Technical University of Munich

Department of Mathematics

Credible Intervals for Causal Effects in Linear Causal Models

Master's Thesis

Jiaqi Lu

Supervisor:	Prof. Dr. Mathias Drton
Advisor:	David Strieder
Submission Date:	15.12.2022

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Garching bei München, 15.12.2022

Acknowledgements

First, I would like to thank Professor Mathias Drton for suggesting the fascinating topic of causal inference that allows me to combine knowledge from books with cutting-edge research, and for outlining directions with his insights and experience when I feel uncertain about my work. I would also want to thank David Strieder for answering all kinds of my questions, having inspiring discussions, and providing professional guidance. I regard them as awesome supervisors and advisors, as well as highly competent researchers.

In addition, I would also like to thank my friends Zhi Liu, Zhishan Li, Qing Hu, and Gen Li for their academic help including but not limited to reviewing my thesis and providing helpful feedback.

Last but not least, I would like to thank my parents, my boyfriend, and all my friends for their love and support during the writing of my thesis. Without them, I might not survive the six months. Thanks!

German Abstract

Eine vielleicht interessantere Aufgabe als die statistische Inferenz ist die kausale Inferenz, bei der nicht nur Korrelationen, sondern auch Kausalitäten untersucht werden. Wir können die kausale Struktur einer Gruppe von Variablen mit Hilfe von gerichteten azyklischen Graphen illustrieren, in denen die kausalen Richtungen und bedingten Unabhängigkeiten gezeigt werden. Die grafische Darstellung hat jedoch einige Einschränkungen, wenn ein Effekt das Ergebnis einer Interaktion ist. Deshalb verwenden wir die Strukturmodelle als eine alternative analytische Standarddarstellung, bei denen jede Variable durch ihre Eltern und etwas Rauschen definiert ist.

Eine wichtige Aufgabe bei der kausalen Inferenz ist die Entdeckung der kausalen Struktur mit reinen Beobachtungsdaten. Dies ist im Allgemeinen unmöglich, da Variablen mit unterschiedlichen Kausalstrukturen genau dieselbe gemeinsame Wahrscheinlichkeitsverteilung haben können und die Entdeckung nur auf Markov-äquivalente Klassen beschränkt werden kann. Aber wenn alle relevanten Variablen beobachtet werden, ist die Aufgabe jedoch in einigen Fällen möglich. Einer davon ist wenn wir annehmen, dass die kausalen Beziehungen linear sind und das Rauschen die gleiche Varianz hat.

Wir betrachten den bivariaten Fall unter dieser Annahme. Wir schlagen A-priori-Wahrscheinlichkeiten für die zugrundeliegende kausale Richtung, die gleiche aber unbekannte Varianz und den von Null verschiedenen kausalen Effekt der richtigen Richtung vor. Wir haben auch einige Kriterien für die Festlegung geeigneter priorisierter Hyperparameter diskutiert. Die Konjugierte A-priori-Verteilungen, die wir setzen, ermöglichen eine geschlossene A-posteriori-Verteilung abzuleiten und sie mit Glaubwürdigkeitsregionen weiter zusammenzufassen. Das A-posteriori-Verteilung wird eine Mischverteilung aus einem Punktmaß an 0 und einer stetigen Verteilung sein. Daher untersuchen und diskutieren wir die Bedingungen von 0 in der Region. Wir schlagen drei Arten von Regionen vor, nämlich das "equal-tailed interval" (ETI), die "highest density region" (HDR) und die "threshold Region". Der dritte Typ enthält 0 nur wenn die Dichte bei 0 höher als eine Schranke ist.

Unser Modell wird sowohl mit simulierten Daten als auch mit Benchmarks getestet und verglichen. Unsere Beobachtung aus den Experimenten zeigt, dass unser Modell dazu tendiert zu sicher über die vorhergesagte grafische Struktur ist. Um das problem zu lösen verwenden wir einen Bootstrap-Durchschnitt. Wir implementieren unser Modell mit `R` und die Ergebnisse zeigen vergleichbare Abdeckungsraten und kleinere Breite im Durchschnitt.

English Abstract

Causal inference is perhaps a more interesting topic than statistical inference, where not only correlations but also causations are studied. We can illustrate the causal structure of a group of variables using directed acyclic graphs (DAGs) that represent the causal directions and conditional independencies. However, the graphical representation has some limitations if an effect is the result of an interaction, so an alternative standard analytical representation is using structural equation models (SEMs), where each variable is defined by its parents and some noise.

An important task in causal inference is to discover the causal structure with only observational data. This is in general impossible since variables with different causal structures can have exactly the same joint probability distribution and the discovery can only be limited to Markov equivalent classes even in the best case. However, under the assumption that all relevant variables are observed, the task is possible in some cases, one of which is if we assume the causal relations to be linear and the noises to have equal variance.

We consider the bivariate case under this assumption. We propose priors for the underlying causal direction, the equal but unknown variance, and the non-zero causal effect of the correct direction. We also discussed some criteria for setting proper prior hyperparameters. The conjugate priors we set allow us to derive a closed-form posterior distribution and further summarize it with credible regions (CRs). The posterior will be a mixture of a point mass at 0 and a continuous distribution, so we especially discussed the conditions of 0 being included in the CR. We suggest three types of CRs, i.e., the equal-tailed interval (ETI), the highest density region (HDR), and the "threshold CR". The third type is proposed by us, where the decision to include 0 is made by comparing the density at 0 with a self-defined threshold.

Our model is tested and compared with both simulated data and benchmarks. In the experiments, we figure out that our model tends to be too certain about predicting the graphical structure, so we use a bootstrap average to reduce the effect of extreme decisions. We implement our model with `R` and the results show comparable coverage rates and smaller average widths.

Contents

1. Introduction **1**

2. Basics **4**
2.1. Graphical Structure 4
2.2. Structural Equation Models 7
2.3. Intervention and Causal Effect 10
2.4. Linear SEM with Homoscedastic Noise 12
2.5. Bayesian Inference in Bivariate Case 13
2.6. Bootstrapping 15

3. Bayesian Inference **17**
3.1. Posterior Distribution of Parameters 17
3.1.1. Given Random Variable X 17
3.1.2. Given i.i.d. Samples X^n 23
3.1.3. Discussion of Hyperparameters 25
3.2. Posterior distribution of Causal Effect 28
3.3. Credible Regions 30
3.3.1. Equal-Tailed Interval 33
3.3.2. Highest Density Region 36
3.3.3. Credible Region with a Threshold 41

4. Algorithms and Experiments **43**
4.1. Simulated Data 43
4.2. Benchmarks 58

5. Conclusion **64**

A. Probability Theory **66**

B. Distributions **68**

References **71**

1. Introduction

Statistical inference Peters et al. (2017) has been used to infer properties of the dependence among random variables. However, sometimes only knowing the correlation can lead us to completely wrong and even absurd results. For example, a previous study Messerli (2012) has verified that there is a significant positive linear correlation between chocolate consumption and winning the Nobel prize. Another study Peters (2013) stated that the rate of homicides increases as ice cream sales increase. Both conclusions from statistical learning sound totally non-logical if any rational person thinks of them with their brain. In fact, they can be having a common cause or even be completely independent of each other. To be able to infer the relation between two or more variables better, we might want to ask questions what is the real cause of what? How will other variables change if we apply actions or interventions to certain variables?

We call the task where we explore causal relations: causal inference. In addition to only looking at the joint distribution of a group of variables, we now also focus on the structures between variables Peters et al. (2017). The very natural illustration of a causal structure is using a causal Bayesian network Spirtes et al. (2000), which we usually represent by a directly acyclic graph (DAG) Spirtes et al. (2000) concerning the conditional independence relations in a probability distribution. Variables are denoted by nodes, and direct causalities are represented by directed edges. However, there are also clear limitations of such kinds of representations when effects do not only come from simple direct causes but also from interactions of variables or indirect causes as explained in Spirtes et al. (2000). A better way to represent interactions is through the probability distribution associated with the graph, which can be defined by a causal structural model (SEM) Pearl (2009). An SEM is defined as a collection of equations and a joint probability distribution, which acts as a powerful analytical tool when we for example apply interventions and analyze causal effects.

There are basically two scenarios corresponding to two types of SEMs concerning how much we know about the whole environment, namely the deterministic or quasi-deterministic case Spirtes et al. (2000); Pearl (2009). The former assumes that the effects are deterministic functions (with independent noises) of their direct causes, while the latter refers to the case where some variables are not determined by their immediate causes or some of their causes are unobserved. Both assumptions have been studied under different needs.

In real-world scenarios, one of the most important and interesting things that catches our attention is to identify the true causal relation purely from the joint distribution of observational data. Being able to do this is quite surprising and appealing because

we could go one step beyond statistical inference without using any extra information. In general, reconstructing the causal graph and recovering the causal structure purely from the joint distribution can be carried out by the PC algorithm Spirtes et al. (2000), yet the procedure is inefficient and the discovery can only be limited up to Markov equivalence class. However, under the assumption of causal sufficiency, i.e., all relevant variables are observed Peters et al. (2014), we might have the chance to recover the SEM from the joint distribution solely if the distribution has causal minimality Peters et al. (2014) and the SEM satisfy some special assumptions Peters et al. (2017). For example, the LiNGAM, i.e., the SEM with linear functions and non-Gaussian noises can be identified from the joint distribution Shimizu et al. (2006). Also if we have an additive noise model (ANM) with non-linear functions and non-Gaussian noises, then the underlying causal structure is identifiable from the joint distribution Hoyer et al. (2008); Peters et al. (2012). If we assume the SEM consists of linear functions with Gaussian noises, then the causal structure is identifiable from the joint distribution if the noises are assumed to have equal variances Peters and Bühlmann (2014). In this thesis, we will pay attention to the last setting i.e., linear function with noises of equal variances.

To dig even deeper than just recognizing the causal structure, we could argue about the causal effect after an intervention on some variables of other variables. Informally speaking, causal effects can be seen as correlations in certain directions. Knowing the causal effect, we can answer questions such as what is the amount of aspirin we need to relieve headaches, or how exactly would smoking more cigarette increase the risk of getting lung cancer. We will see if we have some data and assume that the SEM is known, we can compute the causal effects as statistics as they are functions of the data.

A natural way to summarize the prediction of a coefficient is to calculate its $(1-\alpha)$-confidence region using some hypothesis tests. For example in Strieder et al. (2021), several statistical tests, e.g., the likelihood ratio tests (LRT) and split likelihood ratio tests (SLRT) are considered, and the out-coming confidence regions highlight the range of values the causal effects most likely will be. Another state-of-the-art approach for estimating the causal effect is using Bayesian inference. Considering some self-set priors, we compute the posterior distribution of the causal effects and summarize the distribution with a $(1-\alpha)$-credible region. There are already plenty of previous works that focused on this direction, for example, Hoyer and Hyttinen (2012) focused on structure discovery of linear graphs, and Cao et al. (2019) dig deep into posterior high-dimensional graph selection. In addition to focusing mainly on the graphical structure, Castelletti and Consonni (2021) proposed a possibility to sample causal effects from its posterior distribution. But to the best of our knowledge, there is still no paper that focused on directly estimating the credible region of the causal effects.

We are going to present our work in the following order. In Chapter 2, we will introduce basic notations and assumptions of the thesis, and also the causal inference theory, e.g., DAG, intervention, and causal effects that are relevant to our purpose, and we will restrict our attention to bivariate cases. Then in Chapter 3, we will set priors

for the causal graph, the common yet unknown variance, and the coefficient under the correct direction. With these, we will derive the credible regions (CR) for causal effects by applying Bayesian inference. We will give closed-form expressions for different types of CRs, including the equal-tailed interval (ETI), the highest density region (HDR), or a CR similar to the HDR except we determine whether to include 0 with a threshold. We will especially discuss under what setting would the hyperparameters make it a good prior, and how we determine whether to include 0 in the CR. In Chapter 4, we will test our model on both simulated data and benchmarks. Since there is not much similar work in computing the CR, we will compare our results with Strieder et al. (2021) in terms of coverage rate, width, and zero percentage. Finally, in Chapter 5 we will have a brief conclusion and discussion.

2. Basics

In this chapter, we will first introduce some definitions of graphical structures, such as directed acyclic graphs (DAGs) and d-separation. Then we will present the causal relation of a group of random variables in another manner, namely using the structural equation models (SEMs). We will show that DAGs and SEMs are equivalent under certain restrictions, and the causal effects under interventions can be computed from the SEMs. We will argue that recovering the causal structure only from observational data is possible when causal sufficiency and some other model assumptions are satisfied. Following one of the identifiable settings, we will focus on bivariate linear SEMs with Gaussian noises of equal variances, and analyze the causal effects after an intervention. The notations and assumptions mentioned in this chapter will be continuously used throughout the thesis.

2.1. Graphical Structure

As one of the mostly preferred illustration tool, we will introduce some graphical terminologies. The definitions introduced in this section regarding graphical structure follow contentwise and logically from Peters et al. (2017), while we also refer to Pearl (2009) for some details and interpretations.

Definition 2.1 (Graph)**.** A **graph** $\mathcal{G} = (V, E)$ consists of a set of **vertices** (or **nodes**) V and a set of **edges** $E \subset V \times V$ with $(v, v) \notin E$ for any $v \in V$.

In our graphs, the vertices often represent variables, whereas the edges represent particular relationships that exist between pairs of variables. More specific interpretation however will differ depending on the application.

Definition 2.2 (Adjacency, Parent and Child)**.** Let $\mathcal{G} = (V, E)$ be a graph with $V = \{1, \ldots, n\}$ and $n \geq 2$. Two vertices i, j are **adjacent** if either $(i, j) \in E$ or $(j, i) \in E$. A vertex i is called a **parent** of j if $(i, j) \in E$ yet $(j, i) \notin E$, and a **child** of j if $(j, i) \in E$ yet $(i, j) \notin E$. The set of parents of vertex i is denoted by $pa_{\mathcal{G}}(i)$ and the set of children of i is denoted by $ch_{\mathcal{G}}(i)$. We sometimes call the elements of $pa_{\mathcal{G}}(i)$ not only parents but also **direct causes** of i, while i is called the **direct effect** of all its direct causes. In bivariate case that we will discuss the most later, we will even omit the word "direct", and just say cause and effect.

In a graph, each edge can be either directed (denoted by a single arrowhead on the edge) or undirected.

Definition 2.3 (Directed Edge)**.** An edge between two vertices i and j is **undirected** if $(i, j) \in E$ and $(j, i) \in E$. Sometimes it is also called "bidirected". An edge is **directed** if it is not undirected, and we write $i \to j$ for $(i, j) \in E$.

Like in a family tree, we are not only interested in nodes that are directly connected by an edge, but also those node groups that are distinct relatives, so we use a path to connect all nodes that are more or less related to each other.

Definition 2.4 (Path, Ancestors and Descendants)**.** A **path** in a graph $\mathcal{G}$ is a sequence of distinct vertices $i_1, \ldots, i_m$ with $m \geq 2$, such that i_k and i_{k+1} are adjacent for all $k = 1, \ldots, m-1$. If further we have $i_k \to i_{k+1}$ for all $k = 1, \ldots, m-1$, then we call it a **directed path** from i_1 to i_m. In this case, we call i_1 an **ancestor** of i_m and i_m a **descendant** of i_1. We further call vertex j a **non-descendant** of vertex i if j is not a descendant of i. As commonly assumed (e.g. in Peters et al. (2017)), a vertex i is neither an ancestor, nor a descendant, nor a non-descendant of itself. We denote the set of ancestors of i by $an_{\mathcal{G}}(i)$, the set of descendants of i by $de_{\mathcal{G}}(i)$, and the set of non-descendants of i by $nd_{\mathcal{G}}(i)$.

Remark 2.5. In later writing, when it is clear to which graph we are referring, or when pointing out the graph is not necessary, we will simplify the notations for the sets of parents, children, ancestors, descendants and non-descendants, and write them as $pa(i)$, $ch(i)$, $an(i)$, $de(i)$, $nd(i)$ instead.

Finally we define a specific type of graph called a DAG, which is used to represent a priori assumptions about the connections between and among variables and has various applications in science and computation.

Definition 2.6 (DAG)**.** We call a graph $\mathcal{G}$ a **directed acyclic graph** (DAG) if the following two properties hold:

1. all edges are directed;
2. there is no directed circle, i.e., no pair of vertices (i, j) with directed path from i to j and from j to i.

The second condition can also be understood as no directed path from a node to itself is permitted. This property is important for being able to arrange vertices in a causal ordering (see Peters et al. (2017)) that is consistent with all edge directions, which also enforces the notion that causes must come before their effects. Example 2.7 shows what a DAG could look like.

Example 2.7. We displayed an example of a DAG with four nodes (X_1, X_2, X_3, X_4) and four edges. It is a DAG since all edges are directed and there is no cycle.

Since it will be used later in this thesis, we will formulate the graphical concepts of blocked paths and d-separations, which could be interpreted as stopping the flow of information between the variables that are transferred by some paths, as defined next.

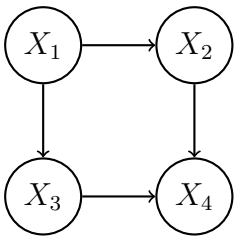


Figure 2.1.: A four-variate DAG with four edges.

Definition 2.8 (d-separation)**.** Let $\mathcal{G} = (V, E)$ be a DAG. A path π between nodes i_1 and i_n is said to be **blocked** by a set of nodes C if and only if it contains a node $i_k \in \pi$ such that either

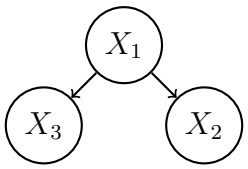
- $i_k \in C$, and directed edges in π do not meet head-to head at i_k, i.e., either $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$, or
- $i_k \notin C$, nor has i_k descendants in C and directed edges in π do meet head-to-head at i_k, i.e., $i_{k-1} \to i_k \leftarrow i_{k+1}$.

Set $A, B \subset V$ are said to be **d-separated** by $C \subset V$ if all paths between $i \in A$ and $j \in B$ are blocked by C, and we write $A \perp\!\!\!\perp_{\mathcal{G}} B|C$.

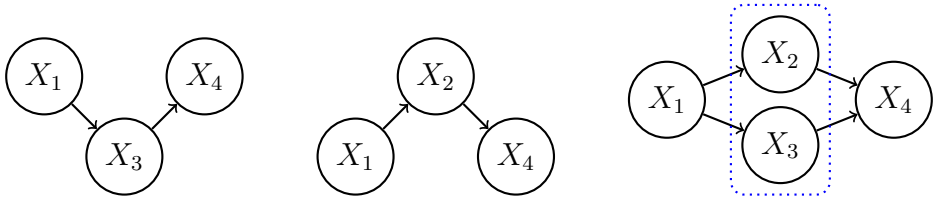
The following example will provide a more intuitive understanding of d-separation.

Example 2.9. We will give examples of blocked paths and d-separations considering the DAG as in Figure 2.1. We do not distinguish here a single element from a singleton set.

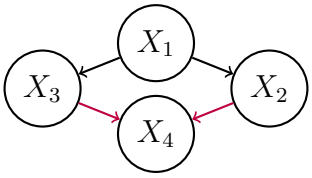
Considering path $\pi_1 : X_2 - X_1 - X_3$ between X_2 and X_3, then π_1 is blocked by X_1 since $X_1 \in \pi_1$ and directed edges do not meet head-to-head at X_1.



For similar reasons, path $\pi_2 : X_1 - X_2 - X_4$ is blocked by X_2 and path $\pi_3 : X_1 - X_3 - X_4$ is blocked by X_3. Since π_2 and π_3 are the only two paths between X_1 and X_4, we say that X_1, X_4 is d-separated by $\{X_2, X_3\}$, i.e., $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_4|X_2, X_3$.



Finally, we look at path $\pi_4 : X_2 - X_4 - X_3$. We say π_4 is blocked by X_1 since the only node on path π_4 is $X_4 \notin \{X_1\}$ and X_1 is not a descendant of X_4 and $X_3 \to X_4 \leftarrow X_2$. However, it is worth noticing that π_4 is not blocked by X_4 since $X_4 \in \pi_4$ and directed path do meet head-to-head at X_4.



The concept of d-separation in a graph will later be related to conditional independence of random variables when we introduce Markov properties.

2.2. Structural Equation Models

Structural equation models (SEMs) have been employed for a very long time in various fields, including agriculture, social sciences (Wright (1921); Haavelmo (1944); Duncan (1975)), as well as physics and engineering. SEMs describe causal relations as deterministic, functional equations. We will introduce semantics of SEMs and its relation to DAGs in this section, then learn how to use them to compute intervention distributions in the next section. The main part of this section follows from Peters et al. (2017); Pearl (2009) and some interpretations follow from Mooij et al. (2016).

Definition 2.10 (Structural Equation Models)**.** A **structural equation model** (SEM) $M := (S, \mathbb{P}_\epsilon)$, sometimes also called a structural causal model in other literature, consists of a collection S of equations of the form

$$X_i = f_i(X_{pa(i)}, \epsilon_i), \quad i = 1, \dots, d, \tag{2.1}$$

where $X_{pa(i)}$ denotes the set of parents of X_i. The joint distribution $\mathbb{P}_\epsilon = \mathbb{P}_{\epsilon_1,\dots,\epsilon_d}$ over the noise variables that are supposed to be jointly independent, i.e., $\mathbb{P}_\epsilon = \mathbb{P}_{\epsilon_1} \times \dots \times \mathbb{P}_{\epsilon_d}$. Given a SEM we could draw its corresponding graph $\mathcal{G}$ by creating a vertex for each X_i and drawing a directed edge from every vertex in $X_{pa(i)}$ to X_i. We further require the graph $\mathcal{G}$ to be a DAG.

Example 2.11. We can represent the DAG in Figure 2.1 as an SEM of the form

$$\begin{aligned} X_1 &= f_1(\epsilon_1) \\ X_2 &= f_2(X_1, \epsilon_2) \\ X_3 &= f_3(X_1, \epsilon_3) \\ X_4 &= f_4(X_2, X_3, \epsilon_4), \end{aligned}$$

where X_i correspond to nodes, f_i are individual functions representing the causal relations, and ϵ_i are jointly independent noises for $i = 1, 2, 3, 4$.

The functional relationship in (2.1) could be interpreted as a method for describing the impact that various possible value combinations of $(X_{pa(i)}, \epsilon_i)$ will have on X_i, indicated by a different equation for each variable. Any subset of the structural equations will remain valid and could be viewed as the conditions for a certain set of interventions (Pearl (2009)).

To define a proper graphical model, we introduce a widely used assumption known as the Markov property that allows us to relate d-separation in DAGs with conditional independence of probability distributions, see Pearl (2009); Peters et al. (2017).

Definition 2.12 (Markov Property)**.** Given a DAG $\mathcal{G} = (V, E)$ and a joint distribution $\mathbb{P}_X$, then $\mathbb{P}_X$ is said to satisfy

- the **local Markov property** with respect to the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents, i.e.,

$$X_i \perp\!\!\!\perp X_{nd_{\mathcal{G}}(i)} | X_{pa_{\mathcal{G}}(i)} \text{ for all } i \in V;$$

- the **global Markov property** with respect to the DAG $\mathcal{G}$ if

$$X \perp\!\!\!\perp_{\mathcal{G}} Y | Z \Rightarrow X \perp\!\!\!\perp Y | Z$$

  for all disjoint sets $X, Y, Z \in V$;

- the **Markov factorization property** with respect to the DAG $\mathcal{G}$ if

$$p(x) = p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i | x_{pa_{\mathcal{G}}(i)}),$$

  where we assume $\mathbb{P}_X$ has density p.

The first two Markov properties describe from the perspective that the separation of nodes in a graph shall be equivalent to the conditional independence of sets of random variables. The third Markov property provides an easier formulation of the joint distribution given the parents of each node and is usually useful for calculations. Adding a simple constraint to the joint probability will make all three Markov properties equivalent as shown in the following theorem (Peters et al. (2017)).

Theorem 2.13. *All Markov properties in Definition 2.12 are equivalent if $\mathbb{P}_X$ has a density function p.*

The proof of Theorem 2.13 can be found in Lauritzen (1996).

Since we will only deal with distributions with a density function in this thesis, we will make mixed use of all three Markov properties from Definition 2.12 and simply refer to all of them as Markov properties.

Example 2.14. Let $X = (X_1, X_2, X_3, X_4)$ be a random vector with joint probability $\mathbb{P}_X$ and density p. Note that we simply refer to all density functions as p. We further assume $\mathbb{P}_X$ to be Markovian with respect to the DAG shown in Figure 2.1. Then according to the *global Markov property*, we have

$$X_1 \perp\!\!\!\perp_{\mathcal{G}} X_4 | X_2, X_3 \qquad \Rightarrow \qquad X_1 \perp\!\!\!\perp X_4 | X_2, X_3.$$

Considering the *Markov factorization property*, we can rewrite the joint density as a multiplication

$$p(x_1, \ldots, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3).$$

As a background knowledge for later discussions regarding identifiability of the causal structure from the joint distribution, we also define faithfulness and Markov equivalence class following Peters et al. (2017). We first introduce faithfulness, which can be seen as the converse of the Markov property.

Definition 2.15 (Faithfulness)**.** Let $\mathcal{G}$ be a DAG and $\mathbb{P}_X$ be a joint distribution, then $\mathbb{P}_X$ is **faithful** to $\mathcal{G}$ if

$$X \perp\!\!\!\perp Y | Z \Rightarrow X \perp\!\!\!\perp_{\mathcal{G}} Y | Z$$

for all disjoint vertex sets X, Y, Z.

Definition 2.16 (Markov Equivalence Class)**.** We denote the set of probability distributions $\mathbb{P}$ that are Markovian with respect to DAG $\mathcal{G}$ as

$$\mathcal{M}(\mathcal{G}) := \{\mathbb{P} : \mathbb{P} \text{ that satisfies Markov properties with respect to } \mathcal{G}\}.$$

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are **Markov equivalent** if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. The set of DAGs that are Markovian with respect to $\mathcal{G}$ is called the **Markov equivalence class** of $\mathcal{G}$.

Now we will give the evidence why we could represent causal graph nicely with an SEM following from Peters et al. (2017); Pearl (2009). First, we will show that an SEM can uniquely define a joint distribution.

Proposition 2.17. An SEM defines a unique joint distribution $\mathbb{P}_X$ over the variables $X = (X_1, \ldots, X_d)^\top$ such that $X_i \overset{d}{=} f_i(X_{pa(i)}, \epsilon_i)$ for all $i = 1, \ldots, d$, where $^\top$ denotes matrix (vector) transpose, and $\overset{d}{=}$ means equal in distribution. We call $\mathbb{P}_X$ the **entailed distribution** of the SEM.

