Scuola di Dottorato
Università degli Studi di Milano-Bicocca

Department of

Economics, Management, and Statistics

Ph.D. program: **Economics and Statistics** Cycle: **XXXV°**
Curriculum: **Statistics**

Bayesian approaches to causal inference and discovery from observational and interventional data

Surname: **Mascaro**
Name: **Alessandro**
Registration number: **854329**

Supervisor: **Federico Castelletti**
Tutor: **Lucia Paci**

Academic Year: 2022-2023

Abstract

The notion of *external intervention* is of fundamental importance within the fields of causal inference and causal discovery from combinations of observational and experimental data. In the former, it serves as a tool for defining, identifying, and estimating causal effects. In the latter, it allows for a precise definition of what experimental data entail, enabling their usage to improve the identifiability of causal structures. This thesis addresses both aspects, investigating them within the Bayesian framework. The manuscript comprises three self-contained chapters, with the first serving as an introduction to the general scope of the thesis, outlining its scientific context and contribution. The main two chapters represent two independent projects, both driven by the objective of broadening the notion of external intervention to reflect the diverse manipulations that scientists may actually implement in their experiments. In Chapter 2, we consider the case of joint interventions which may simultaneously affect several variables. In particular, we present a unified Bayesian approach for causal discovery and causal effect estimation in the Gaussian setting. This leads to a Bayesian model averaging strategy for estimating the *joint* causal effects associated with such interventions when the causal structure of the data-generating process is unknown. In Chapter 3, we instead consider interventions modifying the causal mechanisms of the intervened variables, which we call *general interventions*. We thus construct a Bayesian model for causal discovery from combinations of observational and experimental data originating from unknown general interventions. In addition, we provide definitions and graphical characterizations of the identifiability limit of causal structures in the new setting and devise a suitable MCMC scheme to sample from the joint posterior distribution over causal structures and unknown general interventions.

Contents

Abstract **iii**

Contents **v**

List of Figures **ix**

List of Algorithms **xiii**

Acknowledgements **xv**

1 Introduction **1**

1.1 Causal inference from experimental and observational data 1

1.2 Structural Causal Models 2

1.3 Causal discovery 4

1.3.1 From observational data 4

1.3.2 From experimental data 6

1.4 Bayesian causal discovery 8

1.5 Outline and contribution 10

Bibliography 11

2 Bayesian causal discovery and joint causal effect estimation **17**

2.1 Introduction 17

2.1.1 Related work 19

2.1.2 Outline 19

2.2 Preliminaries 20

2.2.1 Gaussian DAG-models 20

2.2.2 Identifying causal effects in Gaussian DAG-models 21

2.2.3 Identifying joint causal effects 23

2.3 Bayesian inference on causal effects under model uncertainty . . . 24
2.3.1 Model formulation . . . 25
2.3.2 Prior on DAG parameters . . . 25
2.3.3 Prior on DAG $\mathcal{D}$ . . . 27
2.4 MCMC scheme and posterior inference . . . 28
2.4.1 Sampling scheme . . . 28
2.4.2 Posterior inference . . . 31
2.5 Simulations and real data analysis . . . 32
2.5.1 DAG selection . . . 32
2.5.2 Causal effect estimation . . . 33
2.5.3 Real data analysis . . . 36
2.6 Discussion . . . 38
2.6.1 Future developments . . . 40
Bibliography . . . 41

3 Bayesian causal discovery from unknown general interventions 45
3.1 Introduction . . . 45
3.1.1 Related work . . . 47
3.1.2 Outline . . . 48
3.2 Identifiability under general interventions . . . 48
3.2.1 Preliminaries . . . 48
3.2.2 DAG identifiability from known general interventions . . . 51
3.2.3 DAG identifiability from unknown general interventions . . . 55
3.3 Bayesian causal discovery . . . 59
3.3.1 Model formulation . . . 59
3.3.2 Parameter prior elicitation . . . 60
3.3.3 Prior on $(\mathcal{D}, \mathcal{I})$ . . . 62
3.4 MCMC scheme and posterior inference . . . 63
3.4.1 Sampling scheme . . . 64
3.4.2 Posterior inference . . . 67
3.5 Simulations and real data analysis . . . 68
3.5.1 Gaussian DAGs . . . 68
3.5.2 Simulation studies . . . 70
3.5.3 Real data analysis . . . 74
3.6 Discussion . . . 76
3.6.1 Future developments . . . 79

Appendix 81
Bibliography 95

List of Figures

1.1 Four DAGs with the same skeleton. $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_3$ have the set of v-structures and are thus Markov equivalent. $\mathcal{D}_4$ belongs to another Markov equivalence class. 5

1.2 Four DAGs with the same skeleton (first row) and their post-intervention DAG after a hard intervention on node 2 (second row). Each colour corresponds to an I-Markov equivalence class. 7

2.1 A DAG with $q = 6$ node/variables and randomly generated edge weights. . . 18

2.2 A DAG $\mathcal{D}$ and three modified graphs of the operators $Insert(4, 1)$, $Delete(1, 3)$, $Reverse(1, 3)$ respectively. Operator $Insert(4, 1)$ is *not* valid since $\mathcal{D}'_1$ is not acyclic. 29

2.3 Simulation study. Distribution over $N = 30$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true CPDAGs. Methods under comparison are: our Bayesian method with output the Median Probability Model (MPM) and Maximum A Posteriori (MAP) graph estimates and the PC algorithm implemented at significance level $\alpha \in \{0.01, 0.05, 0.10\}$, respectively PC 0.01, PC 0.05, PC 0.10. 34

2.4 Simulation study. Distribution of the absolute-value distance d between estimated and true causal effects for size of the target $s \in \{2, 4\}$, number of variables $q \in \{10, 20\}$ and sample size $n \in \{50, 100, 200, 500\}$. Methods under comparison are: our BMA-based approach (BMA) and the *Joint-IDA* method (IDA). 35

2.5 Real data analysis. Heat maps with estimated posterior probabilities of edge inclusion obtained from two independent MCMC chains 37

2.6 Real data analysis. Comparison between estimated graphs. From left to right: maximum a posteriori, median probability and modified-PC DAG estimates 38

2.7 Real data analysis. Left panel: BMA (top) and joint-IDA (bottom) estimates of the causal effect of X_k on Y in a joint intervention on $\{X_h, X_k\}$. Right panel: sum of absolute-value BMA (top) and joint-IDA (bottom) estimates obtained from joint interventions on $\{X_h, X_k\}$. 39

3.1 Three DAGs resulting from different types of interventions: a) a hard intervention on $\mathbf{TR}_t$; b) simultaneous hard (on $\mathbf{TR}_t$) and soft (on $\mathbf{AQ}_t$) interventions; c) a general intervention on $\mathbf{TR}_t$. Target nodes are depicted in blue, while structural modifications induced by the interventions are colored in red. 46

3.2 A DAG $\mathcal{D}$ and the post-intervention DAG $\mathcal{D}_{T,P}$ for intervention target $T = \{3\}$ and induced parent set $P = \{2\}$. 53

3.3 A collection of $\mathcal{I}$-DAGs for DAG $\mathcal{D}$ and a collection of targets and induced parent sets such that $T^{(2)} = \{3\}$, $P^{(2)} = \{1, 2\}$ and $T^{(3)} = \{4\}, P^{(3)} = \{1, 2, 3\}$. Blue nodes represent the intervention targets, while red edges correspond to the induced parent sets. 53

3.4 Three Markov equivalent DAGs and their post-intervention graphs after a general intervention with $T^{(2)} = \{3\}$, $P^{(2)} = \{2\}$. The intervention is not valid for $\mathcal{D}_3$. 56

3.5 Two unidintifiable combinations of DAGs and general interventions. 57

3.6 Simulations. Distribution (across 40 simulations) of the Structural Hamming Distance (SHD) between true DAG and graph estimate, under scenarios $q \in \{10, 20\}$ (number of variables), $K \in \{2, 4\}$ (number of experimental contexts), and for increasing samples sizes $n_k \in \{100, 500, 1000\}$. Methods under comparison are: GIES and GnIES (dark and light blue), UT-IGSP (yellow) and our Bayesian approach (red). 72

3.7 Simulations. Distribution (across 40 simulations) of the number of false positives and false negatives (# of errors) between true and estimated targets, under scenarios $q \in \{10, 20\}$ (number of variables), $K \in \{2, 4\}$ (number of experimental contexts), and for increasing samples sizes $n_k \in \{100, 500, 1000\}$. Methods under comparison are: GnIES (light blue), UT-IGSP (yellow) and our Bayesian approach (red). 73

3.8 Simulations. Distribution (across 40 simulations) of the sum of falsely identified and non-identified varying edges between context $k = 1$ and $k = 2$, under scenarios $q \in \{10, 20\}$ (number of variables) and for increasing samples sizes $n_k \in \{100, 500, 1000\}$. Method under comparison are: DCI and DCI with stability selection (dark and light blue) and our Bayesian approach (red). . 74

3.9 AML data. Estimated marginal posterior probabilities of target inclusion, computed for each node $v \in [q]$ across AML subtypes, each corresponding to an experimental context k. Subtype M2 corresponds to the reference (observational) context. 75
3.10 AML data. Estimated marginal posterior probabilities of edge inclusion, computed for each possible directed edge (u, v), $u, v \in [q]$ and group-specific post-intervention DAG, each corresponding to one of the four AML subtypes. 77
3.11 AML data. Median Probability graph Model (MPM) estimates obtained under each AML subtype. Graph corresponding to Subtype M2 is the representative I-EG. 78

List of Algorithms

1 PAS algorithm for posterior inference. 29
2 PAS algorithm: construction of $\mathcal{O}_{\mathcal{D}}$ 30

3 Random-scan MH for posterior inference 65
4 Random-scan MH: Construction of $\mathcal{O}_{\mathcal{D}}$ 66
5 Random-scan MH: Construction of $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$ 66
6 Find-Edge (Chickering, 1995) 86
7 Markov chain defined by the proposal distribution of Algorithm 3 94

Acknowledgements

First and foremost, I must thank my supervisor, Federico Castelletti. His brilliant scientific guidance, unwavering patience, and generosity with his time have been invaluable throughout these years. I owe a debt of gratitude to Guido Consonni, whose mentorship and advice have been fundamental in introducing me to the academic world. His extensive knowledge and passion for Bayesian statistics have been a constant source of inspiration. Finally, I would also like to thank David Rossell for hosting me at Universitat Pompeu Fabra and for his brilliant mentoring during my stay there.

On a more personal level, a special mention goes to my fellow PhD students Alice, Federico and Vincenzo. Our journey was challenging and full of unexpected events, but it would have been harder without them, and I keep wonderful memories of our time together. I also thank my family for their love and sacrifices throughout my entire educational and academic career. Last but certainly not least, I want to thank Chiara for her caring support during these years. Her unwavering companionship has been an invaluable source of strength and has made everything much, much easier.

Chapter 1

Introduction

1.1 Causal inference from experimental and observational data

Many of the most critical questions of scientific and societal interest revolve around causality. For instance, in recent years, marked by the need to coexist with the COVID-19 pandemic, the ability to assess the causal effects of infection-reduction policies has been crucial in ensuring safety and minimizing the pandemic's social costs [Bonvini et al., 2022]. In the coming decades, addressing the challenge of climate change requires a profound understanding of its causal mechanisms. Such understanding is essential for evaluating the most effective policies to mitigate its impact [Nowack et al., 2020].

In statistics, particularly in causal inference, distinguishing causal effects from mere statistical associations is a fundamental challenge. Randomized Controlled Trials (RCTs) [Fisher, 1935] provide the primary and oldest tool for this purpose. In an RCT, researchers randomly assign a selected sample of subjects to either a control or a treatment group. The causal effect is then defined as the difference in the observed outcome variable between the two groups. Despite being considered the gold standard of causal inference, RCTs can often be infeasible or unethical. For example, if we wanted to assess the effect of smoking on the onset of respiratory issues, implementing an RCT would require randomly chosen subjects to start smoking, which would be unethical and thus infeasible. In such cases, the only option is to rely on observational data, which are non-experimental in nature. Observational data are inherently less informative than experimental data because they lack information about how the system responds to external stimuli. Consequently, causal inference from observational data is more complex and requires stronger assumptions to validate its conclusions.

The fields of statistics, econometrics, and computer science have developed various conceptual frameworks for causal inference from observational data. In this thesis, we

adopt the *do-calculus* framework based on graphical models [Pearl, 2009]. This framework requires specifying the scientific knowledge about the problem under analysis through a causal diagram, a graphical object in which every node corresponds to a variable and every edge to a direct causal relationship. From this causal diagram, the do-calculus provides the functional of the observed distribution corresponding to the causal effect of interest.

