture a good causal model. Comparering the proposed method with the IPW estimator, the IPW estimator has a better estimation accuracy than the proposed estimator for IHDP data; the IPW and proposed estimators are competitive for the first LBIDD data; the proposed estimator outperforms IPW estimator for another LBIDD data. The experimental results imply that our proposed method is robust to misspecification of the causal model. We believe that this robustness comes from the fact that the proposed method is build on the idea of weighting the estimates computed under multiple models.

6 Conclusion

We proposed a Bayes optimal estimator of the MIE which minimizes the squared error loss on average when the data generating model is an unknown random variable. The proposed estimator has to estimate the causal effects under the all candidate causal graphs and it is hard to compute when the number of candidate causal graphs is large. We also proposed an approximation algorithm for the optimal estimator by using a sparse modeling technique. Some numerical experiments corroborated the effectiveness of our proposed methods. The proposed methods can help analysts when there is some ambiguity in the knowledge of the data generating model.

We have made some strong assumptions in this study:

- Gaussian assumption on the prior or the error terms
- Assumption that there is no hidden confounders
- Assumption that causal orders are known
- Linear assumption on the SCM

These assumptions might limit the range of the real-world applications of the proposed methods. We think that we can relax some assumptions by combining the proposed methods with previously known results. For example, if we can assume non-Gaussian distributions on the error terms, we can relax the assumption on causal orders since models with opposite causal directions would have different posterior probabilities. In such cases, the method proposed in this paper needs to

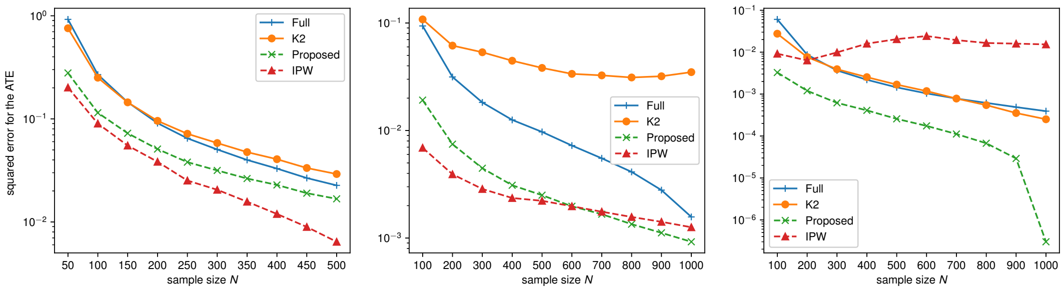

Figure 6: Curves of the squared errors for the MIE for semi-synthetic data (left: IHDP, center and right: LBIDD). The means of the squared errors of 1000 experiments are depicted.

be modified, and the constrution of such an algorithm would be an attractive research direction.

Acknowledgments

This research is partially supported by Information Services International-Dentsu (ISID), Ltd. Research Grant, the Kayamori Foundation of Informational Science Advancement, and No. 19K12128 of Grant-in-Aid for Scientific Research Category (C) and No. 18H03642 of Grant-in Aid for Scientific Research Category (A), Japan Society for the Promotion of Science.

References

[Andrews and Mallows, 1974] Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102.

[Angrist et al., 1996] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

[Babacan et al., 2014] Babacan, S. D., Nakajima, S., and Do, M. N. (2014). Bayesian group-sparse modeling and variational inference. *IEEE transactions on signal processing*, 62(11):2906–2921.

[Berger, 2013] Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

[Boris Choy and Chan, 2008] Boris Choy, S. and Chan, J. S. (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zealand Journal of Statistics*, 50(2):135–146.

[Carvalho et al., 2009] Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80.

[Cooper and Herskovits, 1992] Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347.

[Figueiredo, 2003] Figueiredo, M. A. (2003). Adaptive sparseness for supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 25(9):1150–1159.

[Fisher, 1951] Fisher, R. A. (1951). The design of experiments.

[Guo and Fraser, 2014] Guo, S. and Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications*, volume 11. SAGE publications.

[Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243.

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

[Hoeting et al., 1999] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

[Horii and Suko, 2019] Horii, S. and Suko, T. (2019). A note on the estimation method of intervention effects based on statistical decision theory. In *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

[Horvitz and Thompson, 1952] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

[Ishwaran et al., 2005] Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

[Ji et al., 2008] Ji, S., Xue, Y., and Carin, L. (2008). Bayesian compressive sensing. *IEEE Transactions on signal processing*, 56(6):2346–2356.

[Kaplan and Chen, 2014] Kaplan, D. and Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate behavioral research*, 49(6):505–517.

[Le Cam, 2012] Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

[Park and Casella, 2008] Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

[Pearl, 2000] Pearl, J. (2000). Causality: Models, reasoning, and inference.

[Rubin, 1978] Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

[Shimizu et al., 2006] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

[Shimoni et al., 2018] Shimoni, Y., Yanover, C., Karavani, E., and Goldschmnidt, Y. (2018). Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*.

[Spirtes and Glymour, 1991] Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72.

[Wermuth and Lauritzen, 1982] Wermuth, N. and Lauritzen, S. L. (1982). *Graphical and recursive models for contigency tables*. Institut for Elektroniske Systemer, Aalborg Universitetscenter.

[Wright, 1921] Wright, S. (1921). Correlation and causation. *J. agric. Res.*, 20:557–580.