Then, we will show that the SEM can be constructed from a joint distribution as long as the Markov property holds.

Proposition 2.18. Consider $X_1, \ldots, X_d$ random variables with strictly positive joint density $\mathbb{P}_X$ with respect to the Lebesgue measure. If we assume that $\mathbb{P}_X$ is Markovian with respect to $\mathcal{G}$, then there exists an SEM $(S, \mathbb{P}_\epsilon)$ with graph $\mathcal{G}$ that generates $\mathbb{P}_X$.

Remark 2.19. From now on when we refer to a SEM, we always assume its corresponding graph $\mathcal{G}$ is a DAG, and its entailed distribution satisfies the Markov property with respect to $\mathcal{G}$.

2.3. Intervention and Causal Effect

We are now prepared to introduce an intervention (Peters et al. (2017)) into an SEM. Intuitively, intervening on a variable means changing its direct causes. For example, a student is playing a game where the only way to win ten euros was to roll a dice to six, and now we change the winning rule to be flipping a coin and getting a tail. We call this an intervention since modifying the rule changes the distribution of winning ten euros. Formally, we construct intervention distributions from an SEM M. They are obtained by making modifications to M and considering the new entailed distribution. In general, intervention distributions are different from the original distributions. This section mainly follows from Peters et al. (2017); Maathuis et al. (2009).

Definition 2.20 (Intervention)**.** Let $M = (S, \mathbb{P}_\epsilon)$ be an SEM with entailed distribution $\mathbb{P}_X$. We call the replacement of one (or more) structural equations (without generating cycles) an **intervention**, and denote the entailed distribution of the new SEM as

$$\tilde{\mathbb{P}}_X = \mathbb{P}_{X|do(X_i=\tilde{f}(\tilde{X}_{pa(i)},\tilde{\epsilon}_i)))},$$

which we call an **intervention distribution**.

In this thesis, we will only work with perfect intervention where $\tilde{f}(\tilde{X}_{pa(i)}, \tilde{\epsilon}_i)$ is just a point mass on some $a \in \mathbb{R}$, hence we will use a shorter notation $\mathbb{P}_{X|do(X_i=a)}$. It is important to understand that an intervention distribution is different from a conditional distribution as explained in the following example.

Example 2.21. Assume we have an SEM

$$X_1 = \epsilon_1,\ X_2 = X_1 + \epsilon_2$$

with $\epsilon_1, \epsilon_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Then we have

$$\mathbb{P}_{X_1|do(X_2=a)} \sim \mathcal{N}(0,1),$$

while

$$\mathbb{P}_{X_1|X_2=a} \sim \mathcal{N}(a,1)$$

and

$$\mathbb{P}_{X_1|do(X_2=a)} \neq \mathbb{P}_{X_1|X_2=a}.$$

Having the Markov property ensures us to compute the intervention distribution of a perfect intervention relatively convenient with a truncated factorization formula, see Maathuis et al. (2009).

Proposition 2.22. Let p be the density of $\mathbb{P}_X$ satisfying the Markov property with respect to some underlying DAG, then the density of the intervention distribution $\tilde{\mathbb{P}}_X$ after intervening on X_i is denoted by

$$p(x_1, \ldots, x_d | do(X_i = \tilde{x}_i)) = \begin{cases} \prod_{j \neq i}^{d} p(x_j | x_{pa(i)})\big|_{x_i=\tilde{x}_i}, & \text{if } x_i = \tilde{x}_i, \\ 0, & \text{otherwise.} \end{cases},$$

We could further compute the marginal distribution and the expectation of a certain variable after intervention following Proposition 2.22 and Maathuis et al. (2009).

Proposition 2.23. The marginal distribution of $Y = X_1$ after the intervention $do(X_i = \tilde{x}_i)$ for $i = 2, \ldots, d$ can be computed by

$$p(y|do(x_i = \tilde{x}_i)) = \begin{cases} p(y), & \text{if } Y \in X_{pa(i)}, \\ \int p(y|\tilde{x}_i, x_{pa(i)}) \cdot p(x_{pa(i)}) \mathrm{d}x_{pa(i)}, & \text{otherwise.} \end{cases}$$

The expectation of Y after intervention $do(X_i = \tilde{x}_i)$ is thus

$$\mathbb{E}[Y|do(X_i = \tilde{x}_i)] = \begin{cases} \mathbb{E}[Y], & \text{if } Y \in X_{pa(i)}, \\ \int \mathbb{E}[Y|\tilde{x}_i, x_{pa(i)}] \cdot p(x_{pa(i)}) \mathrm{d}x_{pa(i)}, & \text{otherwise,} \end{cases}$$

where $\mathbb{E}[\cdot]$ denotes the expectation of a certain random variable under the corresponding probability distribution, and $\mathbb{E}[\cdot|\cdot]$ denotes the conditional expectation.

It is interesting to see that the distributions of the parents of a variable X_i will not change after an intervention on X_i. This leads us to our definition of the causal effect of an intervention given the SEM, see Maathuis et al. (2009).

Definition 2.24 (Causal Effect)**.** Following the notations from Proposition 2.23, we define the **causal effect** of $do(X_i = \tilde{x}_i)$ on Y as

$$\frac{\mathrm{d}}{\mathrm{d}x} \mathbb{E}[Y|do(X_i = x)] \Big|_{x=\tilde{x}_i}.$$

It is worth noticing that the existence of a causal effect is related to the existence of a directed path. The latter statement is a necessary condition of the former, yet not a sufficient condition, see Peters et al. (2017).

Proposition 2.25. Given a SEM with corresponding graph $\mathcal{G}$ drawn upon variables $X_1, \ldots, X_d$. If there is no directed path from X_i to X_j, then there is no causal effect of X_i on X_j. Sometimes there is still no causal effect even if a directed path exists.

The first statement follows directly from the Markov factorization property. The second statement could be shown by a counter example.

Example 2.26. Assume we have an SEM

$$X_1 = \epsilon_1, \quad X_2 = -X_1 + \epsilon_2, \quad X_3 = X_1 + X_2 + \epsilon_3$$

with $\epsilon_1, \epsilon_2, \epsilon_3 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Plugging X_2 in X_3 we have $X_3 = X_1 - X_1 + \epsilon_2 + \epsilon_3 = \epsilon_2 + \epsilon_3$. Because the noises are jointly independent, we have

$$\mathbb{P}_{X_3} = \mathbb{P}_{X_3|X_1=a} = \mathbb{P}_{X_3|do(X_1=a)},$$

and the causal effect of X_1 on X_3 is by definition

$$\frac{\mathrm{d}}{\mathrm{d}x} \mathbb{E}[X_3|do(X_1 = x)] \Big|_{x=a} = \frac{\mathrm{d}}{\mathrm{d}x} \left[\int X_3 \mathrm{d}\mathbb{P}_{X_3|do(X_1=x)} \right] \Big|_{x=a} = \frac{\mathrm{d}}{\mathrm{d}x} \left[\int X_3 \mathrm{d}\mathbb{P}_{X_3} \right] \Big|_{x=a} = 0,$$

since the integral is independent of x.

2.4. Linear SEM with Homoscedastic Noise

Now we look at a special case of the Gaussian linear SEM as an example of the definitions in previous sections, and as a foundation for later chapters. The characterizations and theories follow mainly from Peters and Bühlmann (2014); Chen et al. (2019).

Assumption 2.27. Without loss of generality, we assume a centered random vector $X = (X_1, \dots, X_d)^\top$. We consider an SEM with DAG $\mathcal{G} = (V, E)$ of the form

$$X_i = \sum_{k \neq i} \beta_{ik} X_k + \epsilon_i, \quad i = 1, \dots d, \tag{2.2}$$

where all ϵ_i's are jointly independent random variables with mean zero and the coefficients β_{ik} are unknown parameters. Following Peters and Bühlmann (2014) we assume that all ϵ_i's have a common unknown variance $\sigma^2 > 0$. Additionally, for each $i \in \{1, \dots, d\}$ we require $\beta_{ik} = 0$ if and only if $(X_k, X_i) \notin E$.

We call the SEM generated by the system of equations (2.2) and the assumed type of distributions of $(\epsilon_i)_{i=1}^d$ in Assumption 2.27 a linear structural equation model (LSEM) with homoscedastic noise. To see the nice properties of such a LSEM, we first recall a definition of matrix similarity from linear algebra following from Horn and Johnson (2012).

Definition 2.28. Let A, B be two $n \times n$ matrices. We say that B is **permutation similar** to A if there is a permutation matrix P such that

$$B = P^\top A P.$$

Recall that a square matrix P is a permutation matrix if exactly one entry in each row and column is equal to one.

We are now ready to characterize a LSEM with homoscedastic noises, see Chen et al. (2019).

Proposition 2.29. We denote the coefficients with a matrix $B = (\beta_{ji})_{j,i=1}^d$ that is permutation similar to a strict lower triangular matrix with all zero diagonal entries. Then, the system of equations (2.2) admit the unique solution $X = (I - B)^{-1}\epsilon$, where $\epsilon = (\epsilon_1, \dots, \epsilon_d)^\top$ and I is the identity matrix. Moreover, X has the covariance matrix

$$\mathbb{E}[XX^\top] = \sigma^2 (I - B)^{-1} (I - B)^{-\top},$$

where $^{-\top}$ denotes the inverse transpose of a matrix.

By Definition 2.24 and Proposition 2.25 we compute the causal effects of X_i on X_j as

$$\mathcal{C}(i \to j) := \frac{\mathrm{d}}{\mathrm{d}x} \mathbb{E}[X_j | do(X_i = x)]\Big|_{x = x_i} = \begin{cases} (I - B)^{-1}_{ji}, & \text{if } i \to j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Structural equation models enable us to exploit different types of restrictions. Specifically, we call a SEM with jointly independent normally distributed noise variables a Gaussian SEM. In general, it is impossible to identify the underlying graphical structure $\mathcal{G}$ uniquely from a SEM with the entailed distribution $\mathbb{P}_X$ even if all relevant variables are observed. We will formalize this concept as causal sufficiency in the following definition (see Spirtes et al. (2000)).

Definition 2.30 (Causal Sufficiency)**.** A set of variables X is **causally sufficient** if there is no hidden common cause $S \notin X$ that is causing more than one variable in X.

The causal sufficiency assumption does not hold in many real world scenarios, however it is the basic assumption for our study, on which the reasoning of identifiability is based. Lemma 2.31 shows that the identifiability is only limited to Markov equivalence class for causally sufficient data, see Peters et al. (2017).

Lemma 2.31 (Identifiability of Markov Equivalence Class)**.** Let us assume distribution $\mathbb{P}_X$ to be Markovian and faithful with respect to DAG $\mathcal{G}_0$, then if $\mathbb{P}_X$ is the entailed distribution of another DAG $\mathcal{G}$, $\mathbb{P}_X$ is also Markovian and faithful with respect to $\mathcal{G}$ if and only if $\mathcal{G}$ is in the Markov equivalence class of $\mathcal{G}_0$.

There are two main drawbacks of this identifiability. First, there might still be plenty of possible DAGs in the Markov equivalence class, and it is hard to decide which is the true one. Second, faithfulness is hard to test in general, see Zhang and Spirtes (2008). Luckily, recent studies have shown that unique identifiability is possible for causally sufficient data for some special SEMs: (i) if we consider a LSEM with non-Gaussian noise variables, then the underlying DAG is identifiable (Shimizu et al. (2006)); (ii) if we consider SEMs with non-Gaussian noises that are added to the non-linear functions, then the underlying graphical structure is identifiable from $\mathbb{P}_X$ (Hoyer et al. (2008)); (iii) if we consider a Gaussian LSEM with noise variables sharing the same variance, then the underlying DAG is identifiable from the joint distribution $\mathbb{P}_X$. We formally describe here the third case before we use it later, see Peters and Bühlmann (2014).

Theorem 2.32 (Identifiability)**.** *Let $\mathbb{P}_X$ be the entailed distribution of the SEM* (2.2) *where we additionally assume that the noise variables are i.i.d. normally distributed, i.e., $\epsilon_1, \ldots, \epsilon_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, then the underlying DAG $\mathcal{G}$ is identifiable from the entailed distribution $\mathbb{P}_X$ and the coefficients β_{ik} can be reconstructed for all $i = 1, \ldots, d$ and $k \in pa_{\mathcal{G}}(i)$.*

Following Theorem 2.32, we are confident that estimating the causal structures and causal effects is at least theoretically possible under the setting of Gaussian LSEM with noises of equal variances.

2.5. Bayesian Inference in Bivariate Case

In this section we will further restrict our attention to bivariate case assuming Gaussian LSEM with homoscedastic noise as we will later work mainly under this setting. We

will look into the two formulations of SEMs, analyze the distributions of the bivariate random vectors, and compute the causal effect of one variable on another. The notations and results in this section follows from Strieder et al. (2021).

Let $\mathcal{G} = (V, E)$ be a DAG with 2 nodes $V = \{X_1, X_2\}$. Consider a cause-effect pair $X = (X_1, X_2)^\top$ as a bivariate random vector on $\mathbb{R}^2$ with zero mean. We write $(X_1, X_2) \in E$ if X_1 is the cause of X_2 and vice versa. Since we only have two variables, the cause and effect are naturally to be direct. We always assume that there is an underlying causal relation between X_1 and X_2, hence $E \neq \emptyset$ and have exactly one element, either (X_1, X_2) or (X_2, X_1). We write the DAG with $(X_1, X_2) \in E$ as $\mathcal{G}_{12}$, and conversely the DAG with $(X_2, X_1) \in E$ as $\mathcal{G}_{21}$. It is clear that $\mathcal{G}_{12}$ and $\mathcal{G}_{21}$ are the only two possible graphical structures, i.e., $\mathcal{G} \in \{\mathcal{G}_{12}, \mathcal{G}_{21}\}$.

Then there are two possible formulations of LSEMs. Let two noise variables $\epsilon_1, \epsilon_2 \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. If we assume the relationship between X_1 and X_2 to be represented by $\mathcal{G}_{12}$, i.e., $1 \to 2$, then we formulate the SEM as

$$(M1) \qquad X_1 = \epsilon_1, \quad X_2 = \beta_{21} X_1 + \epsilon_2.$$

On the other hand, if we assume the causal relation to be the opposite direction, i.e., $2 \to 1$, then we have the SEM

$$(M2) \qquad X_1 = \beta_{12} X_2 + \epsilon_1, \quad X_2 = \epsilon_2.$$

In both systems of equations $\beta_{12}, \beta_{21} \in \mathbb{R}$ are unknown coefficients. We can represent the coefficients with one variable by setting

$$\beta = \beta_{12} \mathbb{1}\{\mathcal{G} = \mathcal{G}_{21}\} + \beta_{21} \mathbb{1}\{\mathcal{G} = \mathcal{G}_{12}\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function that attains value 1 when the statement inside the curly bracket is true, and 0 otherwise. Under this assumption, we can express the distribution of X given $\mathcal{G}, \beta$ and σ^2.

Following Proposition 2.29, we now characterize the entailed distribution of model (M1) and model (M2). Writing the system of equations of model (M1) in matrix form, we have

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \beta_{21} & 1 \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix},$$

hence we can identify the distribution of X given all other coefficients as

$$X | \beta_{21}, \sigma^2, \mathcal{G}_{12} \sim \mathcal{N}_2(0, \Sigma) \tag{2.4}$$

with covariance matrix

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \beta_{21} \\ \beta_{21} & \beta_{21}^2 + 1 \end{pmatrix},$$

where $\mathcal{N}_2(\mu, \Sigma)$ stands for bivariate normal distribution with mean vector μ and covariance matrix Σ. Similarly, reformulating model (M2) in matrix form we have

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & \beta_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix},$$

and the entailed distribution of (M2) is

$$X|\beta_{12}, \sigma^2, \mathcal{G}_{21} \sim \mathcal{N}_2(0, \Sigma) \tag{2.5}$$

with covariance matrix

$$\Sigma = \sigma^2 \begin{pmatrix} \beta_{12}^2 + 1 & \beta_{12} \\ \beta_{12} & 1 \end{pmatrix}.$$

Following the notations of Pearl (2009) and Strieder et al. (2021), an intervention on X_i $(i = 1, 2)$ is expressed as $\mathrm{do}(X_i = x_i)$. The causal effects of both directions can be expressed by β_{12} and β_{21} respectively as

$$\mathcal{C}(1 \to 2) = \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}[X_2|\mathrm{do}(X_1 = x)]\Big|_{x=x_1} = \begin{cases} \beta_{21}, & \text{if } 1 \to 2, \\ 0, & \text{otherwise,} \end{cases} \tag{2.6}$$

and

$$\mathcal{C}(2 \to 1) = \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}[X_1|\mathrm{do}(X_2 = x)]\Big|_{x=x_2} = \begin{cases} \beta_{12}, & \text{if } 2 \to 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

2.6. Bootstrapping

To end the theoretical background chapter, we now briefly introduce what is bootstrap as a resampling method following Rizzo (2019) since we will be needing this later. In this section, we will mainly focus on introducing what this method is and how it can be proceeded. We will evaluate the goodness of the estimation in later chapter directly with empirical results.

Bootstrap methods are a class of non-parametric Monte Carlo methods that estimate the distribution of a population by resampling. Resampling methods treat an observed sample as a finite population, and random samples are generated from it to estimate population characteristics and make inferences about the sampled population. Bootstrap methods are often used when the distribution of the target population is not specified, i.e., the sample is the only information available, which is usually the case in real-world scenarios.

The empirical distribution of the observed samples can be considered to have similar characteristics as the true underlying distribution. We can estimate the statistic of the true distribution by repeatedly generating random samples from the observed data, which we also call resampling.

Suppose that $x = (x_1, \ldots, x_n)$ is an observed random sample from a distribution with cumulative distribution function $F(x)$. If X^* is selected at random from x, then

$$\mathbb{P}(X^* = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

Bootstrap generates random samples $X_1^*, \ldots, X_n^*$ by sampling with replacement from x, i.e., the random variables X^* are i.i.d., uniformly distributed on the set $\{x_1, \ldots, x_n\}$.

We could also see this as generating samples from the empirical distribution. The empirical cumulative distribution function $F_n(x)$ is an estimator of $F(x)$. It can be shown that $F_n(x)$ is a sufficient statistic for $F(x)$. In fact, the empirical cumulative distribution function F_n is the cumulative distribution function of X^*. Thus in bootstrap, there are two approximations. Function F_n is an approximation of function F. Function F_n^* of the bootstrap replications is an approximation of the function F_n. Resampling from the sample x is equivalent to generating random samples from the distribution $F_n(x)$.

Strategy. Suppose θ is the parameter of interest , and $\hat{\theta}$ is an estimator of θ. Then the bootstrap estimate of the distribution of $\hat{\theta}$ is obtained as follows.

1. For each bootstrap replication indexed $b = 1, \ldots, B$:
   a) generate sample $x^*(b) = (x_1^*, \ldots, x_n^*)$ by sampling with replacement from the observed sample $x_1, \ldots, x_n$;
   b) compute the b-th replication $\hat{\theta}^{(b)}$ from the b-th bootstrap sample.
2. The bootstrap estimate of $F_{\hat{\theta}}(\cdot)$ is the empirical distribution of the replications $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$.

3. Bayesian Inference

From now on, we only compute the results for estimating $\mathcal{C}(1 \to 2)$ no matter how the true causal direction is unless otherwise stated. The causal effect of the other direction $\mathcal{C}(2 \to 1)$ can be easily computed analogously. Recall from the last chapter that the causal effect given by (2.6) only depends on β_{21} given the DAG structure. In other words, if $1 \to 2$ is the true causal direction, we would expect the estimated $\mathcal{C}(1 \to 2)$ to be close to β_{21}, otherwise the estimated $\mathcal{C}(1 \to 2)$ is expected to be approximately 0. Furthermore, instead of having a fixed estimate of $\mathcal{C}(1 \to 2)$ given observational data (Mooij et al. (2016)), we are interested in providing a credible interval for $\mathcal{C}(1 \to 2)$ by doing Bayesian inference. For this purpose, we will first define and compute prior distributions of the DAG structure $\mathcal{G}$, the variance σ^2 and the causal effect β_{21}, then derive the posterior distribution of $\mathcal{C}(1 \to 2)$ such that we could sample from it, and conclude a credible interval that we believe $\mathcal{C}(1 \to 2)$ will most likely lie in. We will first start with deriving the distributions of the bivariate random vector $X = (X_1, X_2)^\top$, then generalize it to the case when we have a set of n data pairs X^n generated i.i.d. from the same distribution as X. Note that we sometimes do not distinguish between a random variable (X_1, X_2) and a sample (x_1, x_2) drawn from the random variable. In the end, we will derive the posterior distributions and the credible intervals for the causal effects. Since there is no formal definition of credible intervals, we will compute two kinds of most commonly used credible intervals, the equal-tailed interval (ETI) and the highest density region (HDR), and a third type that we defined ourselves through a threshold.

3.1. Posterior Distribution of Parameters

Let us assume X_1, X_2 to be variables either following model (M1) or model (M2), i.e., we always assume that there exists a causal relation between the two variables, yet the direction and the strength is unclear.

3.1.1. Given Random Variable X

Our goal in this section is to sample $\mathcal{G}, \sigma^2$ and β from their posterior distributions given data X. More formally speaking, we need to compute the posterior distribution $p(\sigma^2, \beta, \mathcal{G}|X)$, and the marginal posterior distributions of each random variable. We first briefly recall the Bayes' theorem, then we make our assumptions about the prior distributions of each parameters, and finally we derive step-by-step the posterior distributions. By Bayes' theorem, we have the posterior distribution of graph structure $\mathcal{G}$

given X as

$$p(\mathcal{G}|X) \propto p(X|\mathcal{G}) \cdot p(\mathcal{G}), \tag{3.1}$$

and the posterior distributions of parameters σ^2 and β given X and $\mathcal{G}$ (von Kügelgen et al. (2019)), which could be expressed as

$$p(\sigma^2, \beta|X, \mathcal{G}) \propto p(X|\sigma^2, \beta, \mathcal{G}) \cdot p(\sigma^2, \beta|\mathcal{G}), \tag{3.2}$$

where $p(\cdot)$ denote a probability density function. Note that here we do not distinguish between a distribution and its density function.

Following the theory from Castelletti and Consonni (2021), it is natural to set the prior distributions of the graph structure $\mathcal{G}$ and parameters β_{21}, σ^2 as

$$p(\mathcal{G}) \sim \frac{1}{2}\mathbb{1}\{\mathcal{G} = \mathcal{G}_{12}\} + \frac{1}{2}\mathbb{1}\{\mathcal{G} = \mathcal{G}_{21}\}, \tag{3.3}$$

$$p(\sigma^2|\mathcal{G}_{12}) \sim \text{I-Ga}\left(\frac{1}{2}a_{12}, \frac{1}{2}U_{11}\right), \tag{3.4}$$

$$p(\beta_{21}|\sigma^2, \mathcal{G}_{12}) \sim \mathcal{N}\left(\frac{U_{12}}{U_{11}}, \frac{\sigma^2}{U_{11}}\right), \tag{3.5}$$

where I-Ga(a, b) stands for Inverse-Gamma distribution with shape $a > 0$ and rate $b > 0$ (see Appendix B.1). The usage of parameters $a_{12}, U_{11}, U_{12} \in \mathbb{R}$ follows from Castelletti and Consonni (2021) for comprehensive and computational convenience. In order to let the above distributions be well defined, we need $a_{12}, U_{11}, U_{12} > 0$. Plugging the prior distributions in the Bayes' theorem, we could split our goal of this section into three different parts, more specifically we need to compute the three distributions in the following order

$$p(\mathcal{G}|X), \tag{3.6}$$

$$p(\sigma^2|X, \mathcal{G}_{12}), \tag{3.7}$$

$$p(\beta_{21}|\sigma^2, X, \mathcal{G}_{12}). \tag{3.8}$$

We first start with deriving a convenient formulation of $p(X|\mathcal{G}_{12})$ such that is easy to refer to and compare with later.

Lemma 3.1. The distribution $p(X|\mathcal{G}_{12})$ can be formualted as

$$p(X|\mathcal{G}_{12}) = p(X, \mathcal{G}_{12}) \int_0^{+\infty} \left(\int_{-\infty}^{+\infty} p(\beta_{12}|\sigma^2, X, \mathcal{G}_{12}) \mathrm{d}\beta_{21} \right) p(\sigma^2|X, \mathcal{G}_{12}) \mathrm{d}(\sigma^2) \tag{3.9}$$

$$= p(X, \mathcal{G}_{12}) \int_0^{+\infty} p(\sigma^2|X, \mathcal{G}_{12}) \mathrm{d}(\sigma^2). \tag{3.10}$$

Proof. We formulate $p(X|\mathcal{G}_{12})$ as the integration of the joint density $p(X, \beta_{21}, \sigma^2|\mathcal{G}_{12})$

over β_{21} and σ^2

$$\begin{aligned}
p(X|\mathcal{G}_{12}) &= \int_{(0,+\infty)\times\mathbb{R}} p(X,\beta_{21},\sigma^2|\mathcal{G}_{12})\mathrm{d}(\beta_{21}\times\sigma^2) \\
&= \int_0^{+\infty}\int_{-\infty}^{+\infty} p(X,\beta_{21},\sigma^2|\mathcal{G}_{12})\mathrm{d}\beta_{21}\mathrm{d}(\sigma^2) \\
&= \int_0^{+\infty}\int_{-\infty}^{+\infty} p(X,\mathcal{G}_{12})p(\beta_{21},\sigma^2|X,\mathcal{G}_{12})\mathrm{d}\beta_{21}\mathrm{d}(\sigma^2) \\
&= p(X,\mathcal{G}_{12})\int_0^{+\infty}\int_{-\infty}^{+\infty} p(\beta_{12}|\sigma^2,X,\mathcal{G}_{12})p(\sigma^2|X,\mathcal{G}_{12})\mathrm{d}\beta_{21}\mathrm{d}(\sigma^2) \\
&= p(X,\mathcal{G}_{12})\int_0^{+\infty}\left(\int_{-\infty}^{+\infty} p(\beta_{12}|\sigma^2,X,\mathcal{G}_{12})\mathrm{d}\beta_{21}\right)p(\sigma^2|X,\mathcal{G}_{12})\mathrm{d}(\sigma^2) \\
&= p(X,\mathcal{G}_{12})\int_0^{+\infty} p(\sigma^2|X,\mathcal{G}_{12})\mathrm{d}(\sigma^2),
\end{aligned}$$

where the third equality could be achieved since $p(X,\mathcal{G}_{12})$ depends neither on σ^2 nor on β_{21}. The first equality of rewriting a doubled integral by a iterated integral follows from Fubini's theorem (see Royden and Fitzpatrick (1988)). □

Equation (3.9) and (3.10) are interesting in the sense that they show how the posterior densities of σ^2 and β_{21} will appear when we compute $p(X|\mathcal{G}_{12})$. It will be cumbersome to separately compute the two posterior densities. Luckily, now we can derive them simply by comparing the density functions. Then, we will compute $p(X|G)$ and derive the marginal posterior densities along the way. We will first get

Lemma 3.2. The marginal posterior distribution of β_{21} is

$$\beta_{21}|\sigma^2,X,\mathcal{G}_{12} \sim \mathcal{N}\left(\mu,\frac{\sigma^2}{\lambda}\right),\ \text{where } \mu=\frac{x_1x_2+U_{12}}{x_1^2+U_{11}},\lambda=x_1^2+U_{11}. \tag{3.11}$$

which gives a closed form expression for (3.8).