Specifying a causal diagram, or even parts thereof, is difficult in many cases. Furthermore, in numerous practical contexts, the causal diagram is of scientific interest in its own right. These considerations led to the development of another body of literature known as *causal discovery* [Spirtes et al., 2001, Peters et al., 2017]. This literature primarily addresses two key aspects. First, it investigates the problem of determining how much can be learned of a causal diagram based on weaker assumptions about the data-generating process. Second, it develops procedures for estimating the identifiable structures. One can then use these estimated structures to define the associated causal effects.

Learning a causal structure without any prior information is highly complicated. Even assuming that we observe all the relevant variables for explaining a phenomenon, the number of possible identifiable structures grows super-exponentially in the number of variables involved [Gillispie and Perlman, 2001]. Consequently, it often becomes necessary to include prior information in the learning process and to quantify the uncertainty around our estimates. For this reason, in this thesis, we adopt a Bayesian approach

The upcoming sections formally introduce several key concepts crucial for understanding the articles that constitute this doctoral thesis. Specifically, in Section 1.2, we introduce the framework of Structural Causal Models (SCMs), which forms the basis of the do-calculus, and highlight its underlying assumptions. Section 1.3 addresses the issue of causal discovery and outlines the primary methodologies used in this context. In Section 1.4, we frame the problem of causal discovery within the Bayesian setting. Finally, Section 1.5 briefly outlines the two main chapters of the thesis, emphasizing their contributions.

1.2 Structural Causal Models

Let $\mathcal{D} = (V, E)$ be a Directed Acyclic Graph (DAG) with vertex set $V = \{1, \dots, q\} \coloneqq [q]$ and edge set $E \subset V \times V$. We denote by $\mathrm{pa}_{\mathcal{D}}(j)$ the parent-set of node j, i.e. $\mathrm{pa}_{\mathcal{D}}(j) = \{i \in V | (i, j) \in E\}$. DAGs are given a causal interpretation when considered as graphical representations of a *Structural Causal Model* (SCM), each node being a variable and each edge a direct causal relationship. An SCM consists of stable and autonomous mechanisms of the form

$$X_j = f_j\left(X_{\mathrm{pa}_{\mathcal{D}}(j)}, \epsilon_j\right), \quad \epsilon_j \sim p_{\epsilon_j} \qquad j \in [q], \tag{1.1}$$

where $X \coloneqq \{X_1, \dots, X_q\}$ is a collection of *endogenous* variables corresponding to the nodes of $\mathcal{D}$, $\{\epsilon_j\}_{j=1}^q$ a collection of *exogenous* variables with joint distribution p_ϵ, and $\{f_j\}_{j=1}^q$ a set of *structural assignments*, i.e. functions connecting the value of each X_j to its parents/causes $X_{\mathrm{pa}_\mathcal{D}(j)}$ and to ϵ_j. If the exogenous variables in ϵ are independent, then the SCM is called *Markovian*, and a distribution $p(\cdot)$ is induced over the endogenous variables X such that

$$p(x) = \prod_{j=1}^{q} p\left(x_j | x_{\mathrm{pa}_\mathcal{D}(j)}\right), \tag{1.2}$$

where by x we denote any realization of X. Through the Markovian assumption, the exogenous variables induce a set of conditional independencies that result in the factorization (1.2). Because of the assumed stability and autonomy of the mechanisms in (1.1), it is possible to conceive external interventions that perturb only a subset of the mechanisms, leaving the others unchanged. In particular, a *hard* intervention on X_i consists in fixing the value of the intervened variable to a chosen constant value $\tilde{x}_i$, and we denote it by $\mathrm{do}(X_i = \tilde{x}_i)$, or $\mathrm{do}(\tilde{x}_i)$ for short. Accordingly, any intervention also induces a post-intervention distribution on X. In the case of a hard intervention, it takes the form of the truncated factorization

$$p_X\left(x \mid \mathrm{do}\left(X_i = \tilde{x}_i\right)\right) = \prod_{j \neq i} p\left(x_j | x_{\mathrm{pa}_\mathcal{D}(j)}\right) \mathbb{1}\{x_i = \tilde{x}_i\}. \tag{1.3}$$

An SCM thus entails not only an observational distribution over X but also how this distribution would change in response to hypothetical external interventions. By leveraging the language of SCMs and the do-calculus, it is possible to define many functionals of interest for causal inference. For instance, the causal effect of the treatment variable X_i on an outcome variable $Y \in X, Y \neq X_i$ can be defined as

$$\gamma_{x_i y} \coloneqq \frac{\partial}{\partial \tilde{x}_i} \mathbb{E}\left(Y \mid \mathrm{do}\left(X_i = \tilde{x}_i\right)\right). \tag{1.4}$$

In other words, the causal effect is the induced variation in the expected value of Y resulting from an infinitesimal change in the value $\tilde{x}_i$ at which we fix the treatment variable X_i. Once the causal effect is defined within the language of SCMs, the do-calculus enables us to ascertain whether this causal effect is identifiable from observational data. If it is, the do-calculus also identifies the corresponding functional of the distribution $p(\cdot)$, thus making the estimation of causal effects from observational data possible.

Finally, it is worth noting that hard interventions do not encompass the full spectrum of possible manipulations that can be applied to a set of variables. For instance, one may consider and perform simultaneous interventions on a set of variables, or interventions that

modify the mechanisms of the target (treatment) variables; see Korb et al. [2004] for a discussion. Consequently, different types of manipulations require different definitions of causal effects and, in turn, different identification and estimation strategies. For this reason, extensions of the do-calculus, such as the σ-calculus of Correa and Bareinboim [2020], have been developed in the literature.

1.3 Causal discovery

As anticipated, in many practical contexts, specifying a causal diagram is not feasible. Therefore, learning such diagram from data becomes essential both for its intrinsic scientific interest and the subsequent possibility of identifying and estimating causal effects. This learning process relies on assumptions about the data-generating process. The question we address in this section is the following: "How can we leverage these assumptions to extract information about the causal diagram $\mathcal{D}$ given knowledge of $p(\cdot)$?" In Section 1.3.1, we elaborate on the Markovian assumption implying the factorization (1.2) and its usage in causal discovery from observational data. In Section 1.3.2, we instead focus on the assumption of stability and autonomy of the mechanisms of an SCM, showing how it can be leveraged to improve DAG identifiability when experimental data are available.

1.3.1 From observational data

Every Markovian SCM admits a DAG implies the factorization (1.2), which entails a set of conditional independence relationships among variables. It is possible to characterize the set of all conditional independencies implied by a DAG $\mathcal{D}$ using a graphical criterion called d-separation [Pearl, 2009]. Such criterion relies on the notion of a *collider*. Given a path $p = (p_1 = i, p_2, \dots, p_M = j)$ from i to j in $\mathcal{D}$, the node p_m is a collider if $p_{m-1} \to p_m \leftarrow p_{m+1}$. A path p d-connects i and j given the set $C \subseteq [q] \backslash \{i, j\}$ if:

1. All non-colliders on the path do not belong to C;
2. All colliders on the path either belong to C, or have a descendant which belongs to C.

Finally, i and j are d-connected given C if there exists any d-connecting path given C; otherwise, they are d-separated. If i and j are d-separated by C in $\mathcal{D}$, we write $i \perp\!\!\!\perp_{\mathcal{D}} j \mid C$. We denote with $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{D})$ the set of d-separation statements implied by a DAG $\mathcal{D}$; i.e.

$$\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{D}) = \{(i, j, C) \mid i, j \in [q], C \subseteq [q] \backslash \{i, j\}, i \perp\!\!\!\perp_{\mathcal{D}} j \mid C\}. \tag{1.5}$$

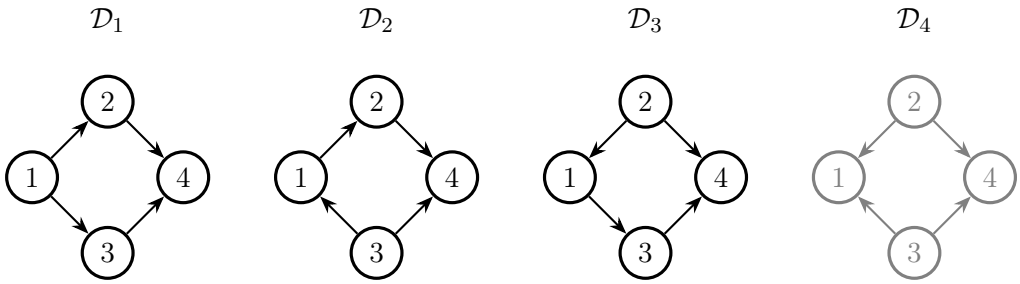


Figure 1.1: *Four DAGs with the same skeleton. $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_3$ have the set of v-structures and are thus Markov equivalent. $\mathcal{D}_4$ belongs to another Markov equivalence class.*

Similarly, we denote with $\mathcal{I}_{\perp\!\!\!\perp}(p)$ the set of conditional independencies in $p(\cdot)$, which is defined as

$$\mathcal{I}_{\perp\!\!\!\perp}(p) = \{(i, j, C) \mid i, j \in [q], C \subseteq [q] \backslash \{i, j\}, X_i \perp\!\!\!\perp X_j \mid C\}. \tag{1.6}$$

We say that $p(\cdot)$ satisfies the *global Markov property* of $\mathcal{D}$ if $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{D}) \subseteq \mathcal{I}_{\perp\!\!\!\perp}(p)$, i.e. if all the d-separation statements in $\mathcal{D}$ imply a conditional independence relation in p [Pearl, 1988]. In addition, we say that p is *faithful* to $\mathcal{D}$ if $\mathcal{I}_{\perp\!\!\!\perp}(p) = \mathcal{I}_{\perp\!\!\!\perp}(\mathcal{D})$, so that it is possible to enumerate all and only the d-separation statements that must hold in the DAG $\mathcal{D}$ and that correspond to the conditional independencies of p. However, even assuming faithfulness it is not possible to distinguish between Markov equivalent DAGs. Two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are *Markov equivalent* if $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{D}_1) = \mathcal{I}_{\perp\!\!\!\perp}(\mathcal{D}_2)$, i.e. they imply the same set of conditional independencies via d-separation. We let $\mathcal{M}(\mathcal{D})$ be the Markov equivalence class of $\mathcal{D}$, that is the set of all DAGs implying the same conditional independencies as $\mathcal{D}$. Markov equivalence classes represent the identification limit when only an observational distribution $p(\cdot)$ is available and the only assumption on the data-generating process is that it can be represented through a Markovian SCM ((1.1)). Different graphical characterizations of Markov equivalence exist. In particular, two DAGs are Markov equivalent if and only if they have the same skeleta and the same set of v-structures [Verma and Pearl, 1990]. See Figure 1.1 for an example.

The task of causal discovery from observational has been tackled from different methodological perspectives, both from a frequentist and from a Bayesian standpoint. A primary distinction among frequentist methods is between *constraint-based* and *score-based* algorithms. The former include algorithms that recover the DAG-equivalence class through sequences of conditional independence tests. The most popular methods are the PC and Fast Causal Inference (FCI) algorithms [Spirtes et al., 2001], together with their extensions rankPC and rankFCI [Harris and Drton, 2013], based on more general (non-parametric)

conditional independence tests. Differently, score-based methods implement a suitable score function which is maximized over the space of DAGs (or their equivalence classes) to provide a graph estimate; an example is the Greedy Equivalence Search (GES) algorithm of Chickering [2002]. Going beyond this distinction, a variety of hybrid methods, i.e. combining features of both the two approaches, have been proposed; see for instance Tsamardinos et al. [2006] and Shimizu et al. [2006], the latter tailored to non-Gaussian linear structural equation models. On the Bayesian side, DAG structure learning has been traditionally tackled as a Bayesian model selection problem. In this framework the target is represented by the posterior distribution of causal structures which is typically approximated through Markov Chain Monte Carlo (MCMC) methods; see Section 1.4 for details. The first work going in this direction is Cooper and Herskovits [1992]. More recent works have focused on sampling from spaces which are "coarser" than the one of DAG structures, such as the space of Markov equivalence classes [Castelletti et al., 2018], the space of orderings [Friedman and Koller, 2003], the space of partitions [Kuipers and Moffa, 2017], or the space of minimal I-MAPs [Agrawal et al., 2018].

For extensive reviews of causal discovery methods, the reader can refer to Heinze-Deml et al. [2018] and Squires and Uhler [2023].