Proof. To get a closed form expression for $p(\mathcal{G}|X)$, we still need to compute $p(X|\mathcal{G})$, which follows from law of total probability

$$\begin{aligned}
p(X|\mathcal{G}_{12}) &= \int_{(0,+\infty)\times\mathbb{R}} p(X|\beta_{21},\sigma^2,\mathcal{G}_{12})p(\sigma^2,\beta_{21}|\mathcal{G}_{12})\mathrm{d}(\sigma^2,\beta_{21}) \\
&= \int_{(0,+\infty)\times\mathbb{R}} p(X|\beta_{21},\sigma^2,\mathcal{G}_{12})p(\beta_{21}|\sigma^2,\mathcal{G}_{12})p(\sigma^2|\mathcal{G}_{12})\mathrm{d}(\sigma^2,\beta_{21}) \\
&= \int_{(0,+\infty)\times\mathbb{R}} \frac{1}{\det(2\pi\Sigma)^{1/2}}\cdot\exp\left(-\frac{1}{2\sigma^2}(x_1,x_2)\begin{pmatrix}\beta_{21}^2+1 & -\beta_{21}\\ -\beta_{21} & 1\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}\right) \\
&\qquad\cdot\frac{1}{\sqrt{2\pi\frac{\sigma^2}{U_{11}}}}\exp\left(-\frac{(\beta_{21}-\frac{U_{12}}{U11})^2}{\frac{2\sigma^2}{U_{11}}}\right)\cdot\frac{\left(\frac{1}{2}U_{11}\right)^{\frac{1}{2}a_{12}}}{\Gamma\left(\frac{1}{2}a_{12}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}a_{12}+1} \\
&\qquad\cdot\exp\left(-\frac{U_{11}}{\sigma^2}\right)\mathrm{d}(\sigma^2,\beta_{21})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\Gamma(\frac{a_{12}}{2})} \int_{(0,+\infty)\times\mathbb{R}} \frac{1}{2\pi\sigma^2\sqrt{2\pi\frac{\sigma^2}{U_{11}}}} \cdot \left(\frac{U_{11}}{2\sigma^2}\right)^{\frac{1}{2}a_{12}+1} \cdot \frac{2}{U_{11}} \\
&\qquad \cdot \exp\Bigg(-\frac{1}{2\sigma^2}\left[(\beta_{21}^2+1)x_1^2 - 2\beta_{21}x_1x_2 + x_2^2\right] \\
&\qquad -\frac{1}{2\sigma^2}\left[U_{11}\beta_{21}^2 - 2U_{11}\beta_{21} + \frac{U_{12}^2}{U_{11}}\right] - \frac{U_{12}}{2\sigma^2}\Bigg) \mathrm{d}(\sigma^2,\beta_{21}) \\
&= \frac{1}{\Gamma(\frac{a_{12}}{2})U_{11}} \int_0^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{\pi\sigma^2\sqrt{2\pi\frac{\sigma^2}{U_{11}}}} \cdot \left(\frac{U_{11}}{\sigma^2}\right)^{\frac{1}{2}a_{12}+1} \\
&\qquad \cdot \exp\Bigg(-\frac{1}{2\sigma^2}\Big[\beta_{21}^2(x_1^2+U_{11}) - 2\beta_{21}(x_1x_2+U_{12}) \\
&\qquad + x_1^2 + x_2^2 + \frac{U_{12}^2}{U_{11}} + U_{11}\Big]\Bigg) \mathrm{d}\beta_{21}\mathrm{d}\sigma^2,
\end{aligned}
\tag{3.12}
$$

where $\det(\Sigma)$ denotes the determinant of a matrix and the last equality of rewriting a doubled integral by a iterated integral follows from Fubini's theorem.

Reformulating the expression inside $\exp(\cdot)$ in the form of a density function of normal distribution with respect to β_{21}, we have

$$
\begin{aligned}
&\exp\left(-\frac{1}{2\sigma^2}\left(x_1^2+x_2^2+\frac{U_{12}^2+U_{11}^2}{U_{11}} - \frac{(x_1x_2+U_{12})^2}{x_1^2+U_{11}}\right)\right) \\
&\qquad \cdot \exp\left(-\frac{1}{2\sigma^2}\left(\beta_{21}\sqrt{x_1^2+U_{11}} - \frac{x_1x_2+U_{12}}{\sqrt{x_1^2+U_{11}}}\right)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(x_1^2+x_2^2+\frac{U_{12}^2+U_{11}^2}{U_{11}} - \frac{(x_1x_2+U_{12})^2}{x_1^2+U_{11}}\right)\right) \\
&\qquad \cdot \exp\left(-\frac{x_1^2+U_{11}}{2\sigma^2}\left(\beta_{21} - \frac{x_1x_2+U_{12}}{x_1^2+U_{11}}\right)^2\right).
\end{aligned}
\tag{3.13}
$$

We further reformulate and simplify (3.12) as

$$
\begin{aligned}
&\frac{1}{\Gamma(\frac{a_{12}}{2})}\int_0^{+\infty} \frac{1}{2\pi\sigma^2}\left(\frac{U_{11}}{x_1^2+U_{11}}\right)^{\frac{1}{2}} \cdot \left(\frac{U_{11}}{\sigma^2}\right)^{\frac{1}{2}a_{12}+1} \\
&\qquad \cdot \left(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{\frac{2\pi\sigma^2}{x_1^2+U_{11}}}} \cdot \exp\left(-\frac{x_1^2+U_{11}}{2\sigma^2}\left(\beta_{21} - \frac{x_1x_2+U_{12}}{x_1^2+U_{11}}\right)^2\right)\mathrm{d}\beta_{21}\right) \\
&\qquad \cdot \exp\left(-\frac{1}{2\sigma^2}\left(x_1^2+x_2^2+\frac{U_{12}^2+U_{11}^2}{U_{11}} - \frac{(x_1x_2+U_{12})^2}{x_1^2+U_{11}}\right)\right)\mathrm{d}(\sigma^2).
\end{aligned}
\tag{3.14}
$$

It is interesting to see that (3.14) has exactly the same form as (3.9), and the integrated part inside the inner integral of (3.14) is actually the density function of a normal distribution. We can make two conclusions out of this. First, the inner integral can be simply reduced to 1. Second, we derive the posterior distribution $\beta_{21}|\sigma^2, X, \mathcal{G}_{12}$. □

Now as we continue to simplify (3.14), we could also derive the marginal posterior of σ^2.

Lemma 3.3. The marginal posterior distribution of σ^2 is

$$\begin{aligned}\sigma^2|X,\mathcal{G}_{12} \sim &\text{I-Ga}(a,b),\\ &\text{where } a=\frac{1}{2}a_{12}+1, b=\frac{1}{2}\left[x_1^2+x_2^2+\frac{U_{12}^2+U_{11}^2}{U_{11}}-\frac{(x_1x_2+U_{12})^2}{x_1^2+U_{11}}\right],\end{aligned} \tag{3.15}$$

which gives a closed form expression for (3.7).

Proof. We further simplify (3.14) to

$$\begin{aligned}&\frac{1}{\Gamma(\frac{a_{12}}{2})}\int_0^{+\infty}\frac{1}{2\pi\sigma^2}\left(\frac{U_{11}}{x_1^2+U_{11}}\right)^{\frac{1}{2}}\cdot\left(\frac{U_{11}}{\sigma^2}\right)^{\frac{1}{2}a_{12}+1}\\ &\quad\cdot\exp\left(-\frac{1}{2\sigma^2}\left(x_1^2+x_2^2+\frac{U_{12}^2+U_{11}^2}{U_{11}}-\frac{(x_1x_2+U_{12})^2}{x_1^2+U_{11}}\right)\right)\mathrm{d}(\sigma^2).\end{aligned} \tag{3.16}$$

The term inside the exponential component that does not depend on σ^2 can be reformulated as

$$\begin{aligned}&\frac{U_{11}x_1^2+2U_{11}^2x_1^2+U_{11}^2x_2^2+U_{12}^2x_1^2+U_{11}^3-2U_{11}U_{12}x_1x_2}{U_{11}(x_1^2+U_{11})}\\ =&\frac{U_{11}(x_1^2+U_{11})^2+(U_{11}x_2-U_{12}x_1)^2}{U_{11}(x_1^2+U_{11})}.\end{aligned} \tag{3.17}$$

We do the reformulation because it will have a better looking multiplicative inverse. Plugging (3.17) into (3.16) and turning what inside the integral sign into the form of a density function of a Inverse-Gamma distribution, we further rewrite (3.16) as

$$\begin{aligned}&\frac{a_{12}}{2\pi}\cdot\frac{U_{12}^{a_{12}+\frac{3}{2}}\cdot(x_1^2+U_{11})^{\frac{1}{2}a_{12}+\frac{1}{2}}}{[U_{11}(x_1^2+U_{11})^2+(U_{11}x_2-U_{12}x_1)^2]^{\frac{1}{2}a_{12}+1}}\\ &\quad\cdot\int_0^{+\infty}\frac{1}{\Gamma(\frac{1}{2}a_{12}+1)}\left(x_1^2+U_{11}+\frac{(U_{11}x_2-U_{12}x_1)^2}{U_{11}(x_1^2+U_{11})^2}\right)^{\frac{1}{2}a_{12}+1}\\ &\quad\cdot(2\sigma^2)^{-(\frac{1}{2}a_{12}+1)-1}\cdot\exp\left[-\frac{1}{2\sigma^2}\left(\frac{U_{11}(x_1^2+U_{11})^2+(U_{11}x_2-U_{12}x_1)^2}{U_{11}(x_1^2+U_{11})}\right)\right]\mathrm{d}(2\sigma^2),\end{aligned} \tag{3.18}$$

with exactly the same form as (3.10). Similar as before, we get two results out of the expression (3.18). First, we can simplify the integral to 1, since it is an integration of a probability density function over the whole range. Second, we derive the posterior distribution $\sigma^2|X, \mathcal{G}_{12}$. □

Now we come the the last step before we can derive the posterior of the graphical structure $\mathcal{G}$.

Lemma 3.4. The density function of the bivariate random variable $X = (x_1, x_2)$ given $\mathcal{G} = \mathcal{G}_{12}$ is

$$p(X|\mathcal{G}_{12}) = \frac{a_{12}}{2\pi} \cdot \frac{U_{11}^{a_{12}+\frac{3}{2}} \cdot (x_1^2 + U_{11})^{\frac{1}{2}a_{12}+\frac{1}{2}}}{[U_{11}(x_1^2 + U_{11})^2 + (U_{11}x_2 - U_{12}x_1)^2]^{\frac{1}{2}a_{12}+1}}. \tag{3.19}$$

The bivariate random variable $X = (x_1, x_2)$ given $\mathcal{G} = \mathcal{G}_{21}$ is

$$p(X|\mathcal{G}_{21}) = \frac{a_{21}}{2\pi} \cdot \frac{U_{22}^{a_{21}+\frac{3}{2}} \cdot (x_2^2 + U_{22})^{\frac{1}{2}a_{21}+\frac{1}{2}}}{[U_{22}(x_2^2 + U_{22})^2 + (U_{22}x_1 - U_{21}x_2)^2]^{\frac{1}{2}a_{21}+1}}. \tag{3.20}$$

Proof. The expression of $p(X|\mathcal{G}_{12})$ follows directly from (3.18) after setting the integral as 1. The density $p(X|\mathcal{G}_{21})$ can be derived analogously by just switching the positions of variable 1 and 2. □

Figure 3.1 displays the conditional density $(X|\mathcal{G}_{12})$ characterized by 30×30 points when we set $a_{12} = 2, U_{11} = 1$, and $U_{12} = 0$. Lemma 3.4 already indicates $p(\mathcal{G}|X)$ since it differs from $p(X|\mathcal{G})$ only by multiplying a constant under the prior assumption that both graphical structures are obtained with the same probability. Yet we still want to compute the exact density probability and represent the expression with simpler notations.

Lemma 3.5. We denote with a new notation

$$r_{12}^{21} := \frac{p(\mathcal{G}_{21}|X)}{p(\mathcal{G}_{12}|X)} = \frac{p(X|\mathcal{G}_{21})}{p(X|\mathcal{G}_{12})}, \tag{3.21}$$

and obtain the marginal posterior distribution of $\mathcal{G}$ as

$$p(\mathcal{G}|X) = p(\mathcal{G}_{12}|X)\mathbb{1}\{\mathcal{G} = \mathcal{G}_{12}\} + p(\mathcal{G}_{21}|X)\mathbb{1}\{\mathcal{G} = \mathcal{G}_{21}\}, \tag{3.22}$$

with

$$p(\mathcal{G}_{12}|X) = \frac{1}{1 + r_{12}^{21}}, \quad p(\mathcal{G}_{21}|X) = \frac{r_{12}^{21}}{1 + r_{12}^{21}}. \tag{3.23}$$

Proof. The equality in (3.21) holds based on our discussion that $p(\mathcal{G}|X)$ and $p(X|\mathcal{G})$ are proportional to each other. We obtain (3.23) by further considering the constraint that the probabilities of obtaining both graphs must add up to one, i.e.,

$$p(\mathcal{G}_{12}|X) + p(\mathcal{G}_{21}|X) = 1 \tag{3.24}$$

□

It is interesting to see that we could compute r_{12}^{21} from (3.19) and (3.20), which only depends on X and hyperparamters a_{12}, U_{11}, U_{12} that we assumed for the prior distribution. Until now, we are confident that we could sample $\mathcal{G}$, σ^2 and β one by one following this order from their corresponding posterior distributions as long as we know X.

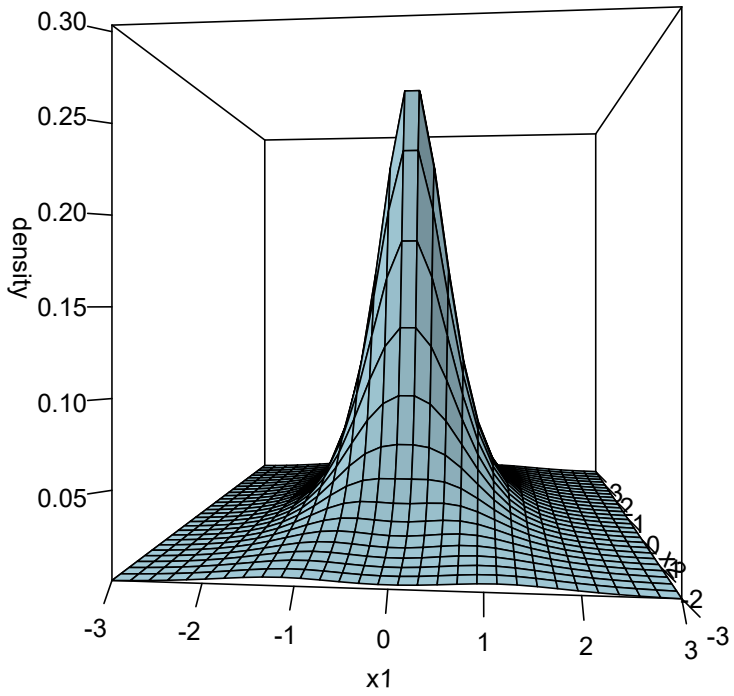


Figure 3.1.: The conditional density function of $X = (x_1, x_2)$ given $\mathcal{G} = \mathcal{G}_{12}$ for $a_{12} = 2, U_{11} = 1$, and $U_{12} = 0$.

3.1.2. Given i.i.d. Samples X^n

Especially for our later purpose of testing the model using a group of samples generated independently from the same distribution, we generalize the results in the previous section from random variable X to a set of n i.i.d. samples $X^n := \{X_1, X_2, \ldots, X_n\}$, where each $X_i = (x_{i,1}, x_{i,2})$ denotes a bivariate sample for $i = 1, \ldots, n$. Our goal remains unchanged, i.e., we still want to sample $\mathcal{G}, \sigma^2$ and β from their posterior distributions, yet given a pile of data sampled i.i.d. from the same distribution as X in the previous section. All the computations are similar as and are half done in the previous section, and now we give the expressions of the three marginal posterior densities.

Lemma 3.6. The posterior density of $\mathcal{G}$ given X^n is

$$p(\mathcal{G}|X) = p(\mathcal{G}_{12}|X)\mathbb{1}\{\mathcal{G} = \mathcal{G}_{12}\} + p(\mathcal{G}_{21}|X)\mathbb{1}\{\mathcal{G} = \mathcal{G}_{21}\} \tag{3.25}$$

with

$$p(\mathcal{G}_{12}|X) = \frac{1}{1+r_{12}^{21}}, \quad p(\mathcal{G}_{21}|X) = \frac{r_{12}^{21}}{1+r_{12}^{21}}, \tag{3.26}$$

where

$$r_{12}^{21} := \frac{\prod_{i=1}^n p(X_i|\mathcal{G}_{21})}{\prod_{i=1}^n p(X_i|\mathcal{G}_{12})} = \exp\left(\sum_{i=1}^n \log p(X_i|\mathcal{G}_{21}) - \sum_{i=1}^n \log p(X_i|\mathcal{G}_{12})\right), \tag{3.27}$$

and $p(X_i|\mathcal{G})$ has the same form as (3.19) and (3.20) for all $i \in \{1,\dots,n\}$, $\mathcal{G} \in \{\mathcal{G}_{12}, \mathcal{G}_{21}\}$.

The posterior distribution of β_{21} is

$$\beta_{21}|\sigma^2, X^n, \mathcal{G}_{12} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\lambda}\right), \text{ where } \mu = \frac{\sum_{i=1}^n x_{i,1}x_{i,2} + U_{12}}{\sum_{i=1}^n x_{i,1}^2 + U_{11}}, \lambda = \sum_{i=1}^n x_{i,1}^2 + U_{11}. \tag{3.28}$$

The posterior distribution of σ^2 is

$$\begin{aligned}&\sigma^2|X^n, \mathcal{G}_{12} \sim \text{I-Ga}(a,b),\\ &\text{where } a = \frac{1}{2}a_{12} + n, b = \frac{1}{2}\left[\sum_{i=1}(x_{i,1}^2 + x_{i,2}^2) + \frac{U_{12}^2 + U_{11}^2}{U_{11}} - \frac{(\sum_{i=1}^n x_{i,1}x_{i,2} + U_{12})^2}{\sum_{i=1}^n x_{i,1}^2 + U_{11}}\right].\end{aligned} \tag{3.29}$$

Proof. Generalizing results from the previous section, we find that the posterior distribution of $\mathcal{G}$ is

$$p(\mathcal{G}|X^n) \propto p(X^n|\mathcal{G})p(\mathcal{G}) = p(\mathcal{G})\prod_{i=1}^n p(X_i|\mathcal{G}), \tag{3.30}$$

which directly derives the r_{12}^{21} for n data. Then we compute the posterior distributions of the two parameters σ^2 and β through $p(X, \beta_{21}, \sigma^2|\mathcal{G}_{12})$. Recall from (3.12) that

$$\begin{aligned}p(X^n, \beta_{21}, \sigma^2|\mathcal{G}_{12}) &= p(X^n|\beta_{21}, \sigma^2, \mathcal{G}_{12})p(\sigma^2, \beta_{21}|\mathcal{G}_{12})\\ &= p(X^n|\beta_{21}, \sigma^2, \mathcal{G}_{12})p(\beta_{21}|\sigma^2, \mathcal{G}_{12})p(\sigma^2|\mathcal{G}_{12})\\ &= \frac{1}{\Gamma(\frac{a_{12}}{2})} \cdot \frac{1}{(2\pi\sigma^2)^n\sqrt{2\pi\frac{\sigma^2}{U_{11}}}} \cdot \left(\frac{U_{11}}{2\sigma^2}\right)^{\frac{1}{2}a_{12}+1} \cdot \frac{2}{U_{11}}\\ &\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n\left[(\beta_{21}^2+1)x_{i,1}^2 - 2\beta_{21}x_{i,1}x_{i,2} + x_{i,2}^2\right]\right.\\ &\quad \left. - \frac{1}{2\sigma^2}\left[U_{11}\beta_{21}^2 - 2U_{11}\beta_{21} + \frac{U_{12}^2}{U_{11}}\right] - \frac{U_{12}}{2\sigma^2}\right\}\end{aligned}$$

$$
\begin{aligned}
=&\frac{2}{\Gamma(\frac{a_{12}}{1})U_{11}}\cdot\frac{1}{(2\pi\sigma^2)^n\sqrt{2\pi\frac{\sigma^2}{U_{11}}}}\cdot\left(\frac{U_{11}}{2\sigma^2}\right)^{\frac{1}{2}a_{12}+1}\\
&\cdot\exp\left\{-\frac{1}{2\sigma^2}\left[\beta_{21}^2\left(\sum_{i=1}^n x_{i,1}^2+U_{11}\right)-2\beta_{21}\left(\sum_{i=1}^n x_{i,1}x_{i,2}+U_{12}\right)\right.\right.\\
&\left.\left.+\sum_{i=1}^n\left(x_{i,1}^2+x_{i,2}^2\right)+\frac{U_{12}^2}{U_{11}}+U_{11}\right]\right\}\\
\propto&\frac{1}{\sqrt{2\pi\frac{\sigma^2}{\sum_{i=1}^n x_{i,1}^2+U_{11}}}}\cdot\exp\left\{-\frac{\left(\beta_{21}-\frac{\sum_{i=1}^n x_{i,1}x_{i,2}+U_{12}}{\sum_{i=1}^n x_{i,1}^2+U_{11}}\right)^2}{2\sigma^2\left(\sum_{i=1}^n x_{i,1}^2+U_{11}\right)^{-1}}\right\}\\
&\cdot(2\sigma^2)^{-[(n+\frac{1}{2}a_{12})-1]}\cdot\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n(x_{i,1}^2+x_{i,2}^2)\right.\right.\\
&\left.\left.+\frac{U_{12}^2+U_{11}^2}{U_{11}}-\frac{\left(\sum_{i=1}^n x_{i,1}x_{i,2}+U_{12}\right)^2}{\sum_{i=1}^n x_{i,1}^2+U_{11}}\right\},
\end{aligned}
\tag{3.31}
$$

where the first line of (3.31) corresponds to the posterior distribution of β_{21}, while the second and third lines correspond to the posterior distribution of σ^2.

□

3.1.3. Discussion of Hyperparameters

Let $X^n=\{(x_{i,1},x_{i,2})\}_{i=1,\dots,n}$ be a group of n centered data pairs such that $(x_{i,1},x_{i,2})\overset{i.i.d.}{\sim}\mathcal{N}(0,\Sigma)$ for $i=i=1,\dots,n$, where Σ is a covariance matrix in general. In particular, We denote

$$\mathbf{X}_1=(x_{i,1})_{i=1,\dots,n}\overset{i.i.d.}{\sim}\mathcal{N}(0,\Sigma_{11})\text{ and }\mathbf{X}_2=(x_{i,2})_{i=1,\dots,n}\overset{i.i.d.}{\sim}\mathcal{N}(0,\Sigma_{22}).$$

It is clear that the normally distributed sequences $(x_{i,1})_{i=1,\dots,n}$ and $(x_{i,2})_{i=1,\dots,n}$ satisfy $\mathbb{E}[|x_{1,1}|]<\infty$ and $\mathbb{E}[|x_{1,2}|]<\infty$, which are the conditions of the strong law of large numbers (strong LLN) (see Theorem A.2). So we have

$$\frac{1}{n}\sum_{i=1}^n x_{i,1}\xrightarrow[n\to\infty]{a.s.}0,\qquad\frac{1}{n}\sum_{i=1}^n x_{i,2}\xrightarrow[n\to\infty]{a.s.}0.$$

Moreover, since $\mathbb{E}[x_{1,1}^2]=\sigma_1^2<\infty$ and $\mathbb{E}[x_{1,2}^2]=\sigma_2^2<\infty$. By the strong LLN we have

$$\frac{1}{n}\sum_{i=1}^n x_{i,1}^2\xrightarrow[n\to\infty]{a.s.}\Sigma_{11},\qquad\frac{1}{n}\sum_{i=1}^n x_{i,2}^2\xrightarrow[n\to\infty]{a.s.}\Sigma_{22}.$$

In addition, by Hölder's inequality (see Appendix A.4) we have $\mathbb{E}[|x_{i,1}x_{i,2}|]<\infty$. By the strong LLN we have

$$\frac{1}{n}\sum_{i=1}^n x_{i,1}x_{i,2}\xrightarrow[n\to\infty]{a.s.}\mathbb{E}[x_{i,1}x_{i,2}]=\Sigma_{12}.$$

Recall the formulas of priors (3.5), (3.4), posteriors (3.28), (3.29), and parameters a, b, μ, λ characterizing the above mentioned distributions. Following Castelletti and Mascaro (2022), we set $U_{11} = U_{22} \neq 0$, $U_{12} = U_{21} = 0$, and $a_{12} = 2$. We want to clarify under which conditions of setting the hyperparameters and rescaling the data would our estimation of the causal effect (asymptotically) make sense. We will support our statements by analyzing the asymptotic (approximated) behaviours of two statistics:

- μ, the posterior mean of β_{21};
- $p(\mathcal{G}|X^n)$, the posterior distribution of graphical structures.

Lemma 3.7. If $n\Sigma_{11} \gg U_{11}$, then μ is asymptotically equal to β_{21}. If $n\Sigma_{11} \ll U_{11}$, then μ is approximately equal to $n\beta_{21}\sigma^2/U_{11}$.

Proof. Discussing β_{21} only makes sense if we assume X_i follows the entailed distribution of model (M1) and the likelihood of X_i to be as (2.4) for all $i \in \{1, \ldots, n\}$, i.e., $x_{i,1} = \epsilon_{i,1}$, $x_{i,2} = \beta_{21}x_{i,1} + \epsilon_{i,2}$ with $\epsilon_{i,1}, \epsilon_{i,2} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Under this assumption we have $\Sigma_{11} = \sigma^2$, $\Sigma_{22} = (1 + \beta_{21}^2)\sigma^2$, and $\Sigma_{12} = \beta_{21}\sigma^2$.

Case 1. If $n\Sigma_{11} \gg U_{11}$, then by continuous mapping theorem (CMT) (see Theorem A.3) we have

$$\mu \xrightarrow[n\to\infty]{a.s.} \frac{\Sigma_{12}}{\Sigma_{11}} = \beta_{21}.$$

Case 2. If $n\Sigma_{11} \ll U_{11}$, then

$$\mu \approx \frac{n\Sigma_{12}}{U_{11}} = \frac{n\beta_{21}\sigma^2}{U_{11}},$$

which depends not only on β_{21}, but also on n and σ^2. □

Case 1 obtains an approximation of μ asymptotically, which is exactly what we would expect. Case 2 presents the case where the estimation is approximately incorrect, i.e., even if n is large, the estimation of μ will still be far from the true causal effect as long as U_{11} dominates over $n\Sigma_{11}$. Hence, we would always want to make sure that case 1 happens, and tries our best to prevent the setting of case 2 to happen.

Lemma 3.8. If $n\Sigma_{11} \gg U_{11}$, then assuming model (M1) to be true, we have

$$p(\mathcal{G}_{12}|X^n) > \frac{1}{2}$$

and $p(\mathcal{G}_{12}|X^n)$ increases as β_{21} increases; assuming model (M2) to be true, we have

$$p(\mathcal{G}_{21}|X^n) > \frac{1}{2}$$

and $p(\mathcal{G}_{21}|X^n)$ increases as β_{12} increases.
On the other hand, if $n\Sigma_{11} \ll U_{11}$, then assuming model (M1) to be true, we have

$$p(\mathcal{G}_{21}|X^n) \approx C \cdot \frac{1}{U_{11}};$$

assuming model (M2) to be true, we have

$$p(\mathcal{G}_{12}|X^n) \approx C \cdot \frac{1}{U_{22}}.$$

Proof. By (3.30) and the CMT, we have

$$p(\mathcal{G}_{12}|X^n) \xrightarrow[n\to\infty]{a.s.} C \cdot \frac{U_{11}^{a_{12}+\frac{3}{2}}(n\sigma^2+U_{11})^{\frac{1}{2}a_{12}+\frac{1}{2}}}{[U_{11}(n\sigma^2+U_{11})^2+U_{11}^2 n(1+\beta_{21}^2)\sigma^2]^{\frac{1}{2}a_{12}+1}},$$

where C is a normalizing constant.

Case 1. If $n\Sigma_{11} \gg U_{11}$. When we assume model (M1) to be true, we have

$$\begin{aligned} p(\mathcal{G}_{12}|X^n) &\approx C \cdot \frac{U_{11}^{a_{12}+\frac{3}{2}}(n\sigma^2)^{\frac{1}{2}a_{12}+\frac{1}{2}}}{[U_{11}(n\sigma^2)^2+U_{11}^2 n(1+\beta_{21}^2)\sigma^2]^{\frac{1}{2}a_{12}+1}} \\ &= C \cdot \frac{1}{(U_{11}n\sigma^2)^{\frac{1}{2}}\left[\frac{n\sigma^2}{U_{11}}+(1+\beta_{21}^2)\right]^{\frac{1}{2}a_{12}+1}} \\ &\approx C \cdot \frac{U_{11}^{\frac{1}{2}a_{12}+\frac{1}{2}}}{(n\sigma^2)^{\frac{1}{2}a_{12}+\frac{3}{2}}} \\ &= C \cdot \frac{1}{U_{11}}\left(\frac{U_{11}}{n\sigma^2}\right)^{\frac{1}{2}a_{12}+\frac{3}{2}}, \end{aligned}$$

where the last approximation is based on our assumption that β_{21} is not that large and hence $\frac{n\sigma^2}{U_{11}} \gg 1+\beta_{21}$. Similarly, we have

$$p(\mathcal{G}_{21}|X^n) \approx C \cdot \frac{1}{U_{22}}\left(\frac{U_{22}}{n(1+\beta_{21}^2)\sigma^2}\right)^{\frac{1}{2}a_{21}+\frac{3}{2}}.$$

Hence,

$$r_{12}^{21} \approx \left(\frac{1}{1+\beta_{21}^2}\right)^{\frac{1}{2}a_{12}+\frac{3}{2}} < 1$$

and decreases as β_{21} increases. More precisely, we have

$$p(\mathcal{G}_{12}|X^n) = \frac{1}{1+r_{12}^{21}} > \frac{1}{2}, \qquad p(\mathcal{G}_{21}|X^n) = \frac{r_{12}^{21}}{1+r_{12}^{21}} < \frac{1}{2},$$

and when β_{21} increases, r_{12}^{21} decreases causing $p(\mathcal{G}_{12}|X^n)$ to increase and $p(\mathcal{G}_{21}|X^n)$ to decrease, i.e., we are more confident to predict $\mathcal{G} = \mathcal{G}_{12}$ for larger β_{21}.

When we assume instead model (M2) to be true, we have analogously

$$r_{12}^{21} \approx \left(\frac{1+\beta_{12}^2}{1}\right)^{\frac{1}{2}a_{21}+\frac{3}{2}} > 1,$$

where r_{12}^{21} increases with β_{12}. More precisely, we have

$$p(\mathcal{G}_{12}|X^n) = \frac{1}{1+r_{12}^{21}} < \frac{1}{2}, \qquad p(\mathcal{G}_{21}|X^n) = \frac{1}{1+r_{12}^{21}} > \frac{1}{2}.$$

When β_{12} increases, $p(\mathcal{G}_{12}|X^n)$ decreases and $p(\mathcal{G}_{21}|X^n)$ increases, i.e., we are more confident to predict $\mathcal{G} = \mathcal{G}_{21}$, leading us to predict $\mathcal{C}(1 \to 2) = 0$.

Case 2. If $n\Sigma_{11} \ll U_{11}$. Assuming model (M1) to be true, we have

$$\begin{aligned} p(\mathcal{G}_{12}|X^n) &\approx C \cdot \frac{U_{11}^{a_{12}+\frac{3}{2}} \cdot U_{11}^{\frac{1}{2}a_{12}+\frac{1}{2}}}{[U_{11} \cdot U_{11}^2 + U_{11}^2 n(1+\beta_{12}^2)\sigma^2]^{\frac{1}{2}a_{12}+1}} \\ &= C \cdot \frac{1}{U_{11}\left[1+\frac{n\sigma^2}{U_{11}}(1+\beta_{21}^2)\right]^{\frac{1}{2}a_{12}+1}} \\ &\approx C \cdot \frac{1}{U_{11}}, \end{aligned}$$

where the last approximation is based on our assumption that $\frac{n\sigma^2}{U_{11}} \approx 0$ and $\frac{U_{11}}{n\sigma^2} \gg 1+\beta_{21}^2$. Similarly we have

$$p(\mathcal{G}_{21}|X^n) \approx C \cdot \frac{1}{U_{22}}.$$

Under the assumption that $U_{11} = U_{22}$, we have $r_{12}^{21} \approx 1$, and further leading to $p(\mathcal{G}_{12}|X^n) \approx p(\mathcal{G}_{21}|X^n) \approx \frac{1}{2}$ regardless of the causal effect β_{21} and data X^n, □

Case 1 obtains a reasonable estimation for $p(\mathcal{G}|X^n)$ while case 2 ends up estimating the probabilities of graphical structures to be approximately the same as coin tossing, which does not make much sense.

Considering Lemma 3.7 and Lemma 3.8, we want the covariance of X to be at least as large as the hyperparameters U_{11} and U_{22}. In other word, we definitely need to avoid the case that $U_{11} \gg n\sigma^2$ when applying the method on datasets. On the other hand, having a too small U_{11}, i.e. for $U_{11} \ll \sigma^2$, will make the Bayesian approach pointless, since the prior belief is too weak and can almost be neglected. So when doing numerical simulations, we set U_{11} to be of the same range as σ^2. This discussion is an important instruction of how we should set hyperparameters reasonably when we test with simulated data and real world benchmark data.

3.2. Posterior distribution of Causal Effect

From (2.6) we recall that the causal effect $\mathcal{C}(1 \to 2)$ equals β_{21} when we assume $\mathcal{G} = \mathcal{G}_{12}$, and $\mathcal{C}(1 \to 2)$ equals 0 when we assume $\mathcal{G} = \mathcal{G}_{21}$. Hence to compute the posterior distribution of the causal effect given data X^n we first need to compute the distribution of β_{21} given $X^n, \mathcal{G}_{12}$, then sum both cases up by a mixture of distributions.

Lemma 3.9. Let t_ν denote a Student's t-distribution with $\nu > 0$ degrees of freedom. Assume we have a random variable $Z \sim t_{2a}$, then the marginal posterior distribution of β_{21} is

$$\beta_{21}|X^n, \mathcal{G}_{12} = \sqrt{\frac{b}{a\lambda}}Z + \mu \sim t_{2a}\left(\mu, \frac{b}{a\lambda}\right)$$

with λ, μ defined in (3.28) and a, b defined in (3.29), which is a generalized Student's t-distribution (see Appendix B.2) of the form $t_\nu(\mu, \sigma^2)$ with location parameter μ, scale parameter $\sigma > 0$, and degrees of freedom $\nu > 0$.

Proof. Recall that from (3.28) and (3.29) we have

$$p(\sigma^2|X^n, G_{12}) = \text{I-Ga}(a, b),$$
$$p(\beta_{21}|\sigma^2, X^n, \mathcal{G}_{12}) = \mathcal{N}\left(\mu, \frac{\sigma^2}{\lambda}\right).$$

Following Marin and Robert (2014) we have the marginal posterior distribution

$$\sqrt{\frac{a\lambda}{b}}(\beta_{21} - \mu)\Big|X^n, \mathcal{G}_{12} \sim t_{2a}. \tag{3.32}$$

□

Formula (3.32) is useful for computing corresponding quantiles and creating credible intervals, just as what we usually do to compute confidence intervals using statistics with known distribution. The posterior distribution of β_{21} that we claimed in the lemma can also be easily derived from (3.32).

Having everything prepared, we could finally derive the posterior distribution of the causal effect $\mathcal{C}(1 \to 2)$.

Proposition 3.10. The posterior distribution of $\mathcal{C}(1 \to 2)$ is a mixture of a point mass and a Student's t-distribution characterized by the density function

$$p(\mathcal{C}(1 \to 2)|X^n) = \frac{1}{1 + r_{12}^{21}}\, t_{2a}\left(\mu, \frac{b}{a\lambda}\right) + \frac{r_{12}^{21}}{1 + r_{12}^{21}}\mathbb{1}\{\mathcal{C}(1 \to 2) = 0\}, \tag{3.33}$$

with λ, μ defined in (3.28) and a, b defined in (3.29).

Proof. From Lemma 3.9 we know that $\mathcal{C}(1 \to 2) = 0$ with probability 1 when $\mathcal{G} = \mathcal{G}_{21}$. Together with Lemma 3.6, expression (2.6) and the law of total probability, we have the posterior distribution of causal effects

$$\begin{aligned} p(\mathcal{C}(1 \to 2)|X^n) &= p(\mathcal{C}(1 \to 2)|X^n, \mathcal{G}_{12})p(\mathcal{G}_{12}|X^n) + p(\mathcal{C}(1 \to 2)|X^n, \mathcal{G}_{21})p(\mathcal{G}_{21}|X^n) \\ &= \frac{1}{1 + r_{12}^{21}}p(\beta_{21}|X^n, \mathcal{G}_{12}) + \frac{r_{12}^{21}}{1 + r_{12}^{21}}p(0|X^n, \mathcal{G}_{21}) \\ &= \frac{1}{1 + r_{12}^{21}}\, t_{2a}\left(\mu, \frac{b}{a\lambda}\right) + \frac{r_{12}^{21}}{1 + r_{12}^{21}}\mathbb{1}\{\mathcal{C}(1 \to 2) = 0\}. \end{aligned}$$

□

An example of the density function of $\mathcal{C}(1 \to 2)$ is illustrated in Figure 3.2.

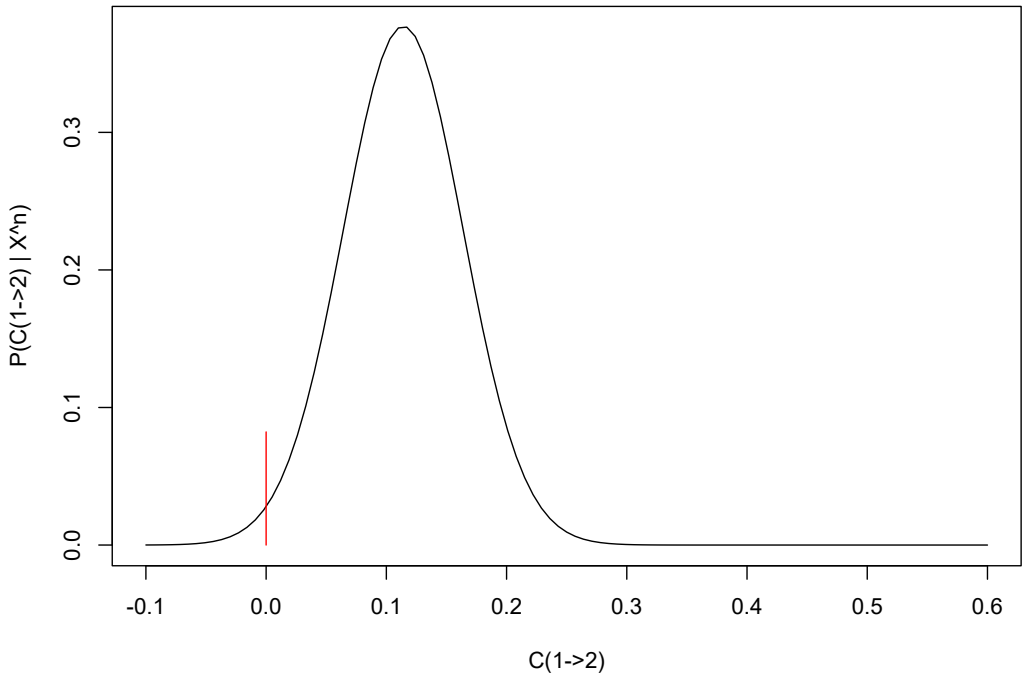


Figure 3.2.: The density function of the posterior distribution of causal effect from 1 on 2, when we set $a_{12} = a_{21} = 2$, $U_{11} = U_{22} = 1$ and $U_{12} = U_{21} = 0$. We simulate 200 data pairs $(x_{i,1}, x_{i,2})$ where $x_{i,1}$ is generated i.i.d. from standard normal distribution, and $x_{i,2} = \beta_{21} x_{i,1} + \epsilon_i$, where we set $\beta_{12} = 0.2$ and ϵ_i i.i.d. sampled from standard normal distribution for $i \in \{1, \ldots, 200\}$. The red vertical line corresponds to the point mass at 0.

3.3. Credible Regions

A way of summarizing the posterior distribution of the causal effect is the credible intervals (CIs) Kruschke (2014); Makowski et al. (2019), or more generally speaking, credible regions (CRs). Note that we will not distinguish between the terminologies CI and CR and will use both interchangeably.

The CR is an important concept in Bayesian statistics, mainly used to describe and summarize the remaining uncertainty when estimating a parameter. It is the Bayesian equivalent of a confidence interval in frequentist statistics. However, they have quite different statistical meanings. As doing Bayesian inference, we would return a posterior distribution that describes our current beliefs of the parameters of interest given the available data. The CR is thus a set of more plausible values that we believe contain the parameter with a particular probability. We could more formally define a CR following Jackman (2009).

Definition 3.11 (Credible Regions)**.** Let $\theta \in \Omega$ denote parameters in general, and let X denote random variables or data, then $C \subset \Omega$ is a $(1-\alpha)$**-credible region (CR)** for θ under posterior distribution given X if

$$\mathbb{P}(\theta \in C|X) \geq 1-\alpha, \quad \text{for } 0 \leq \alpha \leq 1, \tag{3.34}$$

and C is the smallest region under certain conditions satisfying (3.34). For single-parameter problems (i.e. $C \subset \mathbb{R}$), if C is not a set of disjoint intervals, then C is a **credible interval (CI)**.

As it was also argued in Jackman (2009), a CR might not be unique.

Remark 3.12. There is trivially only one 100%-credible region that is the entire support of $\theta|X$. But non-trivial CRs may not be unique, since any region spanning the $(1-\alpha)$ percentiles could be a CR according to our definition. We could solve this problem by restricting attention to CRs that have certain desirable properties. We will discuss them in details in the following sections.

Now we will give a common setup for a CR as an example.

Example 3.13. As in frequentist statistics, we usually set $\alpha = 5\%$ and search for a 95%-credible region. This means that given the observed data, there is a 95% probability that the true value of θ lies within the CR.

In our case, we want to estimate the CRs of $\mathcal{C}(1 \to 2)$ based on its posterior distribution. Since it is a mixture distribution with a point mass at 0 and continuous distribution in $\mathbb{R}$, the CR has three possible formats: an interval, the union of an interval and point $\{0\}$, or the difference of an interval and point $\{0\}$. So we always determine the CR with a two-step approach:

1. determine whether 0 is included in the CR;
2. compute the interval part by defining its lower bound l and upper bound u.

This approach will be used repeatedly later when we compute different types of CRs. In general, we could determine the lower and upper bounds with the following lemma.

Lemma 3.14. Set C is a CR of $\mathcal{C}(1 \to 2)$ if it is the smallest region under certain conditions satisfying

$$\begin{aligned}
&\mathbb{P}(\mathcal{C}(1 \to 2) \in C|X^n) \\
&= p(\mathcal{G}_{21}|X^n)\mathbb{P}(0 \in C|X^n, \mathcal{G}_{21}) + p(\mathcal{G}_{12}|X^n)\mathbb{P}(\beta_{21} \in C|X^n, \mathcal{G}_{12}) && (3.35)\\
&= p(\mathcal{G}_{21}|X^n)\mathbb{1}\{0 \in C\} + p(\mathcal{G}_{12}|X^n)\left(F\left(\sqrt{\frac{a\lambda}{b}}(u-\mu)\right) - F\left(\sqrt{\frac{a\lambda}{b}}(l-\mu)\right)\right) && (3.36)\\
&\geq 1-\alpha.
\end{aligned}$$

We decide whether to include 0 in the CR or not by the posterior distribution of the graphical structure $p(\mathcal{G}|X^n)$. When $\mathcal{G}|X^n$ has probability densities of special values, e.g., one DAG has probability 0 or the two DAGs have the same probabilities, we would obtain some interesting results in these extreme cases. Here we need to make a basic assumption that $0.5 < \alpha \leq 1$ since we are not interested in credible regions with a theoretical coverage of less than 50%. We will observe the following special cases from a quantitative perspective, where we will use approximations without formal definition. Please also note that the following computations are valid only under the assumption that we have the true priors.

Remark 3.15. When $p(\mathcal{G}_{12}|X^n) \approx p(\mathcal{G}_{21}|X^n) \approx 1/2$ and the true causal effect $\beta_{21} = 0$, then the probability of the credible interval C containing the true causal effect is $\mathbb{P}(0 \in C|X^n) = 1 - \alpha$.

Proof. Solving (3.35)$= 1 - \alpha$ under the assumption $\beta_{21} = 0$, we obtain the results. □

Remark 3.16. When $p(\mathcal{G}_{12}|X^n) \approx p(\mathcal{G}_{21}|X^n) \approx 1/2$ and the true causal effect β_{21} is very close to but not equal to 0, then the probability of the credible interval C containing the true causal effect is $\mathbb{P}(\beta_{21} \in C|X^n) \approx 1 - 2\alpha$.

Proof. First of all, we need to have $\mathbb{P}(0 \in C|X^n, \mathcal{G}_{21}) = 1$. Otherwise, we would destroy the assumption of a CI to have at least probability $1 - \alpha$, since $p(\mathcal{G}_{21}|X^n) \approx 1/2$. Under this setting, we could rewrite (3.35) as

$$\begin{aligned} &\frac{1}{2}\mathbb{P}(\beta_{21} \in C|X^n, \mathcal{G}_{12}) + \frac{1}{2} \approx 1 - \alpha \\ \Rightarrow\ &\mathbb{P}(\beta_{21} \in C|X^n, \mathcal{G}_{12}) \approx 1 - 2\alpha. \end{aligned} \tag{3.37}$$

□

We see that even if in the worst case when the model can hardly tell the correct causal direction, the $(1 - \alpha)$-credible interval we computed shall contain the true causal effect at least with probability $1-2\alpha$, e.g., when $\alpha = 0.05$, we have $\mathbb{P}(\beta_{21} \in C|X^n, \mathcal{G}_{12}) \approx 90\%$.

Remark 3.17. When $p(\mathcal{G}_{12}|X^n) \approx 1, p(\mathcal{G}_{12}|X^n) \approx 0$, then the probability of the credible interval C containing the true causal effect is $\mathbb{P}(\beta_{21} \in C|X^n) \approx 1 - \alpha$.

Proof. In this case, we are very sure that $1 \to 2$ is the correct causal direction, and solving (3.35) we get the result. □

For example, when $\alpha = 0.05$, we have $\mathbb{P}(\beta_{21} \in C|X^n, \mathcal{G}_{12}) \approx 95\%$.

Remark 3.18. When $p(\mathcal{G}_{12}|X^n) \approx \alpha, p(\mathcal{G}_{21}|X^n) \approx 1 - \alpha$, then the probability of the credible interval C containing the true causal effect is $\mathbb{P}(0 \in C|X^n, \mathcal{G}_{21}) = 1$.

Proof. Solving (3.35)=$1-\alpha$ we have

$$\begin{aligned}\mathbb{P}(\beta_{21}|X^n,\mathcal{G}_{12}) &= \frac{(1-\alpha)-P(\mathcal{G}_{21}|X^n)}{p(\mathcal{G}_{12}|X^n)}\\ &= \frac{(1-\alpha)-(1-P(\mathcal{G}_{21}|X^n))}{p(\mathcal{G}_{12}|X^n)}\\ &= \frac{P(\mathcal{G}_{21}|X^n)-\alpha}{p(\mathcal{G}_{12}|X^n)}.\end{aligned} \tag{3.38}$$

If $(\mathcal{G}_{12}|X^n)\to\alpha$, then (3.38) converges to 0, thus a non-zero causal effect will be almost surely not included in the credible interval. □

Remark 3.19. When $0<p(\mathcal{G}_{12}|X^n)<\alpha$ and $1-\alpha<p(\mathcal{G}_{21}|X^n)<1$, we would expect $\mathbb{P}(\beta_{21}\in C|X^n,\mathcal{G}_{12})=0$, with C being the $(1-\alpha)$-credible interval.

Proof. Solving (3.35)$\geq 1-\alpha$ we have

$$\mathbb{P}(\beta_{21}\in C|X^n,\mathcal{G}_{12})\geq\frac{(1-\alpha)-p(\mathcal{G}_{21}|X^n)}{p(\mathcal{G}_{12}|X^n)}, \tag{3.39}$$

of which the right-hand side is less than 0. Since $\mathbb{P}(\beta_{21}\in c|X^n,\mathcal{G}_{12})$ is a probability, it has to be greater or equal to 0. Hence the inequality (3.39) always holds. □

The above five remarks show the theoretical coverage rates of the CRs when we have different posterior distributions of $\mathcal{G}$, which could be used as instructions for evaluating our model.

Before we continue, we define some notations:

- let F_{t_ν} and f_{t_ν} denote the cumulative distribution function and the probability density function of a Student's t-distribution with ν degrees of freedom resp.;
- let F and f denote the cumulative distribution function and the probability density function of $\sqrt{\frac{b}{a\lambda}}Z+\mu$ resp., where Z a Student's t-distribution with ν degrees of freedom, and a,b,λ,μ are constants we computed before;
- let $q_{t_\nu}(p)$ denote the p-quantile of a Student's t-distribution with ν degrees of freedom, whose value could not be computed analytically, but can be found in the quantile table.

3.3.1. Equal-Tailed Interval

The first type of CI that we consider is the equal-tailed interval (ETI) (see Kruschke (2014)), which is used by some authors and software because it is easy to compute. As already indicated by its name, we require the tails on both sides to be equal and we get indeed an interval. We define the ETI more formally as the following.

Definition 3.20 (Equal-Tailed Interval)**.** An interval $C = [l, u]$ is the **equal-tailed interval** for a parameter θ under its posterior distribution given some data X if

1. $\mathbb{P}(\theta \in [l, u]|X) \geq 1 - \alpha$;
2. there exists some $\epsilon \geq 0$ s.t. $\mathbb{P}(\theta > u|X) = \mathbb{P}(\theta < l|X) \leq \alpha/2 - \epsilon$;
3. l, u are values such that ϵ in 2. reaches its minimum.

Next, we will discuss how ETI could be analytical represented by first analyzing whether 0 is included, then deciding the lower and upper bounds of the interval.
Case 1. ETI includes 0.