1.3.2 From experimental data

We now consider the case when we observe X in K different experimental settings, and we denote the associated distributions as $\{p_k(x)\}_{k=1}^K$. Each experimental setting corresponds to a specific perturbation performed by an external experimenter on a set of target variables indexed by $T^{(k)} \subseteq [q]$. In the framework of SCMs, this corresponds to an intervention that, because of the stability and autonomy assumption, modifies the generating mechanism of the target nodes $T^{(k)}$ without affecting the others and accordingly, as shown in (1.3), results in a *local* change in the post-intervention distribution.

In what follows, we assume that the performed interventions are hard, but *stochastic*, that is we allow for randomness in the process of fixing the values of the target nodes. Given an observational distribution $p(\cdot)$, the post-intervention distribution induced in the k-th experimental context is thus

$$p_k(x) = \prod_{j \notin T^{(k)}} p\left(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}\right) \prod_{j \in T^{(k)}} p_k\left(x_j\right). \tag{1.7}$$

The assumption of stable and autonomous mechanisms translates into a set of invariances between the observational and the post-intervention distributions. Moreover, a hard

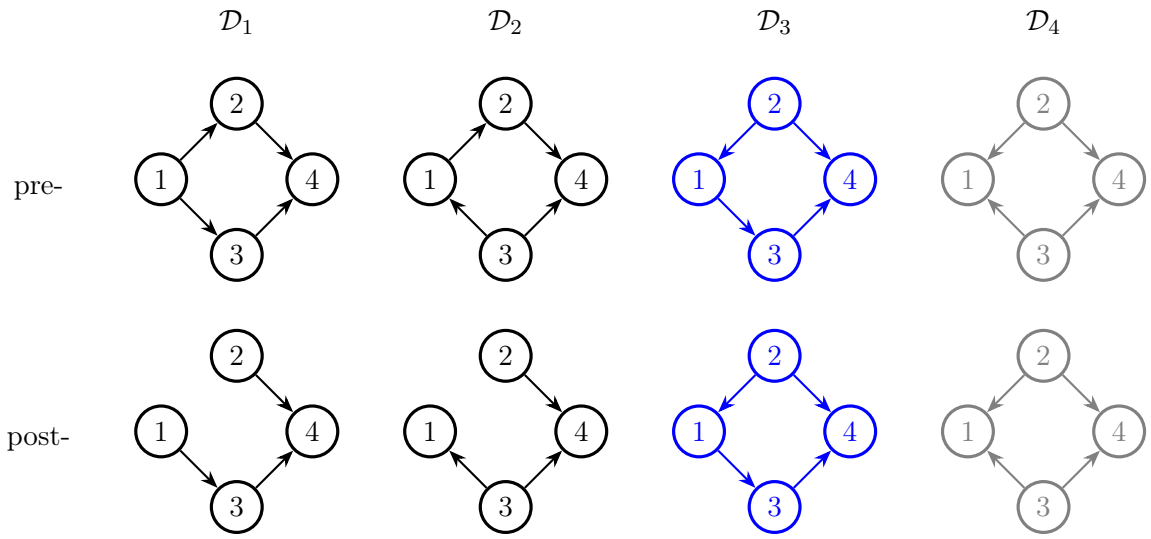


Figure 1.2: *Four DAGs with the same skeleton (first row) and their post-intervention DAG after a hard intervention on node* 2 *(second row). Each colour corresponds to an I-Markov equivalence class.*

intervention removes the dependence between the intervened node and its parent nodes in the DAG. This information can then be used to enhance the identifiability of the DAG, inducing a partition of the DAG space into *I-Markov equivalence classes*, which is finer than the one implied by Markov equivalence.

Figure 1.2 illustrates three Markov equivalent DAGs and their corresponding post-intervention DAGs resulting from a hard intervention on the variable X_3. Notice that, in the post-intervention DAGs of $\mathcal{D}_1$ and $\mathcal{D}_2$, we have that $X_1 \perp\!\!\!\perp X_3$, while in the post-intervention DAG of $\mathcal{D}_3$, the dependence relationship between the two variables is preserved. This information can be used to distinguish between $\mathcal{D}_1$ and $\mathcal{D}_2$, on one hand, and $\mathcal{D}_3$, on the other. DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ entail both the same invariances and the same conditional independences in both the observational and experimental context. Therefore, they are I-Markov equivalent. As for the observational case, there exist graphical characterizations of I-Markov equivalence for the case of hard interventions. In particular, Hauser and Bühlmann [2012] proposed the following. For any DAG $\mathcal{D}$ and any collection of K experimental settings, we denote with $\{\mathcal{D}^{(k)}\}_{k=1}^{K}$ the collection of post-intervention DAGs. Two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are I-Markov equivalent if, for any $k \in [K]$, $\mathcal{D}_1^{(k)}$ and $\mathcal{D}_2^{(k)}$ have the same skeleta and the same set of v-structures.

The first article on causal discovery from combinations of observational and experimental data is Cooper and Yoo [1999], who proposed a Bayesian methodology for data

arising from hard interventions with known targets. In the same Bayesian framework, Tian and Pearl [2001] presented a similar methodology using a different notion of intervention called *mechanism change*. Eaton and Murphy [2007] extended the previous methodologies to the case of categorical data and of interventions with unknown targets. An objective Bayesian methodology in the Gaussian setting was then proposed by Castelletti and Consonni [2019]. The same authors provided an extension to the unknown targets case [Castelletti and Peluso, 2023]. On the frequentist side, Hauser and Bühlmann [2012] developed the Greedy Interventional Equivalence Search (GIES) algorithm as an extension of GES that can handle combinations of experimental data. Recently, Gamella et al. [2022] extended the same methodology to the case of unknown targets in the Gaussian setting. A different score-based approach was proposed by Wang et al. [2017], who developed the Interventional Greedy Sparsest Permutation (IGSP) method, later extended to the case of soft interventions by Yang et al. [2018] and to the case of unknown targets by Squires et al. [2020]. Finally, Mooij et al. [2020] developed the Joint Causal Inference framework, which encodes unknown interventions through additional indicator variables in a pooled dataset and establishes under which assumptions constraint-based methods conceived for observational settings can be applied to the pooled dataset to learn the DAG when the intervention targets are unknown.

1.4 Bayesian causal discovery

In the Bayesian setting, causal discovery can be cast as a Bayesian model selection problem. Let $\boldsymbol{X}$ be a (n, q) matrix collecting n i.i.d. measurements of a random vector X generated by a Markovian SCM with unknown causal diagram $\mathcal{D}$. Let also Θ be the set of parameters associated with the corresponding parametric DAG-model $p(\boldsymbol{X} \mid \Theta, \mathcal{D})$. The factorization (1.2) implies that:

$$p(\boldsymbol{X} \mid \Theta, \mathcal{D}) = \prod_{i=1}^{n} \prod_{j=1}^{q} p(\boldsymbol{X}_{ij} \mid \boldsymbol{X}_{i,\mathrm{pa}_{\mathcal{D}}(j)}, \theta_j, \mathcal{D}) \tag{1.8}$$

where $\theta_j \subset \Theta$ is the subset of parameters associated with the j-th node. When coupled with a prior distribution on the space of all DAGs $p(\mathcal{D})$ and a prior on the associated DAG-model parameters $p(\Theta \mid \mathcal{D})$, the statistical model of Equation (1.8) implies a posterior distribution on the DAG space

$$p(\mathcal{D} \mid \boldsymbol{X}) \propto p(\boldsymbol{X} \mid \mathcal{D}) p(\mathcal{D}) \tag{1.9}$$

where $p(\boldsymbol{X}|\mathcal{D})$ is the *marginal likelihood* or *Bayesian model evidence*, defined as:

$$p(\boldsymbol{X} \mid \mathcal{D}) = \int p(\boldsymbol{X} \mid \Theta, \mathcal{D}) p(\Theta \mid \mathcal{D}) \, d\Theta. \tag{1.10}$$

Although conceptually straightforward, the task of Bayesian causal discovery presents many specific challenges. First, the number of parameters associated with each DAG-model can be very large, and defining a posterior distribution over the space of DAGs requires to specify a prior on the parameters of any DAG-model. Moreover, the possibility that some of these models share parts of their structures brings forth a compatibility requirement [Roverato and Consonni, 2003], so that for any two DAGs $\mathcal{D}_1, \mathcal{D}_2$

$$p(\theta_j \mid \mathcal{D}_1) = p(\theta_j \mid \mathcal{D}_2) \quad \text{for each } j \in [q] \mid \mathrm{pa}_{\mathcal{D}_1}(j) = \mathrm{pa}_{\mathcal{D}_2}(j)$$

In addition to that, a further compatibility requirement emerges as a consequence of the existence of Markov equivalence classes of DAGs. In particular, for any two Markov equivalent DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ we require

$$p(\boldsymbol{X} \mid \mathcal{D}_1) = p(\boldsymbol{X} \mid \mathcal{D}_2) \tag{1.11}$$

that is, two indistinguishable models remain such even after we have specified our prior distributions on their parameter space [Peluso and Consonni, 2020].

Despite these challenges, the Bayesian approach to causal discovery has significant advantages. First, the output of a Bayesian causal discovery procedure is a whole posterior distribution over the model space, which is inherently much richer than a single DAG-estimate and naturally incorporates uncertainty quantification. Second, within the same approach one may also obtain a joint posterior distribution over DAG and DAG-parameters:

$$p(\mathcal{D}, \Theta \mid \boldsymbol{X}) \propto p(\boldsymbol{X} \mid \Theta, \mathcal{D}) p(\Theta \mid \mathcal{D}) p(\mathcal{D}) \tag{1.12}$$

and sample from it using a suitable MCMC scheme. In some parametric models, the causal effects identified by the do-calculus are in the end just functions of the parameters Θ. Denote with $\gamma(\Theta)$ the causal effect of interest, as defined in (1.4). If S samples from the posterior distribution (1.12) are available, than one may easily produce a Bayesian Model Averaging (BMA) estimate of γ as:

$$\hat{\gamma}^{\mathrm{BMA}} = \frac{1}{S} \sum_{s=1}^{S} \gamma\left(\Theta^{(s)}\right) \tag{1.13}$$

where $\Theta^{(s)}$ is the s-th sample from the MCMC scheme used. The above estimate would

incorporate the uncertainty on the causal discovery procedure. More in general, one may consider the samples $\gamma(\Theta^{(1)}), \dots, \gamma(\Theta^{(S)})$ as samples from the posterior distribution of the causal effect of interest and use them to provide estimates and quantify the uncertainty surrounding those estimates.

1.5 Outline and contribution

This manuscript is composed of three self-contained chapters, Chapter 1 being this technical introduction to causal inference and causal discovery in the Bayesian setting.

In Chapter 2, we consider the case of joint interventions which may simultaneously affect several variables. We thus specialize the general approach presented in Section 1.4 to the Gaussian setting and provide a Bayesian methodology for estimating causal effects of joint interventions when the DAG is unknown. We show how the do-calculus identifies these causal effects as DAG-specific functions of the elements of the covariance matrix of a Gaussian DAG-model. We thus derive the joint posterior distribution over DAGs and DAG-parameters and implement an MCMC scheme to sample from it. A posterior distribution over DAGs and causal effects is then obtained by transforming the sampled values of the covariance matrix. Our Bayesian model specification is based on a *compatible* variation of the DAG-Wishart distribution [Ben-David et al., 2015], which assigns equal marginal likelihood to Markov equivalent DAGs. Our proposal has the advantage, over its frequentist counterparts, of naturally accounting for the uncertainty in learning the DAG from data and it has been shown to outperform its competitors in simulation studies. Chapter 2 is based on the articles *"Structural learning and estimation of joint causal effects among network-dependent variables"* [Castelletti and Mascaro, 2021] and the accompanying article *"BCDAG: An R package for Bayesian structure and Causal learning of Gaussian DAGs"* [Castelletti and Mascaro, 2022], detailing a public available implementation of the methods proposed in the package `BCDAG`.

In Chapter 3, we shift our focus on interventions that modify the parent sets of the intervened nodes in the DAG, which we call *general interventions*. We thus propose a Bayesian methodology for causal discovery from experimental data arising from general interventions with unknown targets. DAGs and unknown general interventions may be identifiable only up to some equivalence class. We provide graphical characterizations of such equivalence classes, and, accordingly, we devise compatible priors that guarantee score equivalence of indistinguishable combinations. Finally, we develop a suitable MCMC scheme to sample from the posterior distribution over DAGs and unknown interventions. We evaluate the proposed methodology on both simulated and real datasets. The performance of the method is

competitive with state-of-the-art methods both on the task of structure learning and on the task of learning the difference between different causal structures. Chapter 3 is based on the working paper "Bayesian causal discovery from unknown general interventions" [Castelletti and Mascaro, 2023+].