Case 1.1. $\mu \geq 0$.
When r_{12}^{21} satisfies the inequality constraint

$$\frac{1}{1 + r_{12}^{21}} F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (l - \mu) \right) + \frac{r_{12}^{21}}{1 + r_{12}^{21}} > \frac{\alpha}{2}, \tag{3.40}$$

ETI will contain 0, since otherwise the left tail will be too heavy. The lower and upper bounds l, u of the ETI satisfies

$$\begin{cases} \mathbb{P}(l \leq \mathcal{C}(1 \to 2) \leq u|X^n) \geq 1 - \alpha \\ \mu - l = u - \mu \end{cases}, \tag{3.41}$$

where

$$\mathbb{P}(l \leq \mathcal{C}(1 \to 2) \leq u|X^n) = \frac{r_{12}^{21}}{1 + r_{12}^{21}} + \frac{1}{1 + r_{12}^{21}} \left(F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (u - \mu) \right) - F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (l - \mu) \right) \right).$$

Plugging the second equation in (3.41) into the first inequality we get

$$\begin{aligned}
& 1 - 2F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (l - \mu) \right) \geq (1 - \alpha)(1 + r_{12}^{21}) - r_{12}^{21} \\
\Rightarrow \quad & 2F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (l - \mu) \right) \leq 1 + r_{12}^{21} - (1 - \alpha)(1 + r_{12}^{21}) = \alpha(1 + r_{12}^{21}) \\
\Rightarrow \quad & F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (l - \mu) \right) \leq \frac{\alpha}{2}(1 + r_{12}^{21}) \\
\Rightarrow \quad & \sqrt{\frac{a\lambda}{b}} (l - \mu) = q_{t_{2\alpha}} \left(\frac{\alpha}{2}(1 + r_{12}^{21}) \right) \\
\Rightarrow \quad & l = \sqrt{\frac{b}{a\lambda}} q_{t_{2\alpha}} \left(\frac{\alpha}{2}(1 + r_{12}^{21}) \right) + \mu.
\end{aligned} \tag{3.42}$$

Analogously, we have

$$u = 2\mu - l = \sqrt{\frac{b}{a\lambda}} q_{t_{2\alpha}} \left(1 - \frac{\alpha}{2}(1 + r_{12}^{21}) \right) + \mu. \tag{3.43}$$

If the interval $[l, u]$ we computed does not contain 0, i.e. if $l > 0$, it is against our assumption, and we need to extend the lower boundary of the ETI to 0. In order to let both tails to have equal probabilities, we extend the upper bound to 2μ because of the symmetry of the Student's t-distribution. On the other hand, if $r_{12}^{21}/(1 + r_{12}^{21}) \geq 1 - \alpha$ we might obtain $u \leq l$, which shall not be the case since upper bound shall always be greater than the lower bound. In this case, 0 itself obtains probability larger than $1 - \alpha$, and the ETI is constructed based on symmetry and equal-tailed property to be $[0, 2\mu]$. In short, if $l > 0$ or $u \leq l$, we set ETI $= [0, 2\mu]$. Otherwise, we simply set ETI $= [l, u]$ with l defined by (3.42) and u defined by (3.43).

Case 1.2. $\mu < 0$.

When r_{12}^{21} satisfies the inequality constraint

$$\frac{1}{1 + r_{12}^{21}} \left(1 - F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (u - \mu) \right) \right) + \frac{r_{12}^{21}}{1 + r_{12}^{21}} > \frac{\alpha}{2}, \tag{3.44}$$

the ETI will include 0. We obtain the same l, u as (3.42) and (3.43) by solving (3.41). With similar discussion as in Case 1.1, we know that if $u < 0$ or $u \leq l$ we set ETI=$[2\mu, 0]$, otherwise we set ETI=$[l, u]$ with l defined by (3.42) and u defined by (3.43).

Case 2. 0 is excluded from ETI.

If constraints of Case 1.1 is not satisfied, i.e. $p(\mathcal{G}_{21}|X^n)$ is a very low probability and 0 is far away from the mean μ, then we could compute the ETI with the following two cases.

Case 2.1. $\mu \geq 0$.

In this case, 0 will be included in the left tail. By the equal-tailed property, both tails have probability $\alpha/2$. Looking at the left tail, we have

$$\begin{aligned} & \frac{1}{1 + r_{12}^{21}} F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (l - \mu) \right) + \frac{r_{12}^{21}}{1 + r_{12}^{21}} = \frac{\alpha}{2} \\ \Rightarrow & F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (l - \mu) \right) = \frac{\alpha}{2} (1 + r_{12}^{21}) - r_{12}^{21} \\ \Rightarrow & l = \sqrt{\frac{b}{a\lambda}} q_{t_{2a}} \left(\frac{\alpha}{2} (1 + r_{12}^{21}) - r_{12}^{21} \right) + \mu. \end{aligned} \tag{3.45}$$

Observing the right tail, we have

$$\begin{aligned} & \frac{1}{1 + r_{12}^{21}} \left(1 - F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (u - \mu) \right) \right) = \frac{\alpha}{2} \\ \Rightarrow & F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}} (u - \mu) \right) = 1 - \frac{\alpha}{2} (1 + r_{12}^{21}) \\ \Rightarrow & u = \sqrt{\frac{b}{a\lambda}} q_{t_{2a}} \left(1 - \frac{\alpha}{2} (1 + r_{12}^{21}) \right) + \mu. \end{aligned} \tag{3.46}$$

The l we just computed is always larger than 0, otherwise it is against the assumption. Hence, we could directly set ETI= $[l, u]$ with l defined by (3.45) and u defined by (3.46).

Case 2.2. $\mu < 0$.

This case is analogous to Case 2.1, except here 0 is located in the right tail. Also by the equal-tailed property, both tails have probability $\alpha/2$. Looking at the left tail, we have

$$\frac{1}{1+r_{12}^{21}} F_{t_{2a}}\left(\sqrt{\frac{a\lambda}{b}}(l-\mu)\right) = \frac{\alpha}{2}$$

$$\Rightarrow l = \sqrt{\frac{b}{a\lambda}} q_{t_{2a}}\left(\frac{\alpha}{2}(1+r_{12}^{21})\right) + \mu. \tag{3.47}$$

The upper bound could be computed from the probability of the right tail

$$\frac{1}{1+r_{12}^{21}}\left(1 - F_{t_{2a}}\left(\sqrt{\frac{a\lambda}{b}}(u-\mu)\right)\right) + \frac{r_{12}^{21}}{1+r_{12}^{21}} = \frac{\alpha}{2}$$

$$\Rightarrow u = \sqrt{\frac{b}{a\lambda}} q_{t_{2a}}\left(\left(1-\frac{\alpha}{2}\right)(1+r_{12}^{21})\right) + \mu. \tag{3.48}$$

For similar reason as in Case 2.1, we obtain ETI= $[l, u]$ with l defined by (3.47) and u defined by (3.48).

We could summarize our procedure of computing the ETI with Algorithm 1, where we denote the hyperparameters of the prior distributions as $\mathbf{a} = (a_{12}, a_{21}), U = (U_{ij})_{i,j=1,2}$, the n i.i.d. samples as X^n, the credible level as α, and posterior probability of the graphical structure as $p(\mathcal{G}|X^n)$. We write $p(\mathcal{G}|X^n)$ as an array of length 2 with $p(\mathcal{G}_{12}|X^n)$ being the first entry and $p(\mathcal{G}_{21}|X^n)$ being the second entry. We will keep using these notations when we later write algorithms.

In our experiments, we will refer to the method describe in Algorithm 1 as `ETI`.

3.3.2. Highest Density Region

Another way of defining a CR is to choose the highest density region (HDR) (see Kruschke (2014); Jackman (2009)). Some authors refer to this as the highest density interval (HDI), since for most continuous distributions, the region with the highest density is indeed an interval. However in our case, the posterior distribution of $\mathcal{C}(1 \to 2)$ is a mixture of a continuous distribution on $\mathbb{R}$ and a point mass at 0, hence it is very likely for the HDR not to be an interval so we stick to the terminology HDR. However, we do not distinguish between HDI and HDR.

Intuitively speaking, the HDR indicates which points of a distribution are most credible and cover most of the distribution. In other words, all values inside the HDR have higher probability densities (i.e., credibility) than any value outside the HDR. We define an HDR following the definitions from Jackman (2009).

Definition 3.21. A region C is the **highest density region** for a parameter θ under its posterior distribution given some data X if

Algorithm 1 Compute the CR as an ETI

Input: $\mathbf{a}, U, X^n, \alpha, p(\mathcal{G}|X^n)$

Output: a $(1-\alpha)$-ETI of $\mathcal{C}(1 \to 2)$ with respect to its posterior distribution given X^n

Compute the posterior mean μ of the posterior distribution of $\beta_{21}|X^n$ by (3.28);
if (3.40) is satisfied when $l = 0$, and $p(\mathcal{G}_{21}|X^n) < 1 - \alpha$ **then**
    Set l as in (3.42) and u as in (3.43);
    if $\mu \geq 0$ **then**
        Set $\tilde{l} = \min(0, l), \tilde{u} = \max(u, 2\mu)$ and let $C = [\tilde{l}, \tilde{u}]$;
    else
        Set $\tilde{l} = \min(2\mu, 0), \tilde{u} = \max(0, u)$ and let $C = [\tilde{l}, \tilde{u}]$;
    end if
else if $p(\mathcal{G}_{21}|X^n) \geq 1 - \alpha$ **then**
    Set $l = \min(0, 2\mu)$, $u = \max(0, 2\mu)$ and $C = [l, u]$;
else
    if $\mu \geq 0$ **then**
        Set $C = [l, u]$ with l in (3.45) and u in (3.46);
    else
        Set $C = [l, u]$ with l in (3.47) and u in (3.48);
    end if
end if
return the credible interval C as a ETI

1. $\mathbb{P}(\theta \in C|X) \geq 1 - \alpha$;
2. $p(\theta_1|X) \geq p(\theta_2|X),\ \forall \theta_1 \in C,\ \theta_2 \notin C$.

Our goal is to find the values with the highest density from the mixture distribution. Since we have an analytical solution of the mixture distribution, we could always compute the density of any point with the help of the probability density function. However, for a mixture distribution, the most difficult part is to decide when 0 should be included in the HDR, since we need to compare the densities of a point mass and a continuous distribution. We will discuss with three cases.

Case 1. HDR includes 0 and not only 0.

The following constraints need to be satisfied

1. the HDR needs to obtain a probability of at least $1 - \alpha$, i.e., the probability of the tail have to be less than α;
2. $p(\mathcal{G}_{21}|X^n) < 1 - \alpha$;
3. $p(\mathcal{G}_{21}|X^n)$ needs to be large enough to be selected into the HDR.

We formulate the constraints mathematically as

$$\frac{r_{12}^{21}}{1+r_{12}^{21}} + \frac{1}{1+r_{12}^{21}}\left(F(l)+1-F(u)\right) \leq \alpha,$$
$$\text{where } l = \left\{ f^{-1}\left(\frac{r_{12}^{21}}{1+r_{12}^{21}}\right)\right\}_{\min},\ u = \left\{ f^{-1}\left(\frac{r_{12}^{21}}{1+r_{12}^{21}}\right)\right\}_{\max}, \tag{3.49}$$

where f^{-1} denotes the general inverse of f, i.e., the inverse mapping of the density function f. Since f is not injective, the inverse image $f^{-1}(x)$ of any well-defined $x \in (0, +\infty)$ is a set of two elements, and we use "max" and "min" to denote the larger and smaller element in the set resp.. We could see an example of the general inverse function f^{-1} from Figure 3.3, which we draw empirically using 200 data pairs from model (M1) with $\beta = 0.2$, $\epsilon_1, \epsilon_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $a_{12} = 2$, $U_{11} = 1$, and $U_{12} = 0$.

What we described above is the boundary condition when $p(\mathcal{G}_{21}|X^n)$ is just as large as $p(\mathcal{G}_{12}|X^n)$ times the density of the upper or lower bound of the interval of the continuous distribution.

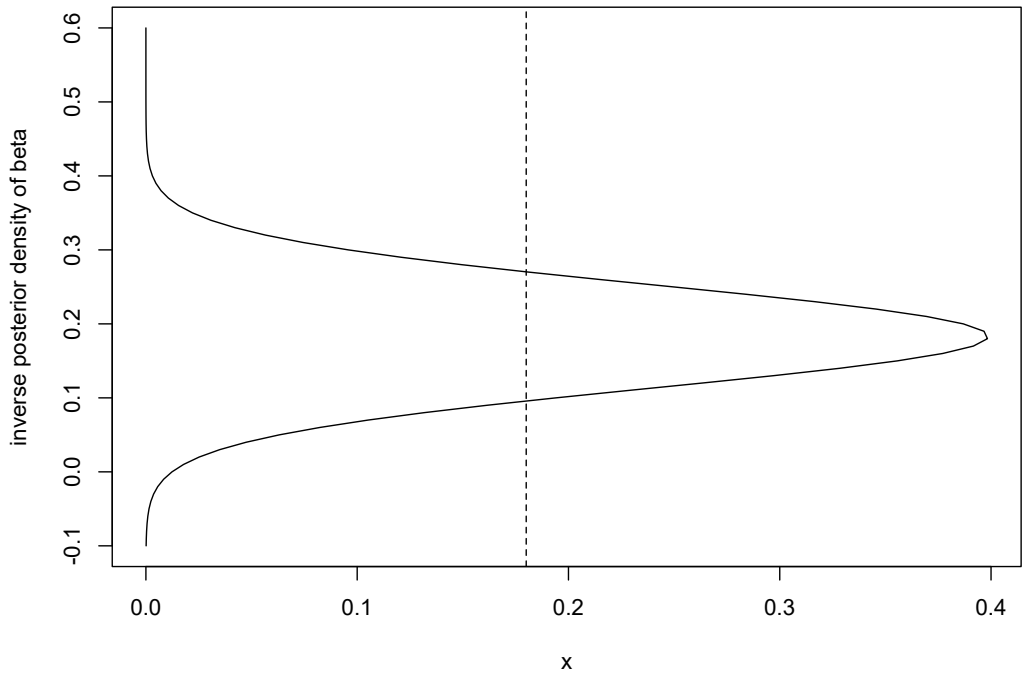


Figure 3.3.: An example of the general inverse of the posterior density of β_{21} given 200 data pairs simulated from model (M1) with $\beta = 0.2$, $\epsilon_1, \epsilon_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $a_{12} = 2$, $U_{11} = 1$, and $U_{12} = 0$.

To get a closed form expression for (3.49), we need to have at least a closed form expression for f^{-1}. To compute this, we first compute $f_{t_{2a}}^{-1}$. Recalling the density of a

Student's t-distribution we have

$$\begin{aligned} f_{t_{2a}}(y) = t \;\Rightarrow\quad & \frac{\Gamma\left(\frac{2a+1}{2}\right)}{\sqrt{2a\pi}\Gamma(a)}\left(1+\frac{y^2}{2a}\right)^{-\frac{2a+1}{2}} = t \\ \Rightarrow & \left(1+\frac{y^2}{2a}\right)^{-\frac{2a+1}{2}} = \frac{\sqrt{2a\pi}\Gamma(a)}{\Gamma\left(\frac{2a+1}{2}\right)}t \\ \Rightarrow & \quad 1+\frac{y^2}{2a} = \left(\frac{\sqrt{2a\pi}\Gamma(a)}{\Gamma\left(\frac{2a+1}{2}\right)}t\right)^{-\frac{2}{2a+1}} \\ \Rightarrow & \quad f^{-1}_{t_{2a}}(t) = \pm\sqrt{2a\left[\left(\frac{\sqrt{2a\pi}\Gamma(a)}{\Gamma\left(\frac{2a+1}{2}\right)}t\right)^{-\frac{2}{2a+1}} - 1\right]}. \end{aligned} \tag{3.50}$$

Next, we compute f^{-1}. By the density transformation formula, we know that

$$f(\beta_{21}) = f_{t_{2a}}\left(\sqrt{\frac{a\lambda}{b}}(\beta_{21}-\mu)\right)\cdot\sqrt{\frac{a\lambda}{b}}.$$

Plugging this into the inverse, we get

$$\begin{aligned} \beta_{21} = f^{-1}(t) \;\Rightarrow & f_{t_{2a}}\left(\sqrt{\frac{a\lambda}{b}}(\beta_{21}-\mu)\right) = \frac{b}{a\lambda}t \\ \Rightarrow & \quad \left(\sqrt{\frac{a\lambda}{b}}(\beta_{21}-\mu)\right) = f^{-1}_{t_{2a}}\left(\sqrt{\frac{b}{a\lambda}}t\right) \\ \Rightarrow & \quad f^{-1}(t) = \sqrt{\frac{b}{a\lambda}}f^{-1}_{t_{2a}}\left(\sqrt{\frac{b}{a\lambda}}t\right) + \mu. \end{aligned} \tag{3.51}$$

Taking (3.50) and (3.51) together, we have

$$\{f^{-1}(t)\}_{\min} = -\sqrt{\frac{2b}{\lambda}\left[\left(\frac{\sqrt{2a\pi}\Gamma(a)}{\Gamma\left(\frac{2a+1}{2}\right)}t\right)^{-\frac{2}{2a+1}} - 1\right]} + \mu \tag{3.52}$$

$$\{f^{-1}(t)\}_{\max} = \sqrt{\frac{2b}{\lambda}\left[\left(\frac{\sqrt{2a\pi}\Gamma(a)}{\Gamma\left(\frac{2a+1}{2}\right)}t\right)^{-\frac{2}{2a+1}} - 1\right]} + \mu \tag{3.53}$$

Hence, although we could not give an analytical solution for the constraints regarding r_{12}^{21}, the constraint (3.49) purely depends on r_{12}^{21}, and we could always tell for any fixed r_{12}^{21} whether we shall include 0 into the HDR. We would only include 0 in the HDR if r_{12}^{21} is a value that satisfies (3.49). If (3.49) is satisfied, the tails on both sides of the HDR are

constructed solely from the Student's t-distribution, which is a symmetric continuous distribution. Hence, we could still obtain equal tails with lower bound l defined by (3.42) and upper bound u defined by (3.43). Since here equal-tailed is not an obliged property, we directly set HDR $= [l, u] \cup \{0\}$.

Case 2. HDR includes only 0.

Under some extreme scenarios, $p(\mathcal{G}_{21}|X^n)$ is very high, i.e., the posterior distribution of the graphical structure tells that $\mathcal{G}_{21}$ is chosen with a very high probability. When $p(\mathcal{G}_{21}|X^n) \geq 1 - \alpha$, we directly set HDR $= \{0\}$ by definition.

Case 3. 0 is excluded from HDR.

If the constraints of Case 1 and Case 2 are both not satisfied, intuitively speaking if $p(\mathcal{G}_{21}|X^n)$ is very low and 0 is far away from μ, then 0 is excluded from the HDR. In this case, it is clear that the HDR is not equal-tailed with respect to to the mixture distribution anymore. However, the HDR is symmetric with respect to μ since two distinct values obtain the same densities of a Student's t-distribution if and only if they are symmetric with respect to μ. We can compute the lower and upper bounds l, u of the HDR by

$$\begin{cases} \frac{1}{1+r_{12}^{21}} \left(F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(u-\mu) \right) - F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(l-\mu) \right) \right) = 1 - \alpha \\ F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(u-\mu) \right) = 1 - F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(l-\mu) \right) \end{cases}. \tag{3.54}$$

Representing l with u using the second equation in (3.54) and plugging it into the first equation, we have

$$\begin{aligned} & 1 - 2F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(l-\mu) \right) = (1-\alpha)(1+r_{12}^{21}) \\ \Rightarrow \quad & 2F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(l-\mu) \right) = \frac{\alpha}{2}(1+r_{12}^{21}) - \frac{r_{12}^{21}}{2} \\ \Rightarrow \quad & l = \sqrt{\frac{b}{a\lambda}} q_{t_{2a}} \left[\frac{\alpha}{2}(1+r_{12}^{21}) - \frac{r_{12}^{21}}{2} \right] + \mu \end{aligned} \tag{3.55}$$

Analogously, representing u with l using the second equation in (3.54) and plugging it into the first equation, we have

$$\begin{aligned} & 2F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(u-\mu) \right) - 1 = (1-\alpha)(1+r_{12}^{21}) \\ \Rightarrow \quad & 2F_{t_{2a}} \left(\sqrt{\frac{a\lambda}{b}}(u-\mu) \right) = 1 - \left[\frac{\alpha}{2}(1+r_{12}^{21}) - \frac{r_{12}^{21}}{2} \right] \\ \Rightarrow \quad & u = \sqrt{\frac{b}{a\lambda}} q_{t_{2a}} \left(1 - \left[\frac{\alpha}{2}(1+r_{12}^{21}) - \frac{r_{12}^{21}}{2} \right] \right) + \mu \end{aligned} \tag{3.56}$$

Since we exclude 0 from HDR by assumption, we have HDR $= [l, u] \setminus \{0\}$ with l defined by (3.55) and u defined by (3.56). We could summarize the process of computing an HDR with Algorithm 2. We will later refer to this method as `HDI`.

Algorithm 2 Compute the CR as an HDR

Input: $\mathbf{a}, U, X^n, \alpha, p(\mathcal{G}|X^n)$

Output: a $(1-\alpha)$-HDR of $\mathcal{C}(1 \to 2)$ with respect to its posterior distribution given X^n

if (3.49) is satisfied and $p(\mathcal{G}_{21}|X^n) < 1 - \alpha$ **then**
    Set $C = [l, u] \cup \{0\}$ with l in (3.42) and u in (3.43);
else if $p(\mathcal{G}_{21}|X^n) \geq 1 - \alpha$ **then**
    Set $C = \{0\}$;
else
    Set $C = [l, u] \setminus \{0\}$ with l in (3.55) and u in (3.56);
end if
return the credible interval C as the HDR

3.3.3. Credible Region with a Threshold

In order to avoid the cumbersome process of deciding whether to include 0 in the credible region when computing an HDR, we innovatively develop an approach to decide whether 0 shall be included in the CI or not. We set in advance a threshold α_0 with $0 \leq \alpha_0 \leq \alpha$. If $p(\mathcal{G}_{21}|X^n) \geq \alpha_0$, then the posterior density of the graph has a high enough probability at $\mathcal{G}_{21}$, and hence we include 0 in the CR. Otherwise, 0 is excluded from the CR. We discuss this method with the following three cases.

Case 1. CR includes 0 and not only 0.

By assumption, when r_{12}^{21} satisfies the constraint $\alpha_0 \leq p(\mathcal{G}_{21}|X^n) = r_{12}^{21}/1+r_{12}^{21} < 1-\alpha$, we include 0 in the HDR. Analogous to computing the CR, we have CR $= [l, u] \cup \{0\}$ with l and u defined by (3.42) and (3.43).

Case 2. CR includes only 0.

Under some extreme scenarios, $p(\mathcal{G}_{21}|X^n)$ is very high, i.e., the posterior distribution of the graphical structure tells that $\mathcal{G}_{21}$ is chosen with a very high probability. When $p(\mathcal{G}_{21}|X^n) \geq 1 - \alpha$, we directly set HDR $= \{0\}$ by definition.

Case 3. 0 is excluded from CR.

By assumption, when r_{12}^{21} satisfies $p(\mathcal{G}_{21}|X^n) = r_{12}^{21}/(1 + r_{12}^{21}) < \alpha_0$, we exclude 0 from the CR. Analogous to the HDR, we have CR $= [l, u] \setminus \{0\}$ with l, u define in (3.55) and (3.56).

We summarize our method of computing a CR by defining a threshold α_0 using Algorithm 3. We need one additional input α_0 in comparison to Algorithm 1 and 2.

Since α_0 is a preset hyperparameter, we could obtain quite different CRs under different settings of α_0. We set special values for α_0 if we want to follow certain accepting

Algorithm 3 Compute the CR by defining a threshold for $p(\mathcal{G}_{21}|X^n)$

Input: $\mathbf{a}, U, X^n, \alpha, \alpha_0, p(\mathcal{G}|X^n)$

Output: a $(1-\alpha)$-CR of $\mathcal{C}(1 \to 2)$ with respect to its posterior distribution given X^n

if $\alpha_0 \leq p(\mathcal{G}_{21}|X^n) < 1-\alpha$ **then**
    Set $C = [l, u] \cup 0$ with l in (3.42) and u in (3.43);
else if $p(\mathcal{G}_{21}|X^n) \geq 1-\alpha$ **then**
    Set $C = \{0\}$;
else
    Set $C = [l, u] \backslash \{0\}$ with l in (3.55) and u in (3.56);
end if
return the credible interval C.

criteria. Setting $\alpha_0 = 0$ indicates that we would always include 0 in the CR as long as $p(\mathcal{G}_{21}|X^n) \neq 0$. On the other hand, setting $\alpha_0 = \alpha$ indicates that we only include 0 in the CR if it becomes too large for the tail, i.e., still excluding 0 from the CR will cause a contradiction to the assumption of the $(1-\alpha)$-credible region. In this case, we are very conservative about including 0 in the CR.

4. Algorithms and Experiments

In this chapter, we will first summarize the essential algorithms, then apply them to datasets for model testing and comparing. For the choice of datasets, we will use simulated data pairs as well as benchmarks of cause-effect pairs. We use the programming language R to do our simulations and evaluations.

4.1. Simulated Data

We first test our methods for deriving the credible intervals for causal effects $\mathcal{C}(1 \to 2)$ with simulated data. The process of generating n i.i.d. bivariate samples can be summarized in the following steps:

1. we generate n noise pairs $(\epsilon_{i,1}, \epsilon_{i,2})_{i=1,\dots,n} \overset{i.i.d.}{\sim} \mathcal{N}(0, D_2)$ where D_2 is a 2×2 diagonal matrix in which all diagonal entries are equal to σ^2;

2. we set the causal effect β (can either be β_{12} or β_{21} depending on the choice of the model);

3. we decide the LSEM we want to sample from; if we sample from model (M1), then we set
$$x_{i,1} = \epsilon_{i,1},\ x_{i,2} = \beta x_{i,1} + \epsilon_{i,2} \quad \text{for } i = 1, \dots, n;$$
on the other hand, if we sample from model (M2), then we set
$$x_{i,2} = \epsilon_{i,2},\ x_{i,1} = \beta x_{i,2} + \epsilon_{i,1} \quad \text{for } i = 1, \dots, n.$$

Using this procedure, we could generate n i.i.d. sample pairs with Gaussian noises of variances σ^2. We will set $\sigma^2 = 1$ for our simulation test so that all noises are standard normal. According to Peters and Bühlmann (2014), the main indication of one variable being the cause is that it has a relatively smaller variance, and the effect will have a larger variance in contrast. This is also obvious from the formulas of model (M1), since X_1 has variance 1, while X_2 has variance $\beta^2 + 1 > 1$.

The main goal of our paper is to generate credible intervals from the posterior distributions of the causal effect $\mathcal{C}(1 \to 2)$ given the data. This would be easy to compute if the posterior distribution is clear, which is a mixture of a point mass at 0 and a continuous distribution $p(\beta_{21}|X^n)$ with $p(\mathcal{G}_{12}|X^n)$ and $p(\mathcal{G}_{21}|X^n)$ being the weights. So the essential ideas and calculations of our model could be separated into two algorithms: the first algorithm computes $p(G|X^n)$, and the second algorithm computes $p(\mathcal{C}(1 \to 2)|X^n)$

Algorithm 4 Compute $p(\mathcal{G}|X^n)$

Input: $\mathbf{a}, U, X^n$

Output: an array $(p(\mathcal{G}_{12}|X^n), p(\mathcal{G}_{21}|X^n))$ representing $p(\mathcal{G}|X^n)$

1: Compute r_{12}^{21} using the log-likelihood and exponential function indicated by (3.28);

2: Set

$$p(\mathcal{G}_{21}|X^n) = \frac{1}{1+r_{12}^{21}},\ p(\mathcal{G}_{21}|X^n) = \frac{r_{12}^{21}}{1+r_{12}^{21}};$$

3: **return** a vector $p(\mathcal{G}|X^n) = (p(\mathcal{G}_{12}|X^n), p(\mathcal{G}_{21}|X^n))$.

based on $p(G|X^n)$ and methods for computing CRs mentioned in Section 3.3. We first use Algorithm 4 to compute the posterior distribution of the graphical structure.