Bibliography

R. Agrawal, C. Uhler, and T. Broderick. Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 89–98. PMLR, 10–15 Jul 2018.

E. Ben-David, T. Li, H. Massam, and B. Rajaratnam. High Dimensional Bayesian Inference for Gaussian Directed Acyclic Graph Models. *arXiv preprint*, 2015.

M. Bonvini, E. H. Kennedy, V. Ventura, and L. Wasserman. Causal Inference for the Effect of Mobility on COVID-19 Deaths. *The Annals of Applied Statistics*, 16(4):2458–2480, 2022.

F. Castelletti and G. Consonni. Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. *The Annals of Applied Statistics*, 13(4):2289–2311, 2019.

F. Castelletti and A. Mascaro. Structural Learning and Estimation of Joint Causal Effects among Network-dependent Variables. *Statistical Methods & Applications*, 30:1289–1314, 2021.

F. Castelletti and A. Mascaro. BCDAG: An R package for Bayesian structure and Causal learning of Gaussian DAGs. *arXiv preprint*, 2022.

F. Castelletti and A. Mascaro. Bayesian Causal Discovery from Unknown General Interventions. *Submitted*, 2023+.

F. Castelletti and S. Peluso. Network Structure Learning Under Uncertain Interventions. *Journal of the American Statistical Association*, 118(543):2117–2128, 2023.

F. Castelletti, G. Consonni, M. L. D. Vedova, and S. Peluso. Learning Markov Equivalence Classes of Directed Acyclic Graphs: An Objective Bayes Approach. *Bayesian Analysis*, 13(4):1235–1260, 2018.

D. M. Chickering. Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3:507–554, 2002.

G. F. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9(4):309–347, 1992.

G. F. Cooper and C. Yoo. Causal Discovery from a Mixture of Experimental and Observational Data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 116–125, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

J. Correa and E. Bareinboim. A Calculus For Stochastic Interventions: Causal Effect Identification and Surrogate Experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 107–114, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

R. A. Fisher. *The Design of Experiments*. Hafner, 1935.

N. Friedman and D. Koller. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, 50(1):95–125, 2003.

J. L. Gamella, A. Taeb, C. Heinze-Deml, and P. Bühlmann. Characterization and Greedy Learning of Gaussian Structural Causal Models under Unknown Interventions, 2022.

S. B. Gillispie and M. D. Perlman. Enumerating Markov equivalence classes of acyclic digraph Models. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 171–177, 2001.

N. Harris and M. Drton. PC Algorithm for Nonparanormal Graphical Models. *Journal of Machine Learning Research*, 14(69):3365–3383, 2013.

A. Hauser and P. Bühlmann. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *Journal of Machine Learning Research*, 13(79):2409–2464, 2012.

C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen. Causal Structure Learning. *Annual Review of Statistics and Its Application*, 5(1):371–391, 2018.

K. B. Korb, L. R. Hope, A. E. Nicholson, and K. Axnick. Varieties of Causal Intervention. In C. Zhang, H. W. Guesgen, and W.-K. Yeap, editors, *PRICAI 2004: Trends in Artificial Intelligence*, pages 322–331, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

J. Kuipers and G. Moffa. Partition MCMC for Inference on Acyclic Digraphs. *Journal of the American Statistical Association*, 112(517):282–299, 2017.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

P. Nowack, J. Runge, V. Eyring, and J. D. Haigh. Causal Networks for Climate Model Evaluation and Constrained Projections. *Nature Communications*, 11(1):1415, 2020.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

J. Pearl. *Causality*. Cambridge University Press, 2009.

S. Peluso and G. Consonni. Compatible Priors for Model Selection of High-dimensional Gaussian DAGs. *Electronic Journal of Statistics*, 14(2):4110–4132, 2020.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

A. Roverato and G. Consonni. Compatible Prior Distributions for Directed Acyclic Graph Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1): 47–61, 12 2003.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A Linear non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7 (72):2003–2030, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 2001.

C. Squires and C. Uhler. Causal Structure Learning: A Combinatorial Perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

C. Squires, Y. Wang, and C. Uhler. Permutation-Based Causal Structure Learning with Unknown Intervention Targets. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pages 1039–1048. PMLR, 03–06 Aug 2020.

J. Tian and J. Pearl. Causal Discovery from Changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 512–521, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65(1):31–78, 2006.

T. Verma and J. Pearl. Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI 90, pages 255–270, New York, NY, USA, 1990. Elsevier Science Inc.

Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation-based Causal Inference Algorithms with Interventions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5822–5831. Curran Associates, Inc., 2017.

K. Yang, A. Katcoff, and C. Uhler. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5541–5550. PMLR, 10–15 Jul 2018.

Chapter 2

Bayesian causal discovery and joint causal effect estimation

2.1 Introduction

Graphical models based on Directed Acyclic Graphs are commonly used to model dependence relationships among variables. When considered as graphical representations of Markovian Structural Causal Models (SCMs), DAGs can be given a causal interpretation, with each node corresponding to an element of the random vector $X \coloneqq (X_1, \dots, X_q)$ and each edge a direct causal influence among these. In this setting, the *do-calculus* [Pearl, 2009] can be used to define, identify and estimate causal effects given only observation data. The do-operator at the base of the do-calculus, denoted as $\text{do}(X_j = \tilde{x}_j)$, represents the action of fixing the value of the random variable X_j to $\tilde{x}_j$. Accordingly, a causal effect can be informally defined as the expected change in an outcome variable $Y \in X$ induced by a unit change of the value at which we fix X_j. The do-calculus consists of a set of rules that, given a DAG, allow the identification of causal effects, even when the corresponding experimental (post-intervention) data are not available.

We consider the problem of identifying and estimating causal effects using observational Gaussian data when the DAG is not known. In particular, we focus on the case of causal effects associated with *joint* interventions on a set of target variables I and which we denote as $\text{do}\{X_j = \tilde{x}_j\}_{j \in I}$. Such causal effects differ from their single-variable counterpart because of the possible interactions that may occur and that depend on the DAG structure. Consider, for instance, the DAG reported in Figure 2.1, where each edge weight corresponds to a randomly generated coefficient of a linear Structural Equation Model (SEM) [Bollen, 1989]. Suppose we are interested in evaluating the causal effect of X_6 on X_1 in i) a single intervention on X_6, and ii) a joint intervention on X_6 and X_4. From the DAG and associated coefficients, one may use the *path method* [Wright, 1934] to identify such causal effects. In

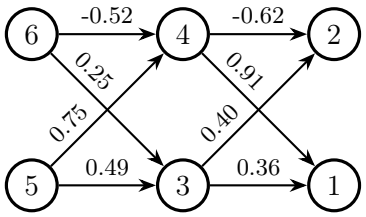


Figure 2.1: *A DAG with $q = 6$ node/variables and randomly generated edge weights.*

the path method, assuming unit node-variance, the first step is to identify all the possible paths leading from node 6 to node 1. Then, a path coefficient is produced by multiplying all the coefficients associated with each edge in the path. Finally, all the path coefficients are summed up to compute the causal effect of interest. In our example, the path method reflects the idea that changing the value at which we fix X_6 would result in a change in node X_4 and X_3 and, in turn, in node X_1. In case i), by the path method, the causal effect equals -0.3832. In case ii), instead, one has to take into account that X_4 is itself a target of intervention, and the dependence between X_6 and X_4 has been destroyed by such intervention. As a consequence, the only "active" path is the one passing through X_3 and the causal effect associated with X_6 in this joint intervention on X_6, X_4 is 0.08.

As the DAG is unknown, it must be estimated from data. When DAGs are given a causal interpretation, the process of learning the DAG structure from data is referred to as *causal discovery* [Spirtes et al., 2001, Peters et al., 2017]. Every DAG model associated with a Markovian SCM encodes a set of conditional independencies that can be read-off from the DAG using a criterion called d-separation [Pearl, 2009]. Under faithfulness, these conditional independencies are exactly those entailed by the joint distribution over X, and it is possible to use them to learn the DAG structure. However, different DAGs may encode the same set of conditional independencies rendering them indistinguishable when relying solely on observational data. Such DAGs are commonly referred to as *Markov equivalent* [Verma and Pearl, 1990]. Correctly accounting for this identifiability limit is of key importance in causal discovery. For instance, in the Bayesian setting, the existence of indistinguishable structure translates into the *compatibility* requirement [Roverato and Consonni, 2003] that Markov equivalent DAGs are assigned equal marginal likelihood, a property which is referred to as *score-equivalence* in the literature on score-based methods for causal discovery [Chickering, 2002].

In this chapter, we present a Bayesian methodology that combines causal discovery and joint causal effect estimation from observational Gaussian data. In particualar, we propose a unified approach that leads to a joint posterior distribution over DAGs, DAG-parameters

and joint causal effects. Our model specification satisfies the compatibility requirement of assigning equal marginal likelihood to Markov equivalent DAGs. In addition, it allows for fast computation of the Bayes factors involved in the MCMC scheme used to sample from the ensued joint posterior distribution.

2.1.1 Related work

The first method dealing with the identification and estimation of single-variable causal effects when the DAG is unknown is the IDA (Identification when the DAG is Absent) method [Maathuis et al., 2009]. IDA consists of a two-step procedure. In the first step, a Completed Partially Directed Acyclic Graph (CPDAG) representing a Markov equivalence class is inferred from data using any causal discovery procedure such as the PC algorithm [Spirtes et al., 2001] or the GES algorithm [Chickering, 2002]. In the second step, all causal effects compatible with the estimated Markov equivalence class of DAGs are enumerated with a fast procedure that only requires local information on the CPDAG. The authors prove that, conditionally on the true CPDAG, their method provides consistent estimates of causal effects associated with interventions on single variables. The same method was then extended to the case of joint interventions by Nandy et al. [2017], who also proved the consistency of the procedure in the new setting. More recently, Perković et al. [2018] showed how any joint causal effect of interest can be identified from a DAG, a CPDAG or a PDAG via covariate adjustment, i.e. via a regression of the outcome node on the treatment nodes and a *valid adjustment sets*. However, all these methods rely on a single estimated Markov equivalence class of DAGs. Consequently, results are highly sensitive to this estimate. Bayesian methods, on the other hand, rely on a unified approach which takes fully into account both the uncertainty over the DAG structure and over the parameter estimation. A Bayesian method combining causal discovery and single-variable causal effect estimation from observational Gaussian data was provided by Castelletti and Consonni [2021a]. Their methodology was also extended to the case of Gaussian data with a binary outcome variable [Castelletti and Consonni, 2021b] and to the case of heterogeneous Gaussian data [Castelletti and Consonni, 2023]. Viinikka et al. [2020] also provided a similar method but adopting a different MCMC scheme with better convergence properties.

2.1.2 Outline

In Section 2.2, we provide background and notation on Gaussian DAG models and we show how, through the do-calculus, it is possible to identify the joint causal effects as DAG-specific functions of the elements of a modified covariance matrix. In Section 2.3, we

introduce our Bayesian model specification, leading to a joint posterior distribution over DAGs, DAG-parameters and joint causal effects. We discuss in Section 2.4 computational details leading to our MCMC scheme for posterior inference. In Section 2.5, we apply our methodology to simulated and real data. Finally, Section 2.6 offers a brief discussion together with possible future developments.

2.2 Preliminaries

In this section, we formally introduce the necessary background on graphical models and causal inference. In Section 2.2.1, we provide an overview of Gaussian DAG-models. In Section 2.2.2, we show how single-variable causal effects can be defined and identified from Gaussian observational data using the do-operator. In Section 2.2.3, we instead focus on the case of joint causal effects.

2.2.1 Gaussian DAG-models

We briefly introduce the graph notation hereinafter adopted. Let $\mathcal{G} = (V, E)$ be a graph, where $V = \{1, \dots, q\} \coloneqq [q]$ is a set of nodes (or vertices) and $E \subseteq V \times V$ a set of edges. In what follows, if $(u, v) \in E$ and $(v, u) \notin E$, $\mathcal{G}$ contains a directed edge $u \to v$, while if both $(u, v) \in E$ and $(v, u) \in E$, then $\mathcal{G}$ contains an undirected edge $u - v$. A graph is called directed if contains only directed edges. Moreover, a Directed Acyclic Graph (DAG) $\mathcal{D}$ is a directed graph which contains no loops, that is sequences of nodes $(u_1, u_2, \dots, u_k)$ with $u_1 = u_k$, such that there exists a path $u_1 \to u_2 \to \cdots \to u_k$. Moreover, if $(u, v) \in E$ we say that u is a parent of v and denote the set of all parents of v in $\mathcal{D}$ as $\mathrm{pa}_{\mathcal{D}}(v)$. Also, if there exists a directed path from u to v we say that v is a descendant of u and let $\mathrm{de}_{\mathcal{D}}(u)$ be the set of all descendants of u in $\mathcal{D}$. Hence, the non-descendants of u are all nodes in the set $\mathrm{nd}_{\mathcal{D}}(u) = V \setminus \mathrm{de}_{\mathcal{D}}(u)$.

Let $X \coloneqq (X_1, \dots, X_q)$ be a random vector. DAGs are given a causal interpretation when considered as graphical representations of a *Structural Causal Model* (SCM) over X, each node being a variable and each edge a direct causal relationship. In particular, in this chapter, we will consider linear Gaussian Structural Equation Models (SEM) of the form

$$X = \boldsymbol{B}^T X + \epsilon, \quad \epsilon \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{D}) \tag{2.1}$$

where $\boldsymbol{B}$ is a (q, q) matrix of regression coefficients with (u, v)-element $\boldsymbol{B}_{uv} \neq 0$ if and only if $u \in \mathrm{pa}_{\mathcal{D}}(v)$, and $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{D}_{11}, \dots, \boldsymbol{D}_{qq})$ is a (q, q) matrix collecting the variances of the

q *exogenous* variables in ϵ. Equation (2.1) implies

$$X \mid \boldsymbol{\Sigma}, \mathcal{D} \sim \mathcal{N}_q\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{D}}\right), \tag{2.2}$$

with $\boldsymbol{\Sigma}_{\mathcal{D}} = (\boldsymbol{I} - \boldsymbol{B})^{-\top}\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{B})^{-1}$, the right-hand side corresponding to the modified Cholesky decomposition of the covariance matrix. The independence of the elements in ϵ makes the SEM in (2.1) Markovian. As a consequence, the induced joint distribution over X satisfies the Markov property of $\mathcal{D}$, meaning that it is possible to factorize it as

$$p(x \mid \boldsymbol{\Sigma}_{\mathcal{D}}, \mathcal{D}) = \prod_{j=1}^{q} p\left(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}, \boldsymbol{\Sigma}_{\mathcal{D}}, \mathcal{D}\right), \tag{2.3}$$

where we denote with x any realization of the random vector X. All the conditional independencies implied by (2.3) can be read off from $\mathcal{D}$ through d-separation. In short, we have that every node is independent of its non-descendants given its parents. When we consider the modified Cholesky decomposition of $\boldsymbol{\Sigma}_{\mathcal{D}}$, it is possible to further specify (2.3) as

$$p(x \mid (\boldsymbol{B}, \boldsymbol{D}), \mathcal{D}) = \prod_{j=1}^{q} d\mathcal{N}\left(x_j \mid x^T_{\mathrm{pa}_{\mathcal{D}}(j)}\boldsymbol{B}_{\mathrm{pa}_{\mathcal{D}}(j)\times j}, \boldsymbol{D}_{jj}\right), \tag{2.4}$$

where by $d\mathcal{N}(x_j \,|\, \mu_0, \sigma_0^2)$ we denote the density corresponding to a Gaussian random variable X_j with mean μ_0 and variance σ_0^2. We refer to any Gaussian model (2.2) whose covariance matrix $\boldsymbol{\Sigma}$ is Markov w.r.t. a DAG $\mathcal{D}$ as a Gaussian DAG-model.

2.2.2 Identifying causal effects in Gaussian DAG-models

We now focus on the identification of the causal effect of a single treatment variable X_i on an outcome variable $Y \coloneqq X_1$ in the context of Gaussian DAG-models. In the language of the *do-calculus*, we can define such causal effect as:

$$\gamma_{x_i y} := \frac{\partial}{\partial \tilde{x}_i}\mathbb{E}\left(Y \mid \mathrm{do}\left(X_i = \tilde{x}_i\right)\right). \tag{2.5}$$

In other words, the causal effect is the induced variation in the expected value of Y resulting from an infinitesimal change in the value $\tilde{x}_i$ at which we fix the treatment variable X_i. To identify the so-defined causal effect, we first need to derive the post-intervention distribution of Y given an intervention fixing the value of X_i. In the Gaussian case, the joint distribution

over $X \backslash X_i$ becomes

$$p\left(x \mid \operatorname{do}\left(\tilde{x}_i\right), (\boldsymbol{B}, \boldsymbol{D}), \mathcal{D}\right) = \prod_{j \neq i} d\mathcal{N}\left(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}^T \boldsymbol{B}_{\mathrm{pa}_{\mathcal{D}}(j) \times j}, \boldsymbol{D}_{jj}\right) \mathbb{1}\{x_i = \tilde{x}_i\}, \tag{2.6}$$

where $\operatorname{do}\left(\tilde{x}_i\right)$ is a shorter notation for $\operatorname{do}\left(X_i = \tilde{x}_i\right)$ that we hereinafter adopt. The marginal distribution of $Y = X_1$ can be obtained via marginalization. In general, if $Y \notin \mathrm{pa}_{\mathcal{D}}(i)$, we have:

$$p\left(y \mid \operatorname{do}\left(\tilde{x}_i\right), \boldsymbol{\Sigma}_{\mathcal{D}}\right) = \int p\left(y \mid \tilde{x}_i, x_{\mathrm{pa}_{\mathcal{D}}(i)}, \boldsymbol{\Sigma}_{\mathcal{D}}\right) p\left(x_{\mathrm{pa}_{\mathcal{D}}(i)} \mid \boldsymbol{\Sigma}_{\mathcal{D}}\right) \, dx_{\mathrm{pa}_{\mathcal{D}}(i)}. \tag{2.7}$$

Accordingly,

$$\mathbb{E}\left(Y \mid \operatorname{do}\left(\tilde{x}_i\right), \boldsymbol{\Sigma}_{\mathcal{D}}\right) = \int \mathbb{E}\left(Y \mid \tilde{x}_i, x_{\mathrm{pa}_{\mathcal{D}}(i)}, \boldsymbol{\Sigma}_{\mathcal{D}}\right) p\left(x_{\mathrm{pa}_{\mathcal{D}}(i)} \mid \boldsymbol{\Sigma}_{\mathcal{D}}\right) \, dx_{\mathrm{pa}_{\mathcal{D}}(i)}. \tag{2.8}$$

By Gaussianity, $\mathbb{E}(Y \mid \tilde{x}_i, x_{\mathrm{pa}_{\mathcal{D}}(i)}, \boldsymbol{\Sigma}_{\mathcal{D}})$ is linear in $\tilde{x}_i, x_{\mathrm{pa}_{\mathcal{D}}(i)}$, and it can be written as

$$\mathbb{E}(Y \mid \tilde{x}_i, x_{\mathrm{pa}_{\mathcal{D}}(i)}, \boldsymbol{\Sigma}_{\mathcal{D}}) = \gamma_0 + \gamma_{x_i y} \tilde{x}_i + \gamma_{\mathrm{pa}_{\mathcal{D}}(i)}^T x_{\mathrm{pa}_{\mathcal{D}}(i)}, \tag{2.9}$$

for some values $\gamma_0, \gamma_i \in \mathbb{R}$ and $\gamma_{\mathrm{pa}_{\mathcal{D}}(i)} \in \mathbb{R}^{|pa_i|}$, where $|\mathrm{pa}_{\mathcal{D}}(i)|$ is the cardinality of the parent set of X_i in $\mathcal{D}$. Substituting Equation (2.9) in Equation (2.8), we thus obtain

$$\mathbb{E}\left(Y \mid \operatorname{do}\left(\tilde{x}_i\right), \boldsymbol{\Sigma}_{\mathcal{D}}\right) = \gamma_{x_i y} \tilde{x}_i + \int \gamma_{\mathrm{pa}_{\mathcal{D}}(i)}^T x_{\mathrm{pa}_{\mathcal{D}}(i)} p\left(x_{\mathrm{pa}_{\mathcal{D}}(i)} \mid \boldsymbol{\Sigma}_{\mathcal{D}}\right) \, dx_{\mathrm{pa}_{\mathcal{D}}(i)}, \tag{2.10}$$

which is clearly linear in $\gamma_{x_i y}$. By the definition given by Equation 2.5, $\gamma_{x_i y}$ corresponds to the causal effect of X_i on Y. Notice that $\gamma_{x_i y}$ is just the regression coefficient associated with X_i in a regression of Y on $(X_i, X_{\mathrm{pa}_{\mathcal{D}}(i)})$. When a causal effect is identified as a regression coefficient in a particular regression, we say that it is identified by *covariate adjustment* and $(X_i, X_{\mathrm{pa}_{\mathcal{D}}(i)})$ is referred to as a *valid adjustment set* [Perković et al., 2018]. The valid adjustment set may not be unique. As a consequence, a recent body of literature has focused on the definition of the *best* valid adjustment set for a given pair of treatment and outcome node [Henckel et al., 2022]. However, covariate adjustment is not the only possible identification strategy. In the next section, we show an alternative strategy for the case of joint causal effects.

2.2.3 Identifying joint causal effects

We now consider the identification of the causal effect on an outcome variable Y of a joint (simultaneous) intervention on more than one variable. We denote with $I \subset \{2, \dots, q\}$ the set of *intervention targets*. Similarly to to the single variable case, we may define such causal effect using the language of the do-calculus. In particular

$$\boldsymbol{\gamma}_y^I \coloneqq (\gamma_{x_h y}^I)_{h \in I}^T, \tag{2.11}$$

where, for each $h \in I$

$$\gamma_{x_h y}^I \coloneqq \frac{\partial}{\partial x_h} \mathbb{E}\left(Y \mid \mathrm{do}\{\tilde{x}_j\}_{j \in I}\right). \tag{2.12}$$

In other words, the causal effect of X_h on Y in a joint intervention on $\{X_j\}_{j \in I}$ is the induced variation in the expected value of Y resulting from an infinitesimal change in the value $\tilde{x}_h$ at which we fix the variable X_j, keeping all the other target variables fixed. The post joint-intervention distribution of X in the Gaussian case can be immediately written as

$$p\left(x \mid \mathrm{do}\left\{\tilde{x}_h\right\}_{h \in I}, \boldsymbol{\Sigma}_{\mathcal{D}}\right) = \prod_{j \notin I} d\mathcal{N}\left(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}^T \boldsymbol{B}_{\mathrm{pa}_{\mathcal{D}}(j) \times j}, \boldsymbol{D}_{jj}\right) \mathbb{1}\{x_h = \tilde{x}_h\}_{h \in I}. \tag{2.13}$$

A hard intervention corresponds to fixing the value of the target variable to a constant and, as a consequence, it destroys the dependence relation of the intervened node with its parents. We thus consider the following modified SEM:

$$X = (\boldsymbol{B}^I)^T X + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}) \tag{2.14}$$

where $\boldsymbol{B}^I$ corresponds to the matrix of regression coefficients that we would observe if we were observing X after an intervention on X_I, that is:

$$\boldsymbol{B}_{uv}^I = \begin{cases} 0 & \text{if } v \in I \text{ and } v \neq u \\ \boldsymbol{B}_{uv} & \text{otherwise.} \end{cases} \tag{2.15}$$

The SEM of Equation 2.14 induces a modified Gaussian DAG-model over X,

$$X \sim \mathcal{N}_q\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{D}^I}\right), \tag{2.16}$$

where $\boldsymbol{\Sigma}_{\mathcal{D}^I} = (\boldsymbol{I} - \boldsymbol{B}^I)^{-T} \boldsymbol{D} (\boldsymbol{I} - \boldsymbol{B}^I)^{-1}$ and $\mathcal{D}^I$ denotes the post-intervention DAG obtained by removing all the edges pointing into nodes in I. The modified DAG-model (2.16) is

related to the post-intervention distribution (2.13). In particular, as for any $j \notin I$, $\mathrm{pa}_{\mathcal{D}}(j) = \mathrm{pa}_{\mathcal{D}^I}(j)$ and $\boldsymbol{B}_{\mathrm{pa}_{\mathcal{D}}(j)\times j} = \boldsymbol{B}^I_{\mathrm{pa}_{\mathcal{D}^I}(j)\times j}$, we have

$$p\left(x \mid \{\tilde{x}_h\}_{h\in I}, \boldsymbol{\Sigma}_{\mathcal{D}^I}\right) = \prod_{j\notin I} d\mathcal{N}\left(x_j \mid x^T_{\mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{B}_{\mathrm{pa}_{\mathcal{D}}(j)\times j}, \boldsymbol{D}_{jj}\right) \mathbb{1}\{x_h = \tilde{x}_h\}_{h\in I}. \tag{2.17}$$