If we have data following model (M1), we would expect Algorithm 4 to have an output of $p(\mathcal{G}_{12}|X^n) > p(\mathcal{G}_{21}|X^n)$. In the worst case when the algorithm cannot predict the graphical structure from the data, we would expect to have an output of $p(\mathcal{G}_{12}|X^n) = p(\mathcal{G}_{21}|X^n) = 0.5$. We will check this by doing 500 Monte Carlo replications and drawing histograms of posterior probabilities $p(\mathcal{G}_{12}|X^n)$ and $p(\mathcal{G}_{21}|X^n)$. Each time we generate 500 sample pairs following model (M1) with $\beta = 0.1$ and we set $\mathbf{a} = (2,2), U = I_2$ with I_2 being the two-dimensional identity matrix. We purposefully set β to be close to zero to make it challenging for the model to estimate the graphical structure. Under this setting, the variances of X_1 and X_2 will be theoretically very close to each other (only differ by 0.01) and the estimation might just be as good as random guesses. Hence we would expect $p(\mathcal{G}_{12}|X^n) \approx p(\mathcal{G}_{21}|X^n) \approx 0.5$.

However, the empirical results do not meet our expectations. We could see from Figure 4.1 that instead of $p(\mathcal{G}_{12}|X^n) \approx p(\mathcal{G}_{21}|X^n) \approx 1/2$, we have something more like

$$\mathbb{E}(p(\mathcal{G}_{12}|X^n)) \approx \mathbb{E}(p(\mathcal{G}_{21}|X^n)) \approx 1/2, \tag{4.1}$$

where with probability $\approx 45\%$ we obtain $p(\mathcal{G}_{12}|X^n) \approx 1$, with probability $\approx 40\%$ we obtain $p(\mathcal{G}_{21}|X^n) \approx 0$, and with probability $\approx 15\%$, $p(\mathcal{G}_{21}|X^n)$ has some values between 0 and 1. Most of the posterior probabilities cluster around 0 or 1, but not around 0.5 as we would be expecting. In other words, the algorithm is too sure about its prediction in each replication.

For example, among the 500 values $p(G_{12}|X)$ we have, more than 175 is < 0.05, that is more than 35%. This would be causing serious problems, since in these over 35% of the cases, $p(G_{21}|X) > 0.95$, so the CIs of $\mathcal{C}(1 \to 2)$ is nothing but $\{0\}$ for all methods except `ETI`. But since β_{21} is 0.1 and not 0, the CIs do not cover the true causal effect in these cases, leading our coverage rate to be always less than 65%. This coverage rate is even worse than when we decide the graphical structure by coin flipping (see Remark 3.16), which will theoretically still lead to a coverage rate of 90%.

Inspired by (4.1), we try to solve this problem by computing a bootstrap mean (see Section 2.6)z for $p(\mathcal{G}|X^n)$ instead of computing each time a single probability. In other words, we bootstrap from data X^n and compute a mean for $p(\mathcal{G}|X^n)$, then use the mean as an estimation for $p(\mathcal{G}|X^n)$ to compute the posterior mixture. Here we use bootstrap

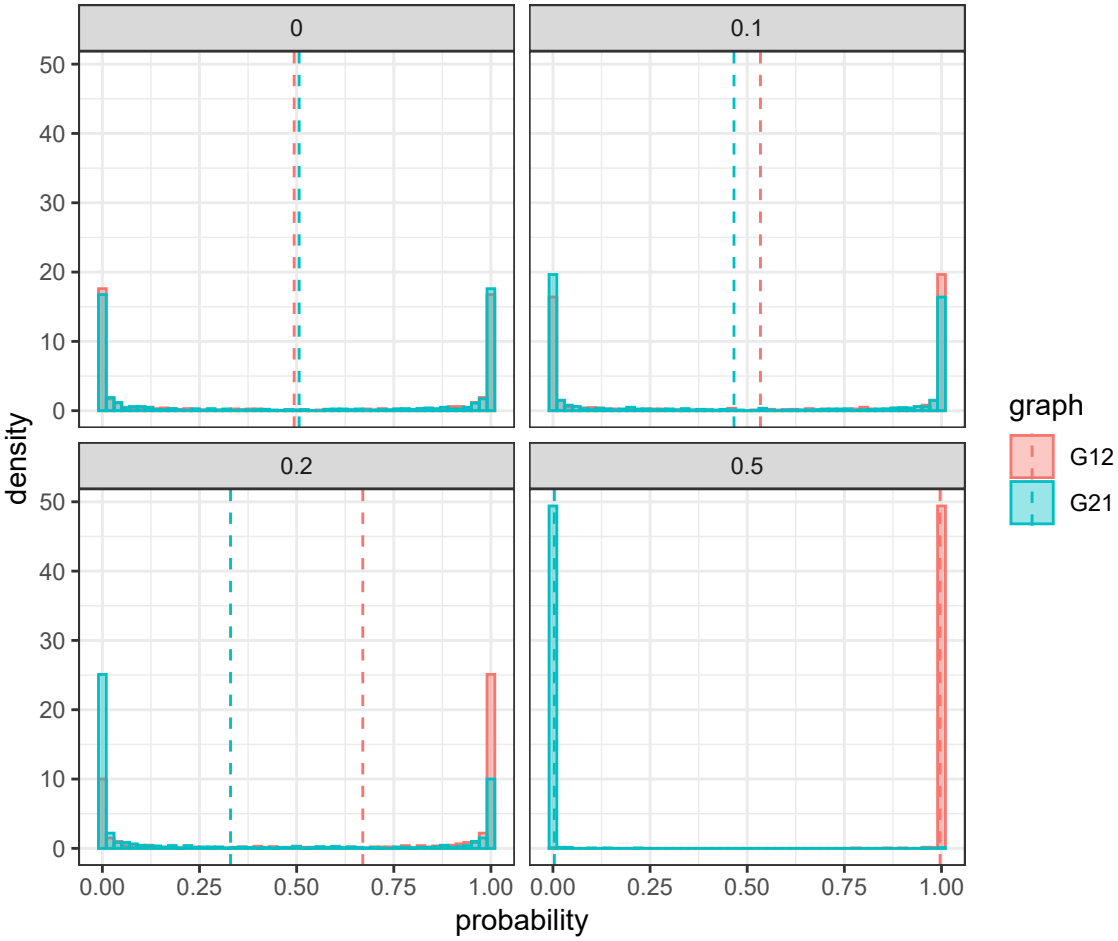


Figure 4.1.: The histogram of $p(\mathcal{G}_{12}|X^n)$ (red) and $p(\mathcal{G}_{21}|X^n)$ (blue) of 500 replications, where the bivariate pairs (x_1, x_2) are generated under model (M1) for $\beta_{21} = 0, 0.1, 0.2, 0.5$. The dashed lines represent the mean values of $p(\mathcal{G}|X^n)$ with $p(\mathcal{G}_{12}|X^n)$ being red and $p(\mathcal{G}_{21}|X^n)$ being blue.

resampling rather than direct sampling from the entailed distribution of model (M1) for later generalization to real-world datasets since their underlying distributions are usually unknown. Here we claim that such estimation for $p(\mathcal{G}|X^n)$ is still a valid probability density for a categorical distribution since the estimated $p(\mathcal{G}_{12}|X^n)$ and $p(\mathcal{G}_{21}|X^n)$ will still sum up to 1.

In Figure 4.2, we show the histograms of the estimated posterior probabilities $p(\mathcal{G}_{12}|X^n)$ and $p(\mathcal{G}_{21}|X^n)$ under the same setting as for Figure 4.1.

We could see a large improvement in the result. For instance, when β is small, the probabilities do not cluster around 0 and 1. Instead, they are kind of uniformly distributed between 0 and 1 and have slightly larger densities in the middle than on the sides. We have certainly improved the problem that the algorithm is too sure about deciding the graphical structure, while still preserving its ability to decide between $\mathcal{G}_{12}$

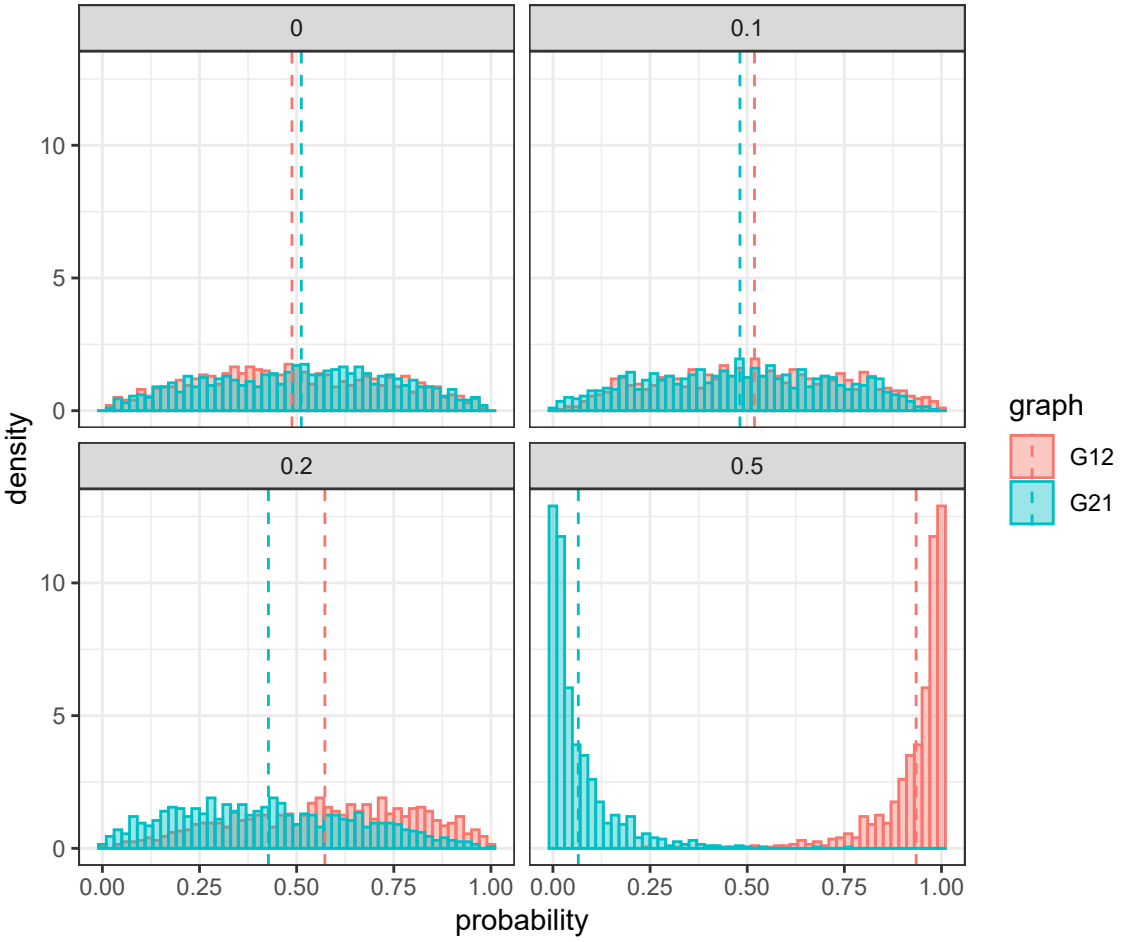


Figure 4.2.: The histogram of bootstrap estimated $p(G_{12}|X^n)$ (red) and $p(G_{21}|X^n)$ (blue) of 500 replications, where 500 bivariate data pairs (x_1, x_2) are generated under model (M1) for $\beta_{21} = 0, 0.1, 0.2, 0.5$. The dashed lines represent the mean posterior distributions with $p(\mathcal{G}_{12}|X^n)$ being red and $p(\mathcal{G}_{21})$ being blue.

and $\mathcal{G}_{21}$, especially for larger β. When $\beta = 0.2$, we can already see that the histogram of estimated $p(\mathcal{G}_{12}|X^n)$ is left skewed while $p(\mathcal{G}_{21}|X^n)$ is right skewed, indicating that the algorithm tends to estimate $p(\mathcal{G}_{12}|X^n) > p(\mathcal{G}_{21}|X^n)$ and in more cases. A similar conclusion can be drawn by observing the average lines, and the distance between the lines grows wider for larger β's. Even in the cases when the estimated $p(\mathcal{G}_{12}|X^n) > p(\mathcal{G}_{21}|X^n)$, i.e., $p(\mathcal{G}_{12}|X^n) > 0.5$, it would most likely not be an extreme value that is very close to 1. Hence when we further compute the CR, we would most likely get the union of an interval and the point 0 rather than having only $\{0\}$, which will remarkably improve the chance of covering the true non-zero causal effect. For $\beta = 0.5$, the bootstrap estimation of $p(\mathcal{G}|X^n)$ can successfully predict the true causal direction in almost all cases, as we can see that almost all red bins are located to the right of 0.5, while the blue bins are almost all on the left side. In comparison, the bootstrap

estimations reduce randomness and provide more trustworthy probabilities. As a trade-off, we need to resample $B(= 200)$ times for computing one estimation of $p(\mathcal{G}|X^n)$, which drastically increases the computational cost. Such a cost is affordable for us since we only deal with small datasets of causal pairs. However, this would become a serious problem if the dataset is large, and would limit the number of Monte Carlo replications we can make when testing the algorithm.

We now show with Algorithm 5 how we generate the posterior CR using three different methods. As discussed before, we will use a bootstrap average to estimate the true $p(\mathcal{G}|X^n)$, so we also need to define as inputs the bootstrap size B, which is the number of bootstrap replications we have, and the bootstrap sample size N, which is the number of samples to select in each bootstrap replication. In addition, we need to decide the desired type of CRs by setting `method` in the input.

Algorithm 5 Compute CR

Input: $\mathbf{a}, U, X^n, \alpha, \alpha_0, \texttt{method}, B, N$
Output: the credible region C

for $b = 1, \ldots, B$ **do**
    Select with replacement N samples from X^n and call them $X^N_{(b)}$;
    Compute $p(\mathcal{G}|X^N_{(b)})$ using Algorithm 4;
end for
Average the bootstrapping results and get an estimated distribution $\widehat{p(\mathcal{G}|X^n)}$ with

$$\widehat{p(\mathcal{G}_{12}|X^n)} = \frac{1}{B}\sum_{b=1}^{B} p(\mathcal{G}_{12}|X^N_{(b)}),\ \widehat{p(\mathcal{G}_{21}|X^n)} = \frac{1}{B}\sum_{b=1}^{B} p(\mathcal{G}_{21}|X^N_{(b)})$$

and use them as inputs for computing the CRs;
if `method`="ETI" **then**
    Compute the credible region C as an ETI using Algorithm 1;
else if `method`="HDI" **then**
    Compute the credible region C as an HDR using Algorithm 2;
else if `method`="threshold" **then**
    Compute the credible region C as an CR defined by a threshold α_0 using Algorithm 3;
end if
return the $(1-\alpha)$-credible region C

Algorithm 5 is the most essential algorithm in our thesis since it is a synthesis of the previous four algorithms and outputs a credible interval of our interest. We will evaluate the performance of Algorithm 5 by comparing the outputs of different types of credible regions with confidence intervals from other algorithms in various aspects. There is no standard criterion for evaluating credible regions, so we will compare the following values following Strieder et al. (2021) when evaluating our algorithms in 1000 Monte Carlo replications:

- coverage rate: the percentage of replications where the true causal effects lie in the CRs;
- width: the average width of the CRs, where we only consider the length of the interval part, i.e., if the CR is the union of $\{0\}$ and an interval, we do not care about the distance from 0 to the interval, but only the width of the interval itself;
- zero percentage: the percentage of replications where 0 is contained in the CRs.

In addition to the above three criteria that directly show the model performance, we will also compare the running time of different algorithms since this is also an important criterion for evaluating an algorithm.

For comparison, we will compute the credible intervals of causal effects from X_1 on X_2 using different methods: `ETI`, `HDI`, and CR determined by a threshold, where the threshold α_0 will be set to 0, 0.01, and 0.05. We denote the methods of producing CRs with these thresholds as `thre0`, `thre01` and `thre05` respectively. It is worth noticing that 0 is the minimum and 0.05 is the maximum value a threshold can take given that $\alpha = 0.05$. Although several previous research studied the causal structure of bivariate cases with Bayesian approaches, they mostly focused on the graphical structure rather than the causal effects, not to say the credible regions of causal effects. This makes it hard to find a proper method to compare with. So instead of comparing with other credible regions, we compare our results with confidence regions of likelihood ratio tests (method `LRT`) and split likelihood ratio tests (method `SLRT`) from Strieder et al. (2021).

As stated in Section 3.3, credible regions can be seen as the Bayesian version of confidence regions. Let us briefly recap what a confidence region (or more commonly confidence interval) is (see Rice (2006)). A $(1-\alpha)$-confidence region for a parameter θ of some probability distribution under a statistical test $\mathcal{H}$ is a random region (interval) that contains θ with probability $(1-\alpha)$ given that the null hypothesis $\mathcal{H}_0$ is true. On the other hand, a $(1-\alpha)$-credible region, constructed based on the posterior distribution of the parameter θ, is a set of more plausible values that we believe would contain θ with probability $(1-\alpha)$. Although both regions (intervals) are constructed from different underlying probability distributions, they are regions that we believe θ would most likely lie in. We also assume that the credible and confidence regions cover the true value with probabilities of at least $(1-\alpha)$ under the posterior distribution and null hypothesis respectively. In short, credible intervals and confidence intervals are very similar in the sense of likelihood and coverage, so it makes sense for us to compare them. Without distinguishing between terminologies, we will also write CR for both credible regions and confidence regions, as well as CI for both credible intervals and confidence intervals.

We first compare the coverage rates of CRs of $\mathcal{C}(1 \to 2)$ constructed via seven different methods in 1000 replications. We also want to see the performance of each method under different settings, so we will vary the sample size and the true causal effect. In addition, data are constructed both from model (M1) and model (M2). For the latter case, the true causal effect is 0 under our assumption (2.6).

The results are reported in Table 4.1. We have a total of two columns where the left column reports the coverage rate from data generated under model (M1) and the right

under model (M2). To evaluate how the size of the true causal effect would affect the accuracy of our prediction, we set β to a list of values from 0 to 0.5. We also set the sample sizes to 100, 500, and 1000.

	model	$1 \to 2$					$2 \to 1$				
method	$n \backslash \beta$	0.00	0.05	0.10	0.20	0.50	0.00	0.05	0.10	0.20	0.50
`ETI`	100	1.00	0.94	0.93	0.94	0.95	1.00	1.00	1.00	1.00	1.00
	500	1.00	0.92	0.93	0.93	0.95	1.00	1.00	1.00	1.00	1.00
	1000	1.00	0.94	0.93	0.95	0.94	1.00	1.00	1.00	1.00	1.00
`HDI`	100	0.99	0.93	0.92	0.94	0.95	0.99	0.99	0.99	1.00	1.00
	500	0.99	0.92	0.92	0.93	0.95	0.99	0.99	1.00	1.00	1.00
	1000	0.99	0.93	0.92	0.94	0.95	0.99	0.99	1.00	1.00	1.00
`thre0`	100	1.00	0.93	0.92	0.94	0.95	1.00	1.00	1.00	1.00	1.00
	500	1.00	0.92	0.92	0.93	0.97	1.00	1.00	1.00	1.00	1.00
	1000	1.00	0.92	0.93	0.94	0.97	1.00	1.00	1.00	1.00	1.00
`thre01`	100	1.00	0.94	0.93	0.94	0.95	1.00	1.00	1.00	1.00	1.00
	500	1.00	0.92	0.92	0.93	0.95	1.00	1.00	1.00	1.00	1.00
	1000	1.00	0.92	0.93	0.94	0.95	1.00	1.00	1.00	1.00	1.00
`thre05`	100	0.99	0.93	0.92	0.94	0.95	0.99	0.99	0.99	1.00	1.00
	500	0.99	0.92	0.92	0.93	0.95	0.98	0.99	1.00	1.00	1.00
	1000	0.99	0.92	0.93	0.94	0.95	1.00	0.99	0.99	1.00	1.00
`LRT`	100	0.95	0.95	0.95	0.95	0.94	0.94	0.95	0.94	0.95	0.94
	500	0.95	0.96	0.94	0.95	0.96	0.94	0.95	0.96	0.95	0.95
	1000	0.96	0.95	0.96	0.96	0.95	0.96	0.96	0.95	0.96	0.95
`SLRT`	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4.1.: The coverage rates of credible intervals and confidence intervals of $\mathcal{C}(1 \to 2)$ estimated with 1000 replications.

Many interesting observations can be made from Table 4.1. We can see that the CRs constructed with `SLRT` achieved the highest coverage rate among all the methods. Method `LRT` achieved the desired coverage rate of approximately 0.95. Besides, all five proposed methods achieved coverage rates of at least 0.90, which is at least better than the worst case explained by Remark 3.16. Also as explained in Remark 3.15, we achieve better coverage rates when $\beta = 0$ than when β is close but not equal to 0. In fact, all methods except for `LRT` achieve an almost 100% coverage rate when $\beta = 0$. In other words, we can almost always predict it when the two variables are completely not related. Apart from that, the estimated CRs have higher coverage rates for data generated from model (M2). For data generated from model (M1), our five methods achieve higher coverage rates for larger β, while for `LRT` and `SLRT` the coverage rates remain unchanged no matter how large β is. For all seven approaches, there is no clear evidence that sample size can influence the coverage rate. Overall, we cannot tell which of our five methods

is the best concerning the coverage rate.

Next, we compare the widths of CRs when we have different sample sizes n and different true causal effects β both under model (M1) and model (M2). Since we might have regions instead of intervals, the way of computing the widths remains open for discussion. There are two types of such regions we might get, either the union of an interval and a point $\{0\}$ or the difference of an interval and a point $\{0\}$. In both cases, the added part or the deleted part is just a single point, which is a null set under the Lebesgue measure. Here we compute the width of a region similar to its Lebesgue measure in the sense that the width of $[l, u] \cup \{0\}$, $[l, u]$ and $[l, u] \backslash \{0\}$ are all $u - l$ for $l \leq u$.

The importance of exploring the widths of CRs is clear. When having the same coverage rates, a CR with smaller width is considered to be more precise. In the most extreme case, a CR with an infinitely large width will surely cover the desired value with 100% probability. However, such a CR does not make much sense since it contributes almost nothing to our estimation. The trend of how the widths change against the sample size is compared in Figure 4.3 with different fixed β's.

Figure 4.3 provides interesting results in multiple aspects. When β is small, the widths of CRs for data from model (M1) and model (M2) do not have large differences if other factors are under the same settings. On the other hand, when β is large simulated samples from model (M2) have in general smaller widths meaning that we can give more precise estimates of the causal effect being 0 satisfying the desired coverage rate (e.g. 95%). When β is small, the widths of CRs of all methods decrease with the sample size. When β is large, the trend remains the same for CRs produced by all other methods except `ETI` for data generated under model (M2). As discussed before, the construction of a CR is based on two parts: the point mass at 0 indicating the probability of predicting the graph wrongly, and the continuous distribution of β in the correct direction. For model (M2) with high β the point mass at 0 is very likely to be over 0.5 as can be seen from Figure 4.1, which ensures that 0 is included in the CR generated by method `ETI`. On the other hand, since 0 can never be the mean of the mixture distribution for large β, CRs produced by `ETI` always have to include something else other than 0 to make sure the equal-tailed assumption holds. Hence the width of an ETI will always end up being 2μ with μ being the mean of the Student's t-distribution defined in (3.28).