The conditional distribution of X given $X_h = \tilde{x}_h$ in the modified DAG-model thus corresponds to the post-intervention distribution (2.13) of the original DAG-model. We can thus consider (2.17) as the post-intervention distribution on which our causal effect of interest is defined, that is

$$p\left(y \mid \mathrm{do}\left\{\tilde{x}_h\right\}_{h\in I}, \boldsymbol{\Sigma}_{\mathcal{D}}\right) = p\left(y \mid \{\tilde{x}_h\}_{h\in I}, \boldsymbol{\Sigma}_{\mathcal{D}^I}\right). \tag{2.18}$$

Taking the expectation on both sides, we obtain

$$\mathbb{E}\left(Y \mid \mathrm{do}\left\{\tilde{x}_h\right\}_{h\in I}, \boldsymbol{\Sigma}_{\mathcal{D}}\right) = \mathbb{E}\left(Y \mid \{\tilde{x}_h\}_{h\in I}, \boldsymbol{\Sigma}_{\mathcal{D}^I}\right). \tag{2.19}$$

By Gaussianity, the right-hand side of (2.19) is linear in $\{\tilde{x}_h\}_{h\in I}$ and can be written as

$$\mathbb{E}\left(Y \mid \{X_h = \tilde{x}_h\}_{h\in I}, \boldsymbol{\Sigma}_{\mathcal{D}^I}\right) = \gamma_0 + \sum_{h\in I} \gamma^I_{x_h y} x_h. \tag{2.20}$$

The vector $\boldsymbol{\gamma}^I_y$ thus corresponds to the regression coefficients associated with the target variables in the regression of Y on $\{X_h\}_{h\in I}$ in the modified DAG model. Such regression coefficients can be computed directly from the modified covariance matrix:

$$\gamma^I_{x_h y} = \frac{(\boldsymbol{\Sigma}_{\mathcal{D}^I})_{h1}}{(\boldsymbol{\Sigma}_{\mathcal{D}^I})_{hh}} \quad \text{for } h \in I. \tag{2.21}$$

It follows that the causal effect $\gamma^I_{x_h y}$ is a function of the covariance matrix $\boldsymbol{\Sigma}_{\mathcal{D}}$ which in turn depends on the underlying DAG $\mathcal{D}$. Therefore, inference on DAG $\mathcal{D}$ and its parameter $\boldsymbol{\Sigma}$ will drive inference of causal effects under model uncertainty; see the next section for details.

2.3 Bayesian inference on causal effects under model uncertainty

In this section, we provide a Bayesian method to estimate $\boldsymbol{\gamma}^I_y$ when the data are generated according to an unknown DAG-model. In Section 2.3.1, we formulate the problem in the Bayesian setting. In Section 2.3.2 we detail the prior specification on the parameters of

the DAG-model, illustrating the ensued marginal likelihood. Finally, in Section 2.3.3, we illustrate our prior specification on the DAG space.

2.3.1 Model formulation

Let $\boldsymbol{X}$ be a (n, q) matrix containing i.i.d. measurements of q variables generated from an unknown Gaussian DAG-model (2.2). Our goal is to derive a posterior distribution over the causal effect $\boldsymbol{\gamma}_y^I$ on $Y = X_1$ from a joint (simultaneous) intervention on $I \subset \{2, \dots, q\}$. As shown in Section 2.11, in Gaussian DAG models $\boldsymbol{\gamma}_y^I$ is identified as a function of a subset of the elements of the covariance matrix $\boldsymbol{\Sigma}_{\mathcal{D}}$. Moreover, the covariance matrix can be written in terms of the modified Cholesky decomposition of its inverse, the precision matrix $\boldsymbol{\Omega}_{\mathcal{D}}$:

$$\boldsymbol{\Omega}_{\mathcal{D}} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^T, \tag{2.22}$$

where $\boldsymbol{L} = (\boldsymbol{I} - \boldsymbol{B})$ and $\boldsymbol{L}_{ij} \neq 0$ if $(i, j) \in E$ or $i = j$. Consequently, our primary interest is in deriving the posterior distribution over $(\boldsymbol{L}, \boldsymbol{D})$ and $\mathcal{D}$

$$p\left((\boldsymbol{L}, \boldsymbol{D}), \mathcal{D} \mid \boldsymbol{X}\right) \propto p\left(\boldsymbol{X} \mid (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}\right) p\left((\boldsymbol{L}, \boldsymbol{D}) \mid \mathcal{D}\right) p(\mathcal{D}), \tag{2.23}$$

from which samples from the posterior distribution over $\boldsymbol{\gamma}_y^I$ can be then be obtained.

The likelihood component in (2.23) can be written as

$$p\left(\boldsymbol{X} \mid (\boldsymbol{L}, \boldsymbol{D}), \mathcal{D}\right) = \prod_{j=1}^{q} d\mathcal{N}_n\left(\boldsymbol{X}_j \mid -\boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}\boldsymbol{L}_{\mathrm{pa}_{\mathcal{D}}(j)\times j}, \boldsymbol{D}_{jj}\boldsymbol{I}_n\right), \tag{2.24}$$

where $\boldsymbol{X}_A$ denotes the sub-matrix of $\boldsymbol{X}$ with columns indexed by $A \subseteq V$. Assigning a prior distribution on DAG $\mathcal{D}$ and its associated Cholesky parameters $(\boldsymbol{L}, \boldsymbol{D})$ requires particular attention and it is the object of the next two sections.

2.3.2 Prior on DAG parameters

Conditionally on $\mathcal{D}$, we assign a *DAG-Wishart* distribution [Ben-David et al., 2015] as a prior on $(\boldsymbol{L}, \boldsymbol{D})$. Let $\mathcal{D}_+^q$ be the set of (q, q) positive matrices with unit main diagonal and $\mathcal{L}_{\mathcal{D}}$ the set of (q, q) matrices with unit main diagonal and whose ij-th entry is non-zero only if $(i, j) \in E$. We call $\Theta_{\mathcal{D}} = \mathcal{D}_+^q \times \mathcal{L}_{\mathcal{D}}$ the Cholesky space associated with the DAG $\mathcal{D}$. The DAG-Wishart distribution $\boldsymbol{\pi}_{\boldsymbol{\alpha},\boldsymbol{U}}^{\Theta_{\mathcal{D}}}$ on $\Theta_{\mathcal{D}}$, with rate hyperparameter $\boldsymbol{U}$ (a (q, q) s.p.d.

matrix) and shape hyperparameter $\boldsymbol{a}(\mathcal{D}) \coloneqq (a_1(\mathcal{D}), \ldots, a_q(\mathcal{D}))$ has density

$$\pi^{\Theta_{\mathcal{D}}}_{\boldsymbol{a},\boldsymbol{U}}(\boldsymbol{L},\boldsymbol{D}) = \frac{1}{\mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{a}(\mathcal{D}),\boldsymbol{U}\right)} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}^{\top}\right)U\right)\right\} \prod_{j=1}^{q} \boldsymbol{D}_{jj}^{-\frac{a_j(\mathcal{D})}{2}}, \tag{2.25}$$

for all $(\boldsymbol{L},\boldsymbol{D}) \in \Theta_{\mathcal{D}}$. $\mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{a}(\mathcal{D}),\boldsymbol{U}\right)$ denotes the normalizing constant, which is finite if $a_j(\mathcal{D}) - |\mathrm{pa}_{\mathcal{D}}(j)| > 2$ for all $j \in [q]$ and can be written as

$$\mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{a}(\mathcal{D}),\boldsymbol{U}\right) = \prod_{j=1}^{q} \Gamma\left(\frac{c_j(\mathcal{D})}{2} - 1\right) 2^{\frac{a_j}{2}-1} (\sqrt{\pi})^{|\mathrm{pa}_{\mathcal{D}}(j)|} \frac{\left|\boldsymbol{U}_{\mathrm{pa}_{\mathcal{D}}(j)}\right|^{\frac{c_j(\mathcal{D})-3}{2}}}{\left|\boldsymbol{U}_{\mathrm{fa}_{\mathcal{D}}(j)}\right|^{\frac{c_j(\mathcal{D})-2}{2}}}, \tag{2.26}$$

where $c_j(\mathcal{D}) = a_j(\mathcal{D}) - |\mathrm{pa}_{\mathcal{D}}(j)|$ for all $j \in [q]$.

The DAG-Wishart distribution presents many useful features, the first one being that it induces local distributions on the non-null elements of $(\boldsymbol{L},\boldsymbol{D})$ that are node-wise independent and such that

$$\boldsymbol{D}_{jj} \mid \mathcal{D} \sim \text{I-Ga}\left(\frac{a_j(\mathcal{D}) - |pa_{\mathcal{D}}(j)|}{2} - 1, \frac{1}{2}\boldsymbol{U}_{j|\mathrm{pa}_{\mathcal{D}}(j)}\right), \tag{2.27}$$

$$\boldsymbol{L}_{\mathrm{pa}_{\mathcal{D}}(j)\times j} \mid \boldsymbol{D}_{jj}, \mathcal{D} \sim \mathcal{N}_{|\mathrm{pa}_{\mathcal{D}}(j)|}\left(-\boldsymbol{U}^{-1}_{\mathrm{pa}_{\mathcal{D}}(j)}\boldsymbol{U}_{\mathrm{pa}_{\mathcal{D}}(j)\times j}, \boldsymbol{D}_{jj}\boldsymbol{U}^{-1}_{\mathrm{pa}_{\mathcal{D}}(j)}\right), \tag{2.28}$$

where $\boldsymbol{U}_{\mathrm{pa}_{\mathcal{D}}(j)} \coloneqq \boldsymbol{U}_{\mathrm{pa}_{\mathcal{D}}(j)\times\mathrm{pa}_{\mathcal{D}}(j)}$ and $\boldsymbol{U}_{j|\mathrm{pa}_{\mathcal{D}}(j)} \coloneqq \boldsymbol{U}_{jj} - \boldsymbol{U}_{j\times\mathrm{pa}_{\mathcal{D}(j)}}\boldsymbol{U}^{-1}_{\mathrm{pa}_{\mathcal{D}}(j)}\boldsymbol{U}_{\mathrm{pa}_{\mathcal{D}}(j)\times j}$.

In addition, the DAG-Wishart distribution is conjugate to the Normal likelihood, meaning that given a matrix $\boldsymbol{X}$ containing n i.i.d. observations from a Gaussian distribution Markov w.r.t. a DAG $\mathcal{D}$, $(\boldsymbol{L},\boldsymbol{D})|\mathcal{D} \sim \boldsymbol{\pi}^{\Theta_{\mathcal{D}}}_{\boldsymbol{a},\boldsymbol{U}}$ implies $(\boldsymbol{L},\boldsymbol{D})|\boldsymbol{X},\mathcal{D} \sim \boldsymbol{\pi}^{\Theta_{\mathcal{D}}}_{\widetilde{\boldsymbol{a}},\widetilde{\boldsymbol{U}}}$, where $\widetilde{\boldsymbol{a}} = \boldsymbol{a} + n$ and $\widetilde{\boldsymbol{U}} = \boldsymbol{U} + \boldsymbol{X}^T\boldsymbol{X}$. We thus also have a closed-form expression for the marginal likelihood of a DAG-model:

$$p(\boldsymbol{X} \mid \mathcal{D}) = (2\pi)^{-(nq)/2} \mathcal{Z}_{\mathcal{D}}\left(\widetilde{\boldsymbol{a}}(\mathcal{D}), \widetilde{\boldsymbol{U}}\right) / \mathcal{Z}_{\mathcal{D}}\left(\boldsymbol{a}(\mathcal{D}),\boldsymbol{U}\right). \tag{2.29}$$

Consequently, it is possible to sample from the posterior distribution over the DAG space using a simple Metropolis-Hastings scheme. Moreover, the marginal likelihood (2.29) is decomposable, meaning that it is a product of q terms corresponding to each parent-child relationship in the DAG. Substituting Equation (2.26) into (2.29), it is immediate to show

that

$$p(\boldsymbol{X} \mid \boldsymbol{D}) = \prod_{j=1}^{q} (2\pi)^{-n/2} \frac{\Gamma\left(\frac{\tilde{c}_j(\mathcal{D})}{2} - 1\right)}{\Gamma\left(\frac{c_j(\mathcal{D})}{2} - 1\right)} \frac{\left|\boldsymbol{U}_{\mathrm{pa}_{\mathcal{D}}(j)}\right|^{1/2}}{\left|\widetilde{\boldsymbol{U}}_{\mathrm{pa}_{\mathcal{D}}(j)}\right|^{1/2}} \frac{\left(\frac{1}{2}\boldsymbol{U}_{j|\mathrm{pa}_{\mathcal{D}}(j)}\right)^{\frac{c_j(\mathcal{D})}{2}-1}}{\left(\frac{1}{2}\widetilde{\boldsymbol{U}}_{j|\mathrm{pa}_{\mathcal{D}}(j)}\right)^{\frac{\tilde{c}_j(\mathcal{D})}{2}-1}}, \tag{2.30}$$

where $\tilde{c}_j = c_j + n$ and, for each $j \in [q]$, each term in the factorization (2.30) corresponds to the conditional marginal likelihood $p(\boldsymbol{X}_j | \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(j)}, \mathcal{D})$. The decomposability property dramatically reduces the computational cost of MCMC schemes that use local moves to explore the DAG space, as the one that we detail in Section 2.4. In addition, sampling from the posterior over DAG-parameters $p\left((\boldsymbol{L}, \boldsymbol{D}) | \boldsymbol{X}, \mathcal{D}\right)$ only requires direct sampling.

Finally, the DAG-Wishart distribution satisfies the assumptions of the DAG-parameter prior construction procedure of Geiger and Heckerman [2002]. It is thus possible to construct a *compatible* version of it that assigns equal marginal likelihood to Markov equivalent DAGs. The compatible DAG-Wishart distribution $\boldsymbol{\pi}^{c,\Theta_{\mathcal{D}}}_{\boldsymbol{a}^c(\mathcal{D}),\boldsymbol{U}}$, firstly derived in Peluso and Consonni [2020], differs from its non-compatible counterpart in the choice of the rate hyperparameter, as for each node $j \in [q]$:

$$\boldsymbol{a}^c_j(\mathcal{D}) = a - q + 2|\mathrm{pa}_{\mathcal{D}}(j)| + 3 \tag{2.31}$$

As a consequence, the compatible DAG-Wishart distribution has only one free rate- hyperparameter a, which must satisfy $a > q - 1$ to guarantee that the prior is proper.

2.3.3 Prior on DAG $\mathcal{D}$

To complete our Bayesian model specification, we finally assign a prior to each DAG $\mathcal{D} \in \mathcal{S}_q$, the set of all DAGs on q nodes. For a given DAG $\mathcal{D} = (V, E)$, let $S^{\mathcal{D}}$ be the adjacency matrix of its skeleton (the underlying undirected graph obtained after removing the orientation of all of its edges), such that for each (u, v)-element in $S^{\mathcal{D}}$, $S^{\mathcal{D}}_{u,v} = 1$ if and only if $(u, v) \in E$ or $(v, u) \in E, 0$ otherwise. For a given probability of edge inclusion $\omega \in (0, 1)$ we then assume $S^{\mathcal{D}}_{u,v} \overset{\text{iid}}{\sim} \mathrm{Ber}(\omega)$ for each $u > v$. Therefore,

$$p\left(\boldsymbol{S}^{\mathcal{D}}\right) = \omega^{\left|\boldsymbol{S}^{\mathcal{D}}\right|}(1-\omega)^{\frac{q(q-1)}{2} - \left|\boldsymbol{S}^{\mathcal{D}}\right|}. \tag{2.32}$$

where $\left|\boldsymbol{S}^{\mathcal{D}}\right|$ is the number of edges in $\mathcal{D}$ (equivalently in its skeleton) and $q(q-1)/2$ corresponds to the maximum number of edges in a DAG on q nodes. Finally, we set

$$p(\mathcal{D}) \propto p\left(\boldsymbol{S}^{\mathcal{D}}\right). \tag{2.33}$$

The resulting prior thus depends on the DAG skeleton only and assigns equal prior weights to DAGs having the same number of edges. As Markov equivalent DAGs have the same skeleta, when combined with a compatible DAG-Wishart prior on the parameters, the prior (2.33) implies equal posterior probability for Markov equivalent DAGs. The hyperparameter ω can be tuned to reflect some prior knowledge on the sparsity of the unknown DAG when this information is available.

2.4 MCMC scheme and posterior inference

In this section, we consider the problem of obtaining and using samples from the posterior distribution (2.23). In Section 2.4.1, we detail the MCMC sampling scheme that we adopt. In Section 2.4.2, we describe how the MCMC output can be used to provide estimates of interest in our setting.

2.4.1 Sampling scheme

Our MCMC scheme is designed as a Partial Analytic Structure (PAS) algorithm [Godsill, 2012]. At each iteration, it proposes a new DAG $\mathcal{D}^*$ by sampling it from a *proposal distribution* $q(\mathcal{D}^* \mid \mathcal{D})$ and then accepts it with probability defined by the Metropolis-Hastings *acceptance ratio*

$$\alpha_{\mathcal{D}^*} = \min\left\{1, \frac{p(\boldsymbol{X} \mid \mathcal{D}^*)p(\mathcal{D}^*)q(\mathcal{D} \mid \mathcal{D}^*)}{p(\boldsymbol{X} \mid \mathcal{D})p(\mathcal{D})q(\mathcal{D}^* \mid \mathcal{D})}\right\}. \tag{2.34}$$

After the first two steps, by the conjugacy of the compatible DAG-Wishart prior specified in our model, it is possible to directly sample from the posterior distribution $p((\boldsymbol{L}, \boldsymbol{D}) \mid \mathcal{D}, \boldsymbol{X})$. Samples from the posterior distribution of $\boldsymbol{\gamma}_y^I$ can then be obtained by constructing $\boldsymbol{L}^I$ and $\boldsymbol{\Sigma}^I$ as in Equations 2.15 and 2.16 and applying (2.21). A high-level description of the sampler is presented in Algorithm 1

We now detail the construction of the proposal distribution $q(\mathcal{D}^*|\mathcal{D})$ and the acceptance ratio $\alpha_{\mathcal{D}^*}$ in our model specification.

As for the proposal distribution, we consider three types of operators: $Insert(u, v)$, $Delete(u, v)$, and $Reverse(u, v)$, corresponding respectively to the insertion, deletion, and reversal of the edge (u, v); see Figure 2.2 for an example.

For any $\mathcal{D} \in \mathcal{S}_q$, we can construct the set of *valid* operators $\mathcal{O}_{\mathcal{D}}$, that is operators whose resulting graph is a DAG. Given a current DAG $\mathcal{D}$ we then propose $\mathcal{D}^*$ by uniformly sampling a DAG in $\mathcal{O}_{\mathcal{D}}$. The construction of $\mathcal{O}_{\mathcal{D}}$ and the DAG proposal are summarized in Algorithm 4. Also notice that because there is a one-to-one correspondence between each operator and resulting DAG $\mathcal{D}^*$, the probability of transition from $\mathcal{D}$ to $\mathcal{D}^*$ (a *direct*

Algorithm 1: PAS algorithm for posterior inference.

Input: Data matrix $\boldsymbol{X}$, number of MCMC iterations S, initial DAG $\mathcal{D}^0$, compatible DAG-Wishart hyperparameters $(a, \boldsymbol{U})$, sparsity hyperparameter ω, intervention targets I

Output: S samples from the posterior distribution $p(\mathcal{D}, (\boldsymbol{L}, \boldsymbol{D}), \boldsymbol{\gamma}_y^I \mid \boldsymbol{X})$

for *s in 1:S* **do**
- Set $\mathcal{D}^{(s)} = \mathcal{D}^{(s-1)}$;
- Sample $\mathcal{D}^*$ from $q\left(\mathcal{D}^* \mid \mathcal{D}^{(s-1)}\right)$;
- Set $\mathcal{D}^{(s)} = \mathcal{D}^*$ with probability $\alpha_{\mathcal{D}^*}$;
- Sample $(\boldsymbol{L}, \boldsymbol{D})^{(s)}$ from the posterior distribution $\boldsymbol{\pi}^{c,\Theta_{\mathcal{D}}}_{\tilde{\boldsymbol{a}}^c(\mathcal{D}),\tilde{\boldsymbol{U}}}$;
- Construct $(\boldsymbol{L}^I)^{(s)}$ as in (2.15) and recover $(\boldsymbol{\Sigma}^I)^{(s)}$ as in (2.16);
- Obtain $(\boldsymbol{\gamma}_y^I)^{(s)}$ as in (2.21)

end

return $\{\mathcal{D}^{(s)}, (\boldsymbol{L}, \boldsymbol{D})^{(s)}, (\boldsymbol{\gamma}_y^I)^{(s)}\}_{s=1}^S$;

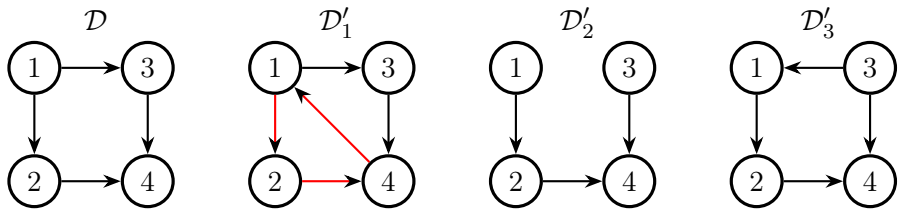


Figure 2.2: *A DAG $\mathcal{D}$ and three modified graphs of the operators $Insert(4,1)$, $Delete(1,3)$, $Reverse(1,3)$ respectively. Operator $Insert(4,1)$ is not valid since $\mathcal{D}'_1$ is not acyclic.*

successor of $\mathcal{D}$) is $q(\mathcal{D}^* \mid \mathcal{D}) = 1/|\mathcal{O}_{\mathcal{D}}|$ and the proposal ratio in (2.34) is

$$\frac{q(\mathcal{D} \mid \mathcal{D}^*)}{q(\mathcal{D}^* \mid \mathcal{D})} = \frac{|\mathcal{O}_{\mathcal{D}}|}{|\mathcal{O}_{\mathcal{D}^*}|}. \tag{2.35}$$

Algorithm 2: PAS algorithm: construction of $\mathcal{O}_{\mathcal{D}}$

Input: A DAG $\mathcal{D} = (V, E)$
Output: A set of valid operators $\mathcal{O}_{\mathcal{D}}$
Set $\mathcal{O}_{\mathcal{D}} = \emptyset$;
Construct $E_I = \{(u, v) : (u, v) \notin E \wedge (v, u) \notin E\}$;
Construct $E_D = \{(u, v) : (u, v) \in E\}$;
for $e \in E_D$ **do**
  Add $Delete(e)$ to $\mathcal{O}_{\mathcal{D}}$;
  if $Reverse(e)$ *is* valid **then** add it to $\mathcal{O}_{\mathcal{D}}$;
end
for $e \in E_I$ **do**
  if $Insert(e)$ *is* valid **then** add it to $\mathcal{O}_{\mathcal{D}}$;
end
return $\mathcal{O}_{\mathcal{D}}$;

Because of the structure of the proposal distribution, at each step of our MCMC algorithm we need to compare two DAGs $\mathcal{D}$ and $\mathcal{D}^*$ which differ by one edge only. Notice that the operator $Reverse(u, v)$ can be also brought back to the same case since is equivalent to the consecutive application of the operators $Delete(u, v)$ and $Insert(v, u)$. We thus consider two DAGs $\mathcal{D}$ and $\mathcal{D}^*$ differing by one edge, so that $(u, v) \in \mathcal{D}$ and $(u, v) \notin \mathcal{D}^*$, and we denote by $(\boldsymbol{L}_{\mathcal{D}}, \boldsymbol{D}_{\mathcal{D}})$ and $(\boldsymbol{L}_{\mathcal{D}^*}, \boldsymbol{D}_{\mathcal{D}^*})$ the corresponding Cholesky parameters. These differ only for their v-th components $\big((\boldsymbol{L}_{\mathcal{D}})_{\mathrm{pa}_{\mathcal{D}}(v)\times v}, (\boldsymbol{D}_{\mathcal{D}})_{vv}\big)$ and $\big((\boldsymbol{L}_{\mathcal{D}^*})_{\mathrm{pa}_{\mathcal{D}^*}(v)\times v}, (\boldsymbol{D}_{\mathcal{D}^*})_{vv}\big)$. By the decomposability property of the marginal likelihood in our model specification (Equation 2.30), also the marginal likelihoods of the two DAGs will differ only for their v-th component and their ratio in (2.34) becomes

$$\frac{p(\boldsymbol{X} \mid \mathcal{D}^*)}{p(\boldsymbol{X} \mid \mathcal{D})} = \frac{p(\boldsymbol{X}_v \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}^*}(v)}, \mathcal{D}^*)}{p(\boldsymbol{X}_v \mid \boldsymbol{X}_{\mathrm{pa}_{\mathcal{D}}(v)}, \mathcal{D})}. \tag{2.36}$$

As a consequence, it only requires the computation of the statistics relative to the v-th component of the two DAGs, resulting in a lower computational cost. From our DAG-prior specification in Equations (2.32) and (2.33) it also follows that

$$\frac{p(\mathcal{D}^*)}{p(\mathcal{D})} = \frac{\omega}{1 - \omega}, \tag{2.37}$$

which is reversed if $\mathcal{D}^*$ differs from $\mathcal{D}$ for an edge deletion. Sampling from the compatible DAG-Wishart distribution $\boldsymbol{\pi}^{c,\Theta_{\mathcal{D}}}_{\tilde{\boldsymbol{a}}^c(\mathcal{D}),\tilde{\boldsymbol{U}}}$ can be achieved by resorting to the component-wise representation of (2.28). The sample from the posterior distribution of the joint causal effect of interest then follows by applying (2.21) to each sampled value.