We will neglect `ETI` from our discussion in this paragraph because of its abnormal behavior. Among all other methods that at least have similar trends, `SLRT` generates CRs with larger widths compared to other methods. In most cases, `LRT` also produces CRs with larger widths than our four methods. The difference is more obvious for smaller β, and not that obvious for large β and large sample size. For model (M1), the widths of CRs generated by all methods except `SLRT` converge to around 0.17. To what value will the widths of `SLRT` converge remains unclear, yet for sample size 1000 the width is around 0.25. For model (M2) on the other hand, the widths of CRs of our four methods converge quickly to 0 as the sample size increases. `LRT` needs to have higher sample sizes to be certain about the estimation and to obtain CRs of widths close to 0, while the performance of `SLRT` remains unclear even with a very large sample size, yet

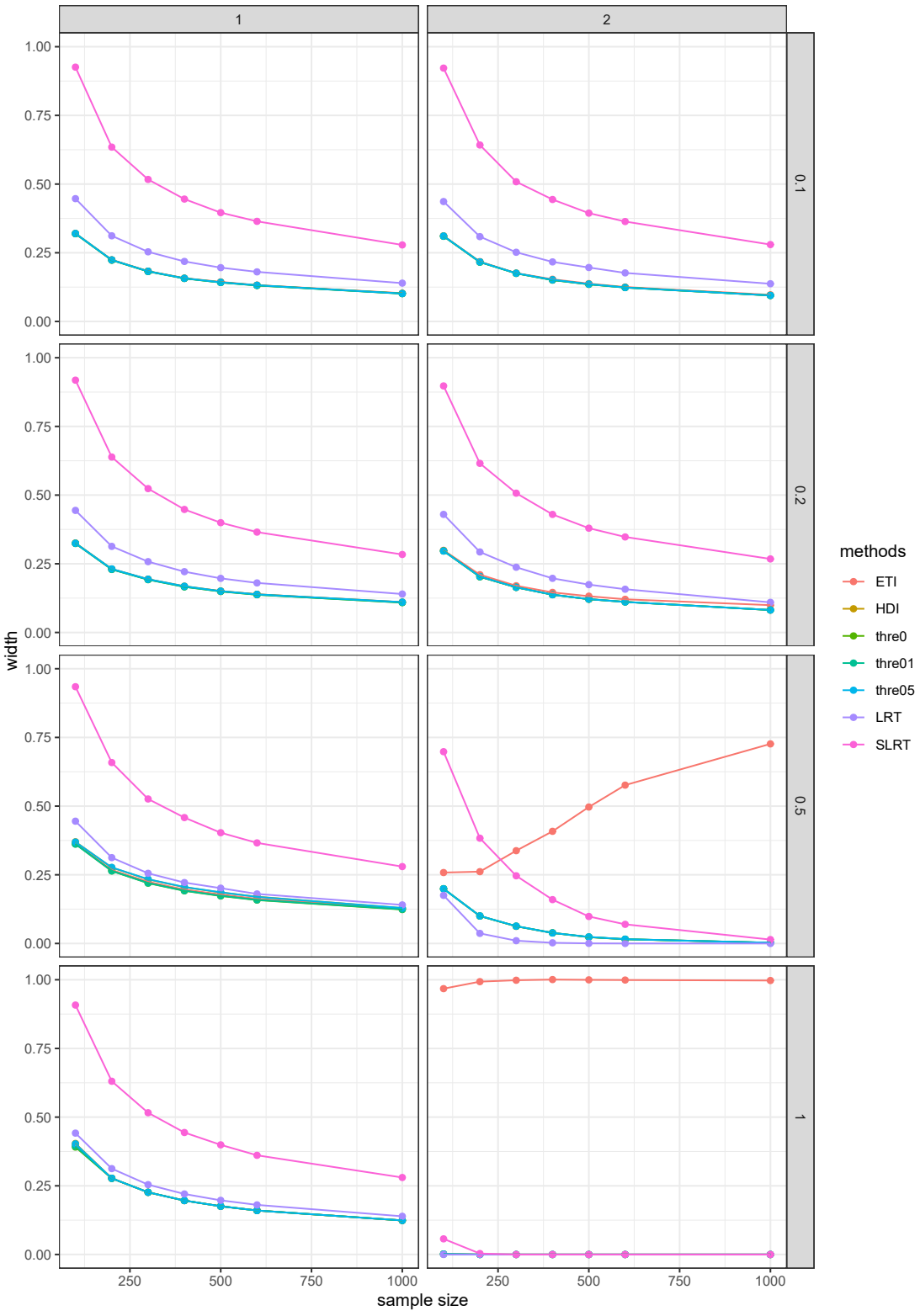


Figure 4.3.: Curves showing the widths of CRs against the sample size for different fixed causal effects estimated with 1000 replications.

we know that it always tends to give wider regions.

From Figure 4.4 we can see the tendency of the widths against β for different fixed sample sizes. We could already have an insight into how it looks like from the series of plots in Figure 4.3, however plotting what we want to observe in the same plot illustrates the changes more intuitively. We set β in both model (M1) and model (M2) to be some values between -0.2 and 1, while fixing the sample sizes to be 100, 200, 500 and 1000.

A general view of Figure 4.4 would already give us some interesting observations. For data generated under model (M1), the widths of CRs are almost constant for all β's and all methods. We could observe a slight increase in the widths of CRs from our five methods when β gets larger. On the other hand, the width decreases for larger β for data generated under model (M2) when we use all other methods except `ETI`. Not observing `ETI`, we see that the width will eventually decrease to 0 for model (M2) despite the sample size. The widths under this setting will be smaller for larger sample sizes and will decrease to 0 when β is large enough. For example, when $\beta = 1$ almost all methods except `ETI` produce CRs with widths very close to 0. The abnormal behavior of `ETI`, as already observed in Figure 4.3, becomes now clearer. The widths of CRs produced by `ETI` seem to start increasing when β exceeds 0.5. However, the width seems to be larger when we have more simulated samples. In general, the trend of how the width changes with β stays the same for different sample sizes. Also for the first time, we observe data with negative causal relations. It is interesting to see that all of the methods provide symmetric behaviors for positive and negative causal effects, showing that all methods can generalize nicely no matter whether the causal effects are positive or negative.

Another measure that we are interested in evaluating is the percentage of zero being included in the CRs. Having a completely non-zero CR can be seen as strong evidence of the existence of a non-zero causal effect. On the other hand, having 0 included in the CR makes our statement less persuasive. In Figure 4.5, we look at how the percentage of zero will change with different sample sizes from 100 to 1000. We also compared the curves under different settings of true causal effects and causal directions.

In general, for data generated under model (M1), the CRs produced by `LRT` have the lowest zero percentages, and they converge also fastest to 0 as the sample size increases, followed by `thre05`, `HDI`, `SLRT`, and `ETI`. The method `thre01` has slightly larger zero percentages than the above-mentioned methods and the CRs of `thre0` always contain 0 with 100% possibility. This could be explained by the assumption we made for method `thre0` that we will include 0 in the CR even if $p(\mathcal{G}_{21}|X^n)$ is just slightly larger than 0. In comparison, `LRT` is much more conservative about including 0. Methods `thre05`, `HDI`, and `SLRT` have very similar zero percentages. For very small $\beta(= 0.1)$, 0 will always be included in CRs constructed by all methods, even with large sample sizes. In other words, no matter how large the sample size is, the algorithm will not be confident enough to state that there is a non-zero causal effect, since the true causal effect is small. For slightly larger $\beta(= 0.2)$, the zero percentages of `LRT` first start to show an obvious decrease with the sample size. For even larger $\beta(= 0.5)$, `LRT` can give a CR without 0 already when the sample size is 500. The zero percentages of all other methods except `ETI` also start to converge against 0 as the sample size increases. Only the zero

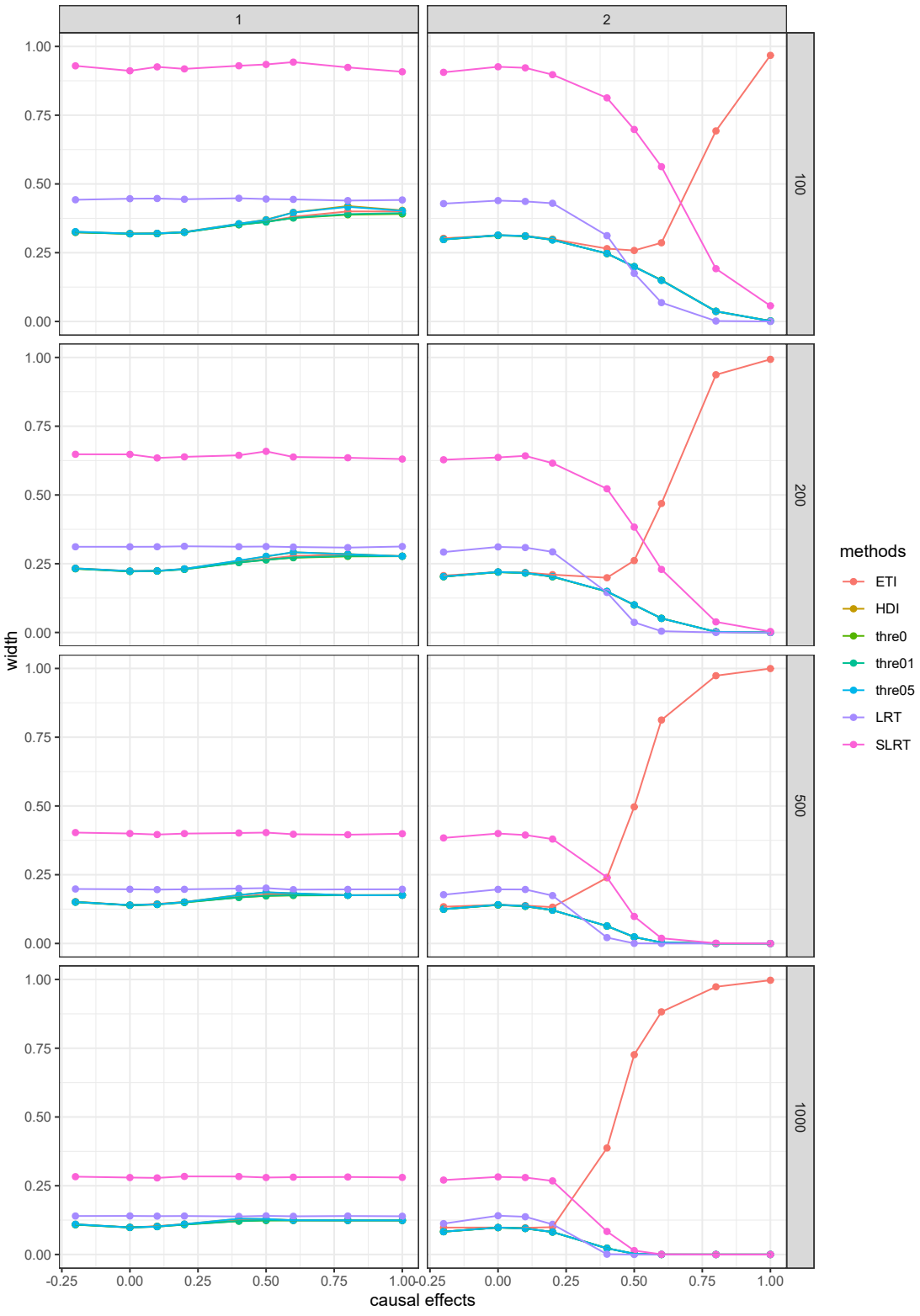


Figure 4.4.: Curves showing the widths of CRs against the causal effects for different fixed sample sizes estimated with 1000 replications.

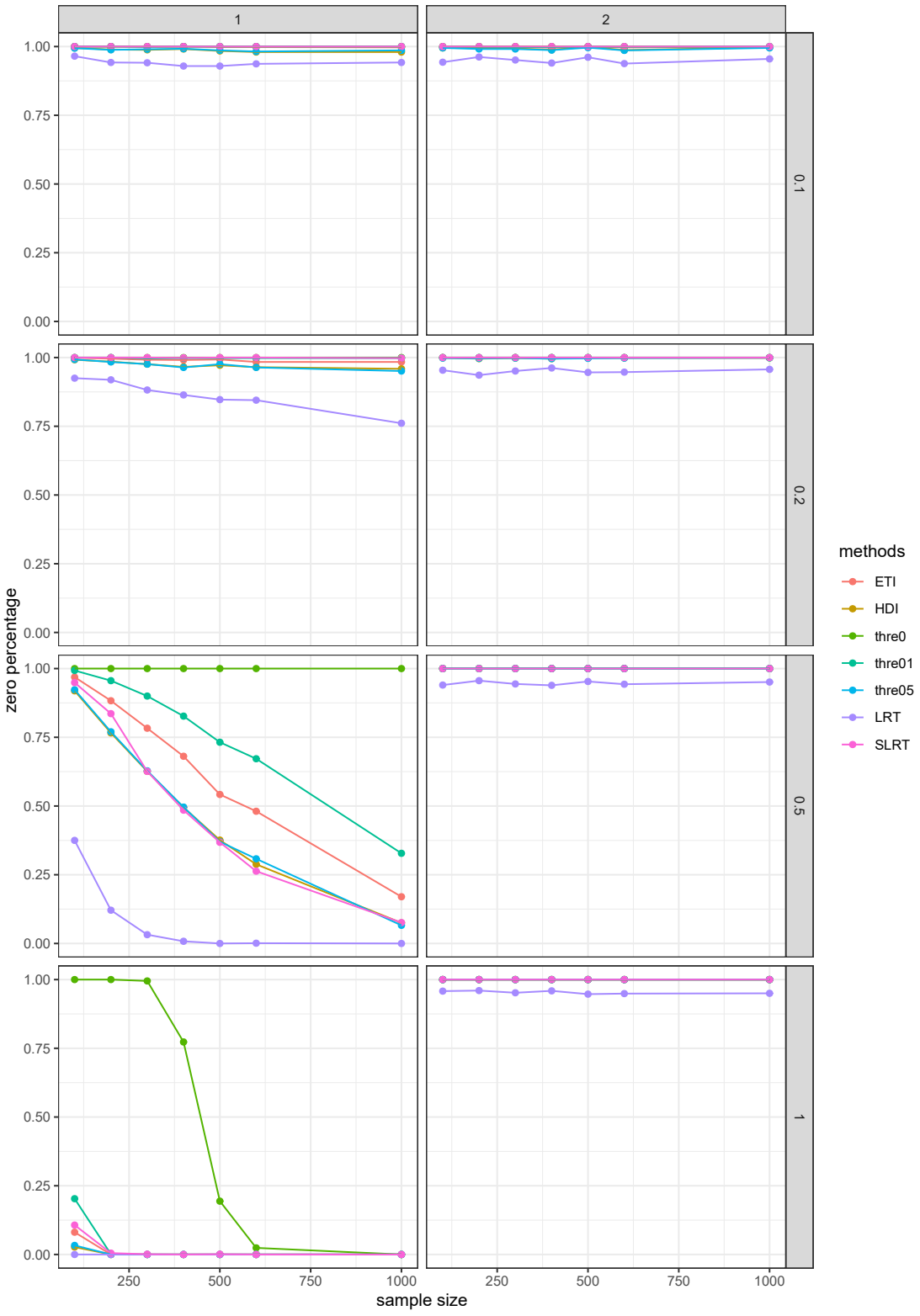


Figure 4.5.: Curves showing the zero percentages of CRs against the sample size for fixed causal effects estimated with 1000 replications.

percentages of ETI remain to be 100% with similar reasons as when we explained why its widths are so large. However, when β is large enough ($= 1$), the zero percentage of ETI will finally drop to 0 as the sample size increases. For data generated under model (M2), the CRs of LRT have zero percentages of about 95% while CRs of all other methods have 100% zero percentages, no matter how large the sample sizes are. On the one hand, we can say that LRT is very conservative about including 0 in comparison to other methods, no matter the true causal direction. On the other hand, having a CR without zero can almost indicate that there is a non-zero causal effect while the converse is not true.

In Figure 4.6 we show more directly how the zero percentage will change with β from -0.2 to 1. The plots illustrate the curves for sample sizes being fixed to 100, 500, and 1000. We still consider data generated under both model (M1) and model (M2).

There is not much to say about model (M2) since it is already clearly illustrated in Figure 4.5. However, we could still make some interesting observations for data generated under model (M1). No matter how large the sample size is, the zero percentage will decrease drastically from 1 for small β to 0 for large β. The only difference might be that data of large sample sizes are more confident with detecting relatively small β than data with smaller sample sizes. We also see a symmetry of negative and positive causal effects from the plots, since the zero percentage for data with negative β are approximately the same as for positive β with the same absolute values.

We summarize the observations in the table and figures presented above and highlight the most interesting findings in short. Among all the seven methods, SLRT achieves the best coverage rate, yet also produces CRs with significantly larger widths. LRT also tends to produce CRs with slightly larger widths than our five methods. It is also the most conservative about including zero. Considering our methods, ETI performs nicely when we can assume the causal direction but will create CIs with very large widths in the opposite direction. However, we usually do not know the causal direction in such a task and our goal is to predict the causal direction. At the same time, ETI does not outperform other methods in terms of coverage rate, so it does not seem to be a nice summary of the mixture distribution. Constructing the CR by method thre0 will strongly encourage the CR to include 0, so having a CR by method thre0 not including 0 is a very strong signal of having a remarkable causal effect, yet it is a bit too conservative in stating the existence of a causal effect when it is small. It is surprising to see that the two methods: HDI and thre05 have very similar characteristics in almost all three values we measured. Since thre05 has the most strict condition of including 0, we could expect it to be almost as strict for HDI. On the other hand, method thre01 also seems to be similar to HDI and thre05, only thre01 seems to have more preference of including 0 than the other two methods, especially for small β.

Finally, we compare the computational cost of each method. For evaluating the running time we use the function system.time() in R. The function takes an expression and evaluates the time in three aspects: user time, system time, and elapsed time (see Becker (2018)). The "user time" is the CPU time charged for the execution of the given expression (i.e., the current R session). The "system time" is the CPU time spent by

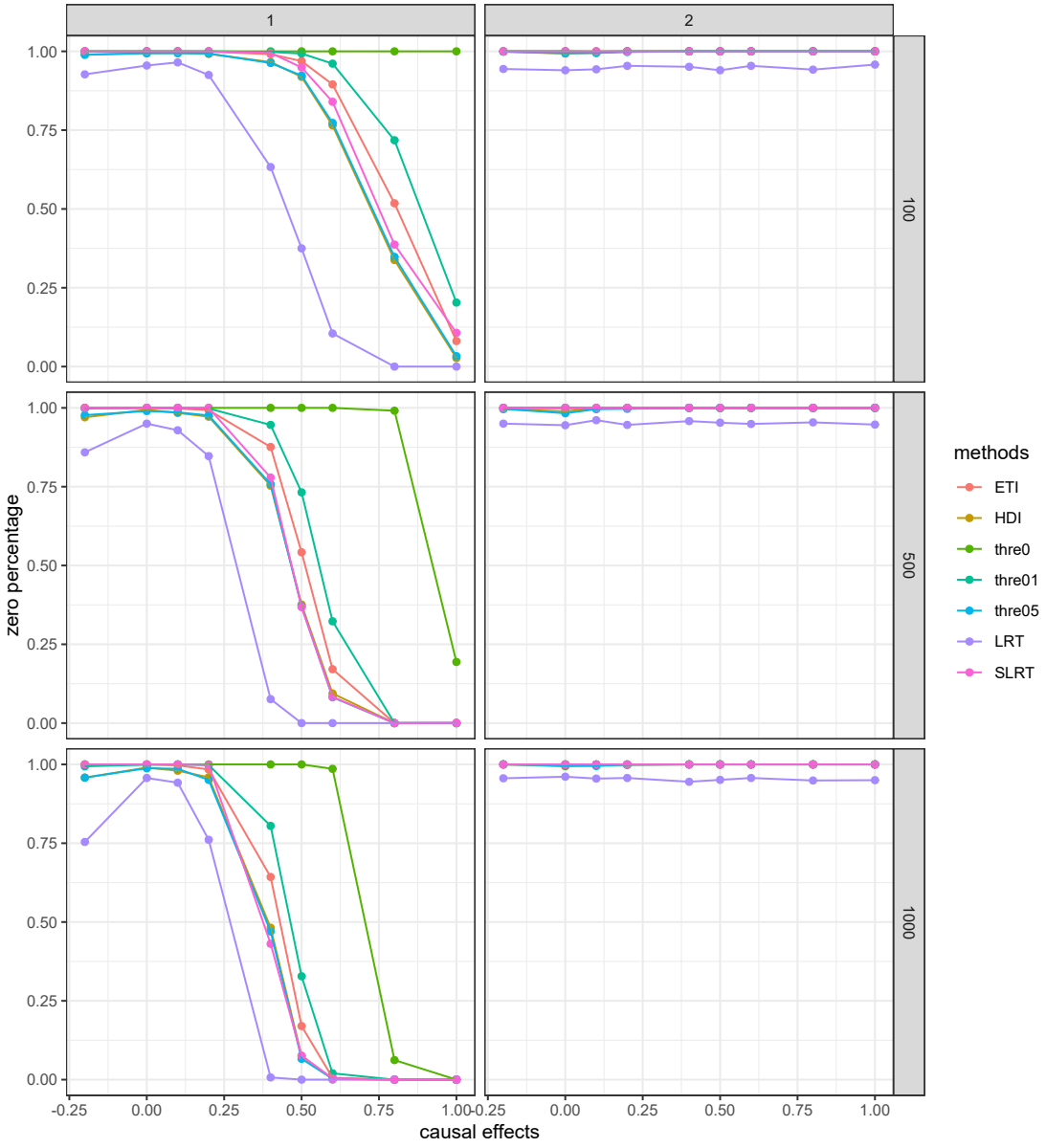


Figure 4.6.: Curves showing the zero percentages of CRs against the causal effect for different fixed sample sizes estimated with 1000 replications.

the kernel (the operating system) on behalf of the current process. The "elapsed time" is the real time that has passed between invocation and termination. Trying to avoid the influence of the system and focus on the running times of the algorithms alone, we compare the "running time" of all methods. Figure 4.7 uses box plots to illustrate the running times (evaluated with different sample sizes) of all seven methods in 1000 replications.

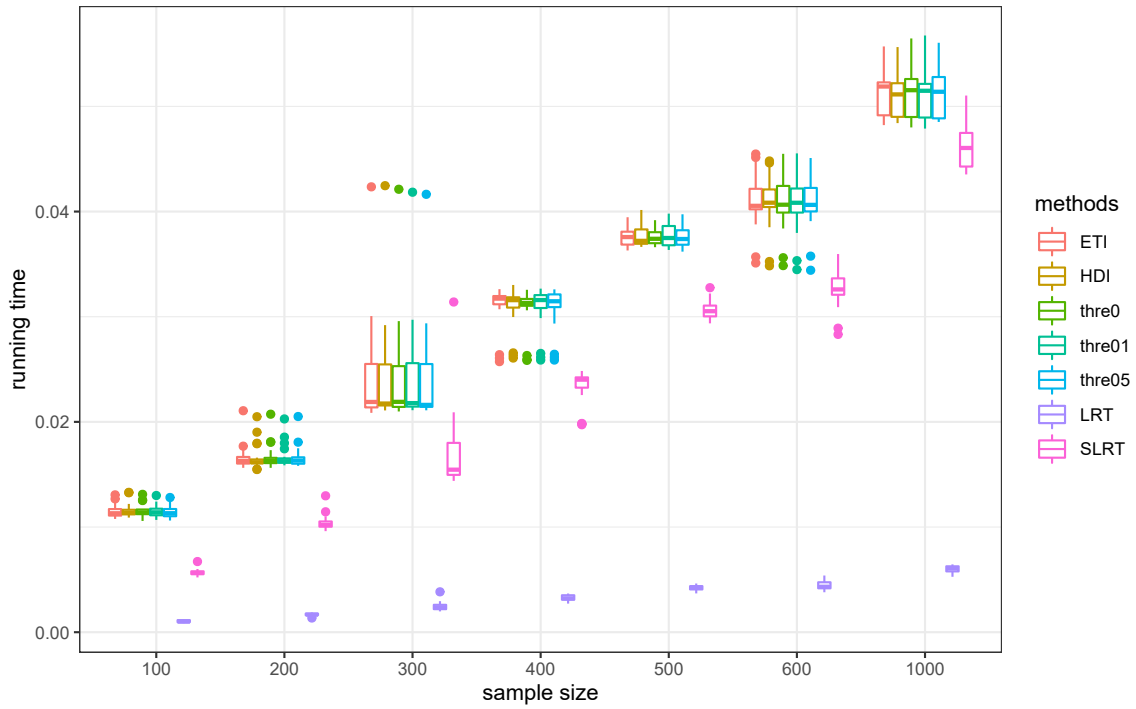


Figure 4.7.: User times of all the seven methods evaluated by 1000 replications with different sample sizes.

The running time of all methods seems to have a positive (sub-)linear relation with the sample size, which indicates that the computational cost of none of the methods will grow exponentially with the sample size, and we could expect an upper bound for the running time we need. Looking at the differences between different methods, we see that `LRT` has an obvious advantage over other methods. For example, for data with 1000 samples, the user time of `LRT` is even less than 0.01 seconds while the user time of all other methods is over 0.04. `SLRT` is the second fastest algorithm, yet does not show much difference from our five methods. All other methods that use the Bayesian approach have very similar running times no matter the sample size. Surprisingly, all the threshold methods are not faster than others, even though they have the simplest decision criteria of including 0. A reasonable explanation for this might be that instead of deciding on whether to include 0, estimating the posterior distribution of $p(\mathcal{G}|X^n)$ with bootstrapping is the most costly performance.

4.2. Benchmarks

For benchmark datasets, we use the collection presented in Mooij et al. (2016). This is a standard choice for testing causal models in bivariate cases since the data is nicely described and the causal directions are clearly provided. It contains 108 datasets from various fields, 102 of which are bivariate datasets. According to our assumptions of having an LSEM with noises of equal variances, we only choose pairs that are at least normally distributed and linearly related. Following the choice of Strieder et al. (2021) for example, we select the following pairs: `pair66`, `pair67`, `pair76`, `pair89`, `pair90`. We will give a brief description of the datasets (see details in Mooij et al. (2016)):

- `pair66` and `pair67` denote daily stock returns from several financial institutions;
- `pair76` denotes the average annual rate of change of total food consumption and the total population;
- `pair89` and `pair90` denote the fine root decomposition rates of different months in German forests and grasslands.

We summarize the true causal directions, the sample sizes, and the sample variances of the five causal-effect pairs in Table 4.2. We specifically list the variances of all variables for two reasons:

1. Due to our assumption of homoscedasticity, we want the added noises of both variables to have equal variances. Although we could not test the variances of the noises directly, we require X_1 and X_2 in each paired data to have variances at least of the same scale, leading to possibly similar noise variances.
2. We will be needing the variances as references for setting hyperparameters U according to our discussion in Section 3.1.3.

Name	True Direction	Sample Size	$\mathrm{Var}(X_1)$	$\mathrm{Var}(X_2)$
`pair66`	$1 \to 2$	1331	3.98×10^{-4}	4.02×10^{-4}
`pair67`	$1 \to 2$	1331	4.02×10^{-4}	5.33×10^{-4}
`pair76`	$1 \to 2$	347	4.18×10^{0}	7.33×10^{0}
`pair89`	$1 \leftarrow 2$	131	1.92×10^{1}	7.01×10^{0}
`pair90`	$1 \leftarrow 2$	126	3.89×10^{1}	1.72×10^{1}

Table 4.2.: The true causal directions, the sample sizes, and the sample variances of both variables of the five selected cause-effect pairs.

We now verify that the datasets are properly chosen satisfying our assumptions. The scatter plots in Figure 4.8 show that the two variables from each dataset can be modeled by a linear relationship.

The first three plots show clear linear relationships between X_1 and X_2. The points in the last two plots look a bit randomly located, yet still lie around some straight lines. It

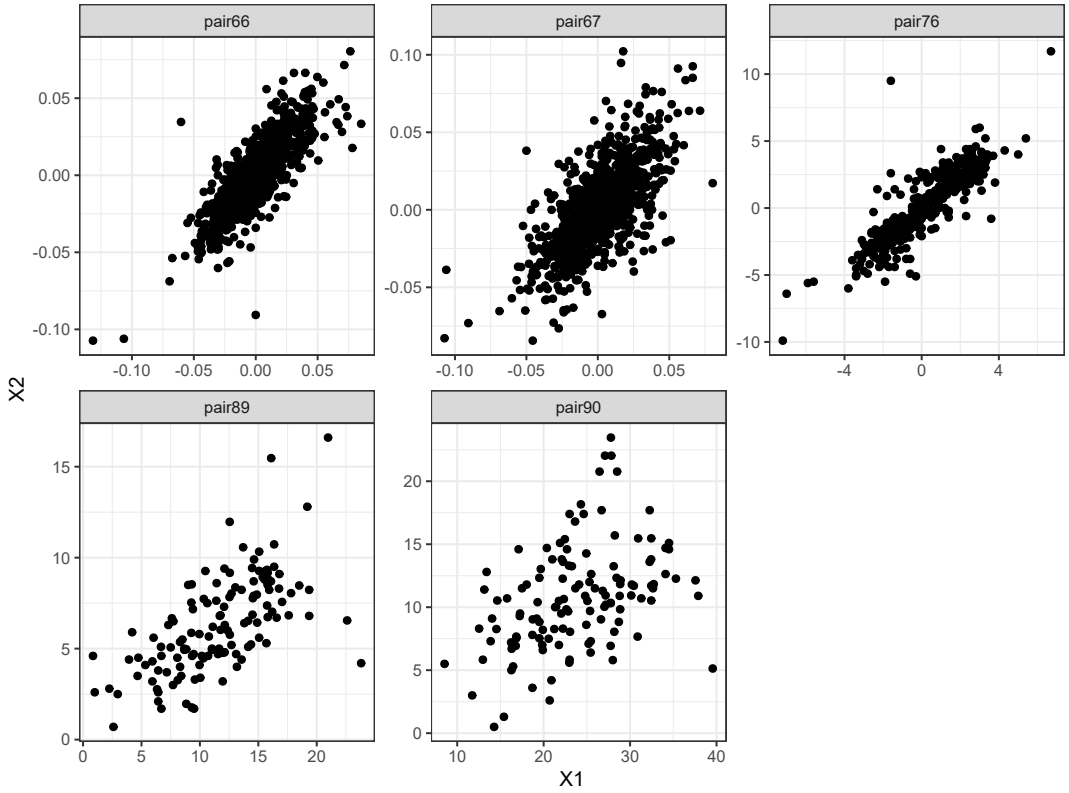


Figure 4.8.: Scatter plots of the five selected cause-effect pairs showing roughly linear relations.

is also interesting to see that all pairs seem to show a positive correlation. Although X_1 and X_2 in each collection of data pairs land within the same scale, there are relatively large differences between datasets. For example, data in `pair66` range mostly between -0.1 and 0.1, yet data in `pair89` range from 0 to 50. Also, our model assumes the data to be centered, yet most of the real-world data are not. Hence, we need to pre-process all data properly by centering them before testing our model with them.

We now use QQ-plots (as shown in Figure 4.9) to illustrate that the empirical distributions of both variables from each dataset are roughly Gaussian distributed.

From the plots, we see that `pair66` and `pair67` are somewhat heavy-tailed while `pair76` is a bit light-tailed. However, they stated around the line in general. We also see from the last two plots that `pair89` and `pair90` are almost perfectly normally distributed. In short, the five selected collections of cause-effect pairs roughly satisfy our assumptions of Gaussianity and linearity and can be considered as chosen properly.

The first step after we obtain some real-world datasets is preprocessing so that they fulfill our model assumptions and basic beliefs. We have already checked with the above two plots that linearity and Gaussianity hold for all datasets and we still need to ensure that the data are centered according to Assumption 2.27, which is not the case as can be seen from Figure 4.8. Hence, we center X_1 and X_2 in each dataset by subtracting the

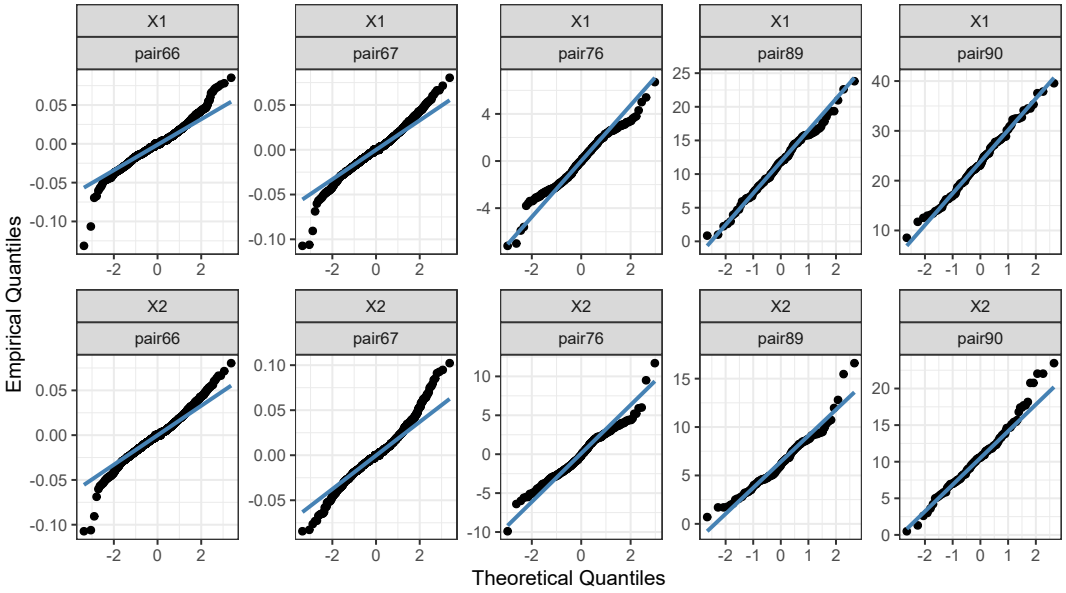


Figure 4.9.: QQ-plots for all variables of the five selected cause-effect pairs for checking whether their empirical distributions are Gaussian.

sample mean. More formally speaking, the transformed data $\tilde{X}^n = \{(\tilde{x}_{i,1}, \tilde{x}_{i,2})\}_{i=1,\dots,n}$ is obtained from the original data $X^n = \{(x_{i,1}, x_{i,2})\}_{i=1,\dots,n}$ by

$$\tilde{x}_{i,1} = x_{i,1} - \overline{x}_1, \quad \tilde{x}_{i,2} = x_{i,2} - \overline{x}_2, \qquad \text{for } i = 1, \dots, n,$$

where

$$\overline{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{i,k}, \qquad \text{for } k = 1, 2.$$

The second step would be to set proper hyperparameters a_{12}, U_{11}, U_{12} as we have discussed in Section 3.1.3. From Table 4.2 we see that the variances of both variables in `pair66` and `pair67` are so small that if we still set $U_{11} = U_{22} = 1$ and $U_{12} = U_{21} = 0$ as in the simulation studies, the hyperparameters will dominate the prediction and lead to undesired results. Hence it is important for us to set proper hyperparameters. As discussed before, we want U_{11}, U_{22} to be of the same range as $\mathrm{Var}(X_1), \mathrm{Var}(X_2)$. Hence, we set $U_{11} = U_{22} = \mathrm{Var}(X_1)/2$ when we apply the algorithm to real-world datasets, where $\mathrm{Var}(X_1)$ is the sample variance of variable 1. We could of course also use the sample variance of variable 2 without causing much difference since at least for the datasets that we selected, the variances of the two variables do not differ much from each other. In fact, there will only be a large difference in the variances if the noises have very large variances and/or there exists a large causal effect. We could also directly set U_{11} and U_{22} to be the sample variances of X_1 or X_2, or to be the variances divided by some other constants as long as they are not too extreme. Under our setting, the hyperparameters will neither dominate the data nor be trivial.

For evaluating the seven methods with the selected benchmark datasets, we computed the causal effects in both directions, i.e., $\mathcal{C}(1 \to 2)$ and $\mathcal{C}(2 \to 1)$ as defined by (2.6) and (2.7). Recall that the posterior distribution of $\mathcal{C}(1 \to 2)$ has the formula (3.10) and is defined by both the data X^n and the hyperparameters a_{12}, U_{11}, U_{12} of the prior belief. The posterior distribution of $\mathcal{C}(2 \to 1)$, although has not been computed analytically, can be obtained directly by changing 1 to 2 and 2 to 1 in the index. We usually set

$$a_{21} := a_{12}, \quad U_{22} := U_{11} \text{ and } U_{21} := U_{12}$$

as done in Castelletti and Mascaro (2022). We plot the posterior CRs of $\mathcal{C}(1 \to 2)$ and $\mathcal{C}(2 \to 1)$ in the following Figure 4.10 and Figure 4.11. Since we know the correct causal direction of each dataset and we assume all of them to have linear relationships, we do linear regressions on all five datasets in the correct direction and use the estimated coefficients as references for our results. In the wrong causal direction, we would simply estimate the coefficient to be 0. We plot the regression coefficients as black dashed horizontal lines on top of the CRs.

Many interesting observations can be made from the results in Figure 4.10 and Figure 4.11. The most unexpected observation might be that `LRT` returns empty sets for all methods, despite its nice performance in simulation studies. An interpretation (see Strieder et al. (2021)) might be that this method constructs a null hypothesis of linear and equal variance against a general Gaussian alternative. Even if the data is almost perfectly Gaussian, equal variance is a characteristic that is very hard to achieve in the real world, so `LRT` performs poorly and can always result in empty confidence intervals. Method `SLRT` on the other hand does not have this problem and can be considered to have a nice performance, in the sense that for the first three datasets where $1 \to 2$ is the correct causal direction, `SLRT` is the only method that returns non-zero intervals for all three datasets while all other methods include 0 at least for `pair66`. However, it seems to produce remarkably wider regions than other methods in general.

Another method that produces extremely wider regions compared to others is `ETI`, especially when the true causal effect is 0 or close to 0. The reason for this has already been discussed when we do simulation studies. The method `thre0` also shows a special characteristic in the sense that it always includes 0. In simulation studies, we see that only when the sample size and the true causal effect are large enough, `thre0` can generate CRs that do not include 0. It seems hard for real-world data to have these characteristics since all CRs from `thre0` contain the point 0. The other 3 methods, `HDI`, `thre01`, and `thre05` give almost the same CRs. This is no surprise since these methods only differ in the criteria of including 0.

Among all the five datasets, `pair66` seems to be the most difficult to learn for all seven methods even though its causal effect is almost as large as `pair67`. We might be able to give a reasonable explanation for this if we look at Table 4.2 that shows the variances. We could tell an obvious difference between $\mathrm{Var}(X_1)$ and $\mathrm{Var}(X_2)$ in `pair67`. More precisely speaking, in `pair67` $\mathrm{Var}(X_2)$ is over 30% larger than $\mathrm{Var}(X_1)$. However, the variances of the two variables in `pair66` are too similar to be distinguished. So even though the causal effect is large, the causal structure is hard to be identified as there is

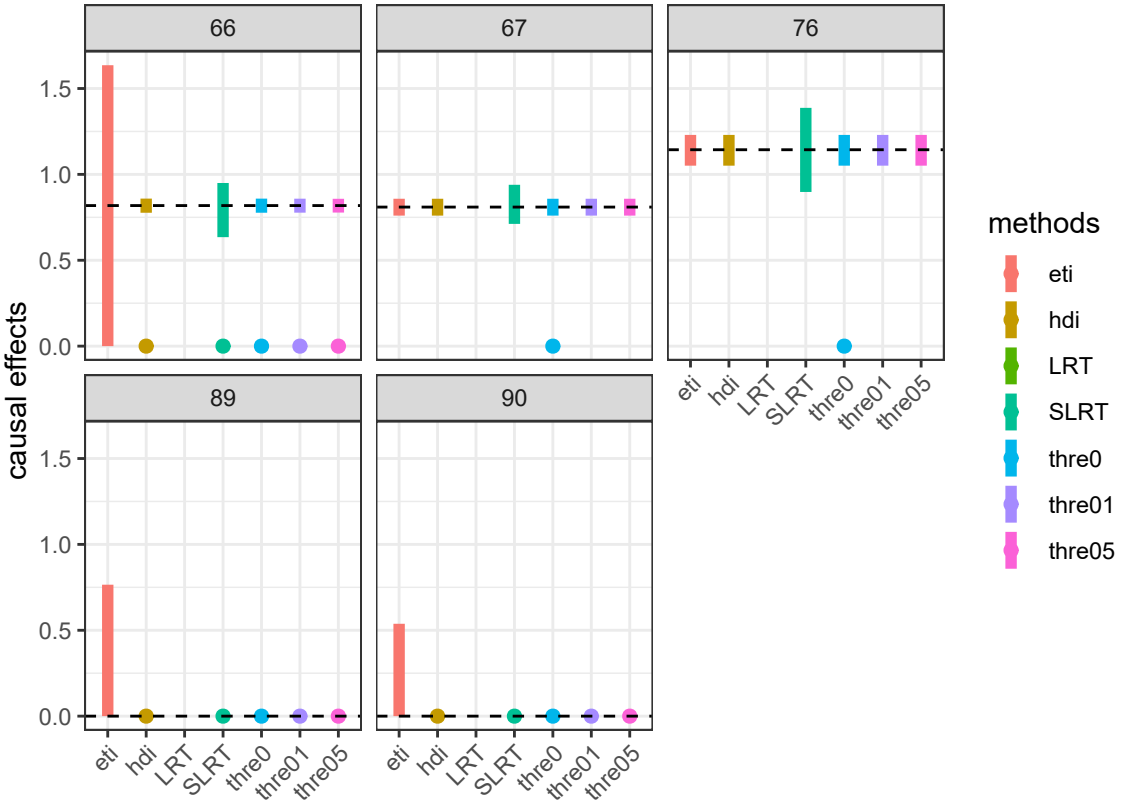


Figure 4.10.: The 95%-CRs of $\mathcal{C}(1 \to 2)$ estimated with five real-world datasets with seven methods. Each method is represented by a unique color. The black dashed lines represent a constant that is equal to the estimated coefficients of linear regressions under the correct direction and equal to 0 under the wrong direction.

not much information bias (see Mooij et al. (2016)). Looking at all other four datasets, we could at least tell whether or not there exists a non-zero causal effect with `HDI`, `SLRT`, `thre01`, and `thre05`, which is quite a nice observation. In addition, the CRs from the above-mentioned methods all include the regression coefficients, which we regard as references of the "true" causal effect given that we know the true causal direction.

Finally, we summarize the performance of our five methods by comparing them with each other and with `LRT` and `SLRT`. `ETI` produces intervals with very large widths if the causal effects are 0. This is due to the fact that the equal-tailed assumption is maybe not suitable for an asymmetric mixture distribution of a continuous distribution and a point mass. Method `thre0` will always ptoduce CRs containing 0 since we set the threshold to be extreme. Hence, it is hard to tell the causal direction if we observe the CRs generated by `thre0`, unless we have an adequate amount of data and/or the causal effect is very significant. The other three methods have quite similar performance, only `thre01` is a bit more careful about leaving 0 out of the CR. Moreover, all five methods can generalize

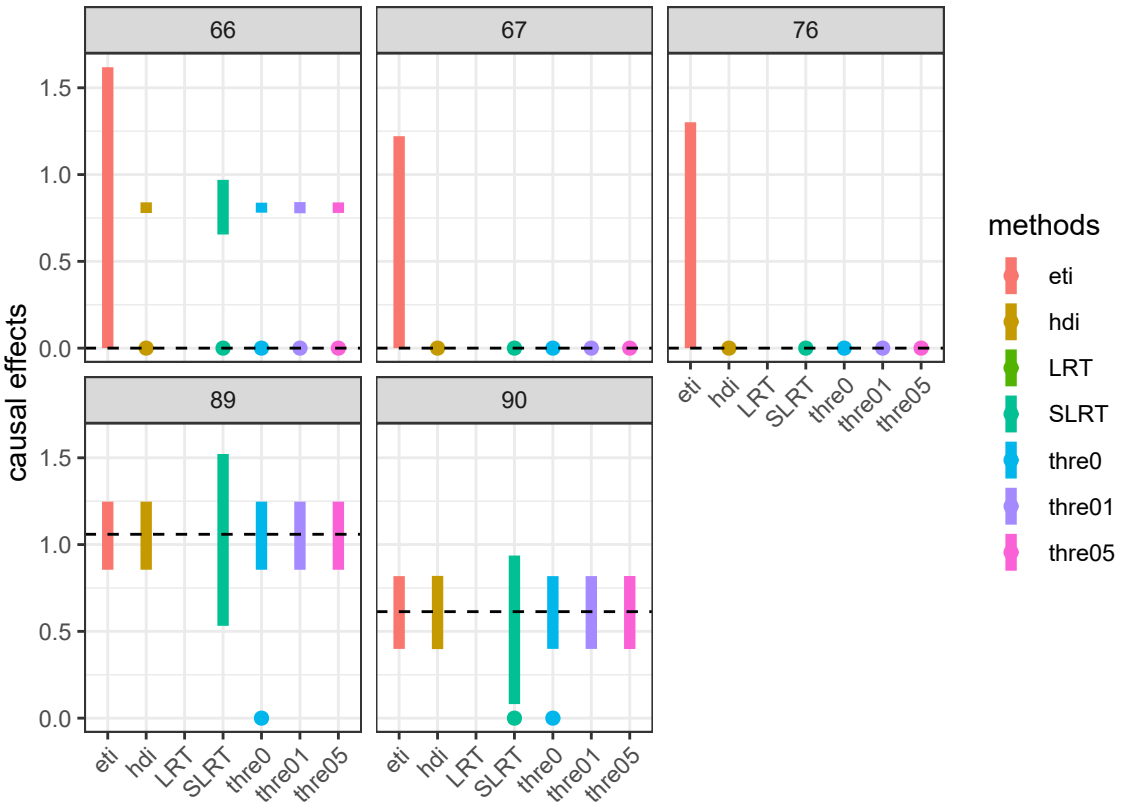


Figure 4.11.: The 95%-CRs of $\mathcal{C}(2 \to 1)$ estimated with five real-world datasets with seven methods. Each method is represented by a unique color. The black dashed lines represent a constant that is equal to the estimated coefficients of linear regressions under the correct direction and equal to 0 under the wrong direction.

nicely to real-world data, even if the assumptions might be slightly broken.

5. Conclusion

In this thesis, we proposed a Bayesian approach for causal inference. Identifying the causal structure of a group of causally efficient variables purely from their joint distribution is usually limited to Markov equivalence classes. However, several studies provide model assumptions that allow the causal structure to be uniquely identified. We follow one of the assumptions proposed by Peters and Bühlmann (2014) that assumed variables to have linear relations and homoscedastic noises. For Bayesian inference, we first set trivial priors for the underlying graphical structures, inverse Gamma prior for the unknown variance, and Gaussian prior for the non-zero causal effect. We also discussed that the hyperparameters of the priors shall be neither too powerful nor too trivial. Based on the priors and the Gaussian assumption, we derived the posterior distribution of the causal effect as a mixture of a point mass at 0 and a Student's t-distribution with the graph posterior being the weight. We proposed then three different types of $(1-\alpha)$-credible region (CR), namely the equal-tailed interval (ETI), the highest density region (HDR), and the "threshold CR". Theoretically, their difference lies in the construction methodology. Reflecting on our model, the discussion focuses on whether to include point 0 in the CR. The including principle for ETI and HDR is clear, and for the last case, it is comparing the posterior density at 0 with a self-set threshold. We discussed the results of setting different thresholds in a theoretical manner and also compared them later with experiments. To solve the problem that our model is always too sure about deciding the graphical structure, we use bootstrap averages to estimate the posterior density of the graphical structure.

We tested and compared the 95%-credible regions produced by our methods with both simulated data and benchmarks. The simulation results shown in Table 4.1 and Figure 4.3-4.6 provided us with interesting insights into how the models perform in an ideal setting. When the true causal effect is small, the models are unsure about predicting the underlying graphical structure, leading to a coverage rate lower than 95% yet still over 90%. On the other hand, except for method `ETI` which produces wide intervals because of its equal-tailed assumption, CRs produced by other methods have relatively small widths of around 0.17. The zero percentages present the differences when setting different thresholds α_0. Setting $\alpha_0 = 0$ tends to include 0 in most cases while setting $\alpha_0 = \alpha$ let the model be very careful about including 0. Our model has not much advantage in terms of running time, which might be a result of using bootstrap averages for computing the posterior distribution of the graphical structure. A positive aspect of our model is that it generalizes nicely to real-world datasets even if not all assumptions are perfectly satisfied. For datasets with larger information bias between the cause and the effect, our model can provide nicer predictions of the existence of causal relations

and the size of causal effects.

The strength of our methods is that we provide a closed-form expression for directly computing the credible region out of the posterior distribution of the causal effect. However, we just consider the very basic Gaussian linear setting, and there remain many more open questions to be studied and discussed. For example, we could generalize from linear to non-linear SEM following Hoyer et al. (2008) or from Gaussian to non-Gaussian noises following Shimizu et al. (2006). It would be also interesting to think about generalizing it from low-dimensional to high-dimensional cases. A possible challenge in the generalization might be the way we define the priors since the number of DAGs will increase tremendously as we increase the number of nodes. Moreover, it is argued in McElreath (2020) that it is better to compute an 89%-credible region rather than a 95%-credible region and this would be an interesting attempt to try out. In addition, we only applied Bayesian inference by taking the common prior from Castelletti and Consonni (2021) with some discussion about the hyperparameters yet without further arguments. Selecting and comparing the prior distributions will always be an important problem in Bayesian inference and remains to be discussed more formally.

A. Probability Theory

First, we recall the weak and strong law of large numbers for i.i.d. random variables that follows from Loeve (1977).

Definition A.1 (Convergence of Random Variables)**.** A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ **converges almost surely** to X, and we write as

$$X_n \xrightarrow[n\to\infty]{a.s.} X,$$

if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

We say a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ **converges in probability** to X, and we write as

$$X_n \xrightarrow[n\to\infty]{p} X,$$

if for every $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Next, we recall the weak and strong law of large numbers for i.i.d. random variables that follows from Loeve (1977).

Theorem A.2 (Law of Large Numbers)**.** *Let $X_1, \ldots, X_n,\ n \in \mathbb{N}$ be a list of i.i.d. random variables. If $\mathbb{E}[X_1^2] < \infty$, then $(X_i)_{i=1}^n$ satisfies the* ***weak law of large numbers****, i.e.,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p} \mathbb{E}[X_1].$$

If $\mathbb{E}[|X_1|] < \infty$, then $(X_i)_{i=1}^n$ satisfies the ***strong law of large numbers****, i.e.,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[X_1].$$

We then recall a part of the continuous mapping theorem following from Van der Vaart (2000).

Theorem A.3 (Continuous Mapping)**.** *Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. If $X_n \xrightarrow[n\to\infty]{a.s.} X$, then $g(X_n) \xrightarrow[n\to\infty]{a.s.} g(X)$.*

Finally, we recall the Höder's inequality following Loeve (1977).

Definition A.4 (Hölder's inequality)**.** Let X, Y be random variables defined on the same probability space then the **Hölder's inequality** regarding the expectation can be expressed as

$$\mathbb{E}[|XY|] \leq \left(\mathbb{E}\left[X^r\right]\right)^{\frac{1}{r}} \left(\mathbb{E}\left[X^s\right]\right)^{\frac{1}{s}},$$

for $r > 1$ and $1/r + 1/s = 1$.

B. Distributions

We characterize two important but not that basic distributions in this part by stating their density functions, expected values, and variances (see Gelman et al. (1995)).

Definition B.1 (Inverse Gamma distribution)**.** We denote an Inverse-Gamma distribution with shape $a > 0$ and scale $b > 0$ as I-Ga(a, b). The probability density function is

$$p(x) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{b/x}, \qquad x \in (0, \infty),$$

with $\Gamma(\cdot)$ denoting the gamma function.

The expected value of a random variable $X \sim$I-Ga(a, b) is given by

$$\mathbb{E}[X] = \frac{b}{a-1}, \qquad a > 1,$$

and the variance is

$$\mathrm{Var}(X) = \frac{b^2}{(a-1)^2(a-2)}, \qquad a > 2.$$

Definition B.2. (Student's t-distribution) We characterize the student's t-distribution $t_\nu(\mu, \sigma^2)$ with degrees of freedom $\nu > 0$, location μ and scale $\sigma > 0$ with probability density function

$$p(x) = \frac{\Gamma\left(\frac{\mu+1}{2}\right)}{\Gamma\left(\frac{\mu}{2}\right)\sqrt{\pi\nu}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2}.$$

A random variable X following the distribution $t_\nu(\mu, \sigma^2)$ has expected value

$$\mathbb{E}[X] = \mu, \qquad \nu > 1,$$

and variance

$$\mathrm{Var}(X) = \frac{\nu}{\nu-2}\sigma^2, \qquad \nu > 2.$$

References

Becker, R. (2018). *The new S language*. CRC Press.

Cao, X., Khare, K., and Ghosh, M. (2019). Posterior graph selection and estimation consistency for high-dimensional bayesian dag models. *The Annals of Statistics*, 47(1):319–348.

Castelletti, F. and Consonni, G. (2021). Bayesian inference of causal effects from observational data in gaussian graphical models. *Biometrics*, 77(1):136–149.

Castelletti, F. and Mascaro, A. (2022). Bcdag: An r package for bayesian structure and causal learning of gaussian dags. *arXiv preprint arXiv:2201.12003*.

Chen, W., Drton, M., and Wang, Y. S. (2019). On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980.

Duncan, O. D. (1975). Introduction to structural equation models.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

Hoyer, P. O. and Hyttinen, A. (2012). Bayesian discovery of linear acyclic causal models. *arXiv preprint arXiv:1205.2641*.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Kruschke, J. (2014). Doing bayesian data analysis: A tutorial with r, jags, and stan.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Loeve, M. (1977). *Probability Theory I*. Graduate Texts in Mathematics. Springer.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164.

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541.

Marin, J.-M. and Robert, C. P. (2014). *Bayesian essentials with R*, volume 48. Springer.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and nobel laureates. *N Engl J Med*, 367(16):1562–1564.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peters, J. (2013). When ice cream sales rise, so do homicides. coincidence, or will your next cone murder you.

Peters, J. and Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2012). Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models.

Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

Rizzo, M. L. (2019). *Statistical computing with R*. Chapman and Hall/CRC.

Royden, H. L. and Fitzpatrick, P. (1988). *Real analysis*, volume 32. Macmillan New York.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Strieder, D., Freidling, T., Haffner, S., and Drton, M. (2021). Confidence in causal discovery with linear causal models. In *Uncertainty in Artificial Intelligence*, pages 1217–1226. PMLR.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

von Kügelgen, J., Rubenstein, P. K., Schölkopf, B., and Weller, A. (2019). Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks. *arXiv preprint arXiv:1910.03962*.

Wright, S. (1921). Correlation and causation.

Zhang, J. and Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271.