2.4.2 Posterior inference

The output of Algorithm 1 consists of a collection of DAGs and Cholesky parameters $\left\{\mathcal{D}^{(s)}, (\boldsymbol{L}, \boldsymbol{D})^{(s)}\right\}_{s=1}^{S}$ sampled from their posterior distribution and a collection of joint causal effects $\left\{\boldsymbol{\gamma}^I_y\right\}_{s=1}^{S}$ for a set of input intervention targets $I \subset [q]$. From this output, the posterior probability of a DAG $\mathcal{D}$ can be approximated as

$$\hat{p}(\mathcal{D} \mid \boldsymbol{X}) \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left\{\mathcal{D}^{(s)} = \mathcal{D}\right\}, \tag{2.38}$$

which corresponds to the DAG frequency of visits in the chain. Alternatively, approximations of posterior model probabilities can be obtained from re-normalized marginal likelihoods; see also García-Donato and Martínez-Beneito [2013] for a discussion. From (2.38), a MAP DAG estimate can be immediately recovered as

$$\hat{\mathcal{D}}_{\text{MAP}} = \underset{\mathcal{D}}{\text{argmax}}\ \hat{p}(\mathcal{D} \mid \boldsymbol{X}). \tag{2.39}$$

Alternatively, for each pair of nodes (u, v) we can compute the (estimated) posterior probability of edge inclusion

$$\hat{p}(u \to v \mid \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left\{u \to v \in \mathcal{D}^{(s)}\right\}, \tag{2.40}$$

from which a Median Probability (DAG) Model $\hat{\mathcal{D}}_{\text{MPM}} = (V, \hat{E}_{\text{MPM}})$ can immediately be recovered, where

$$\hat{E}_{MPM} = \{(u, v) : \hat{p}(u \to v \mid \boldsymbol{X}) \geq 0.5\}. \tag{2.41}$$

Finally an overall summary of the posterior distribution of each causal effect coefficient $\gamma^I_{x_h y}, h \in I$ can be computed as

$$\hat{\gamma}^I_{x_h y} = \frac{1}{S} \sum_{s=1}^{S} \left(\gamma^I_{x_h y}\right)^{(s)}, \tag{2.42}$$

which corresponds to a Bayesian Model Averaging (BMA) estimate wherein posterior model probabilities are approximated through their MCMC frequencies of visits. We underline that (2.42) naturally incorporates the uncertainty around both the underlying causal DAG model and the allied DAG-dependent parameters.

2.5 Simulations and real data analysis

In this section, we evaluate the performance of our methodology on both simulated and real data. In Section 2.5.1, we evaluate our method on the task of causal discovery by comparing it with the PC algorithm. In Section 2.5.2, we instead focus on the task of causal effect estimation and provide a comparison with the IDA method. Finally, in Section 2.5.3 we illustrate a real data application of our methodology.

2.5.1 DAG selection

We construct different simulation scenarios by varying the sample size $n \in \{50, 100, 200, 500\}$ and the number of nodes $q \in \{10, 20\}$. Under each simulation scenario by $n \times q$ we generate $N = 30$ datasets, each obtained as follows. We first randomly sample a sparse DAG $\mathcal{D}$ by fixing a probability of edge inclusion equal to 0.2. We then generate the corresponding (true) Cholesky parameters $(\boldsymbol{L}, \boldsymbol{D})$ by drawing the non-zero elements of $\boldsymbol{L}$ from $[-1, -0.1] \cup [0.1, 1]$ while fixing $\boldsymbol{D} = \boldsymbol{I}_q$. We finally construct the covariance matrix $\boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{L}^{-T}\boldsymbol{D}\boldsymbol{L}^{-1}$ and generate n multivariate i.i.d. observations, representing an (n, q) dataset $\boldsymbol{X}$, from the Gaussian DAG-model $\mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma}_{\mathcal{D}})$. For each simulated dataset we run $S = 25000$ iterations of Algorithm 1 to approximate the posterior distribution over DAGs and Cholesky parameters.

To assess the accuracy of our method in recovering the underlying causal structure we compare DAG point estimates that can be retrieved from our MCMC output with the corresponding true DAGs. Similarly, we evaluate the performance of the frequentist *PC* algorithm, the structural learning method underlying the IDA approach of Maathuis et al. [2009]. Specifically, with regard to our method, we consider both $\hat{\mathcal{D}}_{\text{MPM}}$, the Median Probability (DAG) Model (MPM), and $\hat{\mathcal{D}}_{\text{MAP}}$, the Maximum A Posteriori (MAP) DAG estimates as defined in the previous section. We implement the PC algorithm at significance level $\alpha \in \{0.01, 0.05, 0.10\}$. In addition, because PC outputs a *CPDAG*, starting from each of our DAG estimates (MPM and MAP) we construct the representative CPDAG, that is the CPDAG representing the equivalence class of the estimated DAG. We compare each CPDAG estimate with the CPDAG representing the equivalence class of the true DAG in terms of *Structural Hamming Distance* (SHD). SHD corresponds to the number of edge

insertions, deletions or flips needed to transform the estimated graph into the true graph. Lower values of SHD correspond to better performances. Our results are summarized in the box-plots of Figure 2.3 which report the distribution of the two indexes over the $N = 30$ simulations. It appears that our MPM and MAP estimates are competitive with PC for moderate sample sizes and perform slightly better than PC as n increases in terms of SHD.

2.5.2 Causal effect estimation

We now evaluate the performance of our method in causal effect estimation. In particular, for each of the scenarios considered in the previous section, we vary the number of intervened nodes (size of the target) $s \in \{2, 4\}$ and we randomly choose a target I consisting of s nodes randomly sampled from $\{2, \dots, q\}$. We then recover the post-intervention parameters $\boldsymbol{L}^I$ using (2.15) and compute the induced $\boldsymbol{\Sigma}_{\mathcal{D}^I}$; the true set of causal effects $\boldsymbol{\gamma}_y^I$ follows from (2.11). We then compute the BMA estimate (2.42) for each intervened node $h \in I$. Each estimated causal effect $\hat{\gamma}^I_{x_h y}, h \in I$, is compared with the corresponding true causal effect $\gamma^I_{x_h y}$ by computing the absolute-value distance

$$d_h^{\text{BMA}} = \left|\hat{\gamma}^I_{x_h y} - \gamma^I_{x_h y}\right|. \tag{2.43}$$

We include in our analysis the joint-IDA method of Nandy et al. [2017]. In particular, for the graph selection step we implement PC algorithm at significance level $\alpha = 0.01$ which has also been shown to perform better in sparse settings Kalish and Buhlmann [2007]. Joint-IDA recovers for each intervened node $h \in I$ the set of distinct causal effects compatible with the input CPDAG. This is then summarized through the arithmetic mean which provides an estimate of $\theta^l_{h,1}$, $h \in I$. The joint-IDA estimate is compared with the true causal effect by computing the absolute-value distance d_h^{IDA} following (2.43). Results are summarized in the box-plots of Figure 2.4 which reports the distribution of d_h^{BMA} and d_h^{IDA} across the 30 simulated datasets and intervened nodes, for increasing values of the sample size n, different number of variables q and different size of the target $s \in \{2, 4\}$. Clearly, lower values of the distance correspond to better performances.

It appears that both methods improve their performances as the sample size increases. However, our BMA-based method outperforms joint-IDA under all scenarios and in particular in the setting $s = 4$. One possible reason is that, differently from our Bayesian method, joint-IDA relies on a given (estimated) equivalence class of DAGs. Indeed, causal inference results strongly depend on the input CPDAG estimate and therefore on the accuracy in the graph selection. By contrast, our MCMC-based method relies on a posterior distribution over a collection of DAGs some of which, although lying outside the true-DAG equivalence

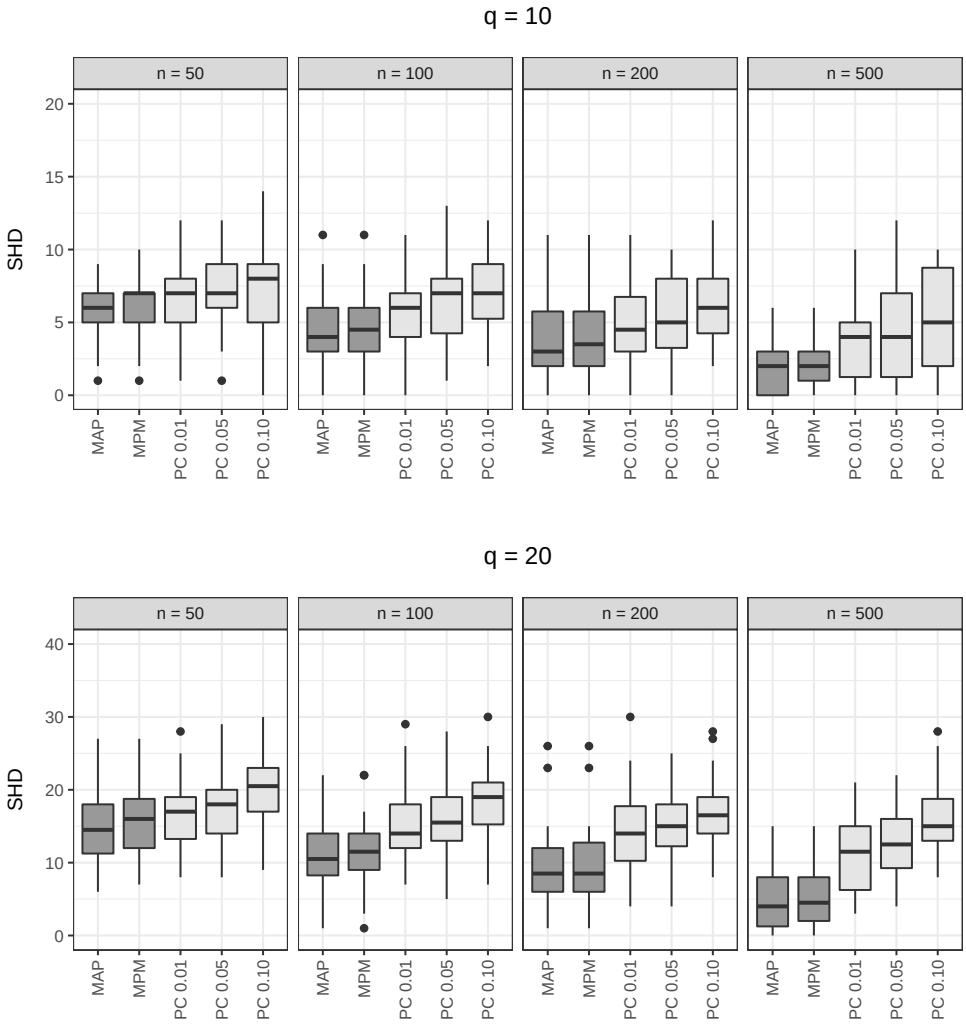


Figure 2.3: *Simulation study. Distribution over $N = 30$ simulated datasets of the Structural Hamming Distance (SHD) between estimated and true CPDAGs. Methods under comparison are: our Bayesian method with output the Median Probability Model (MPM) and Maximum A Posteriori (MAP) graph estimates and the PC algorithm implemented at significance level $\alpha \in \{0.01, 0.05, 0.10\}$, respectively PC 0.01, PC 0.05, PC 0.10.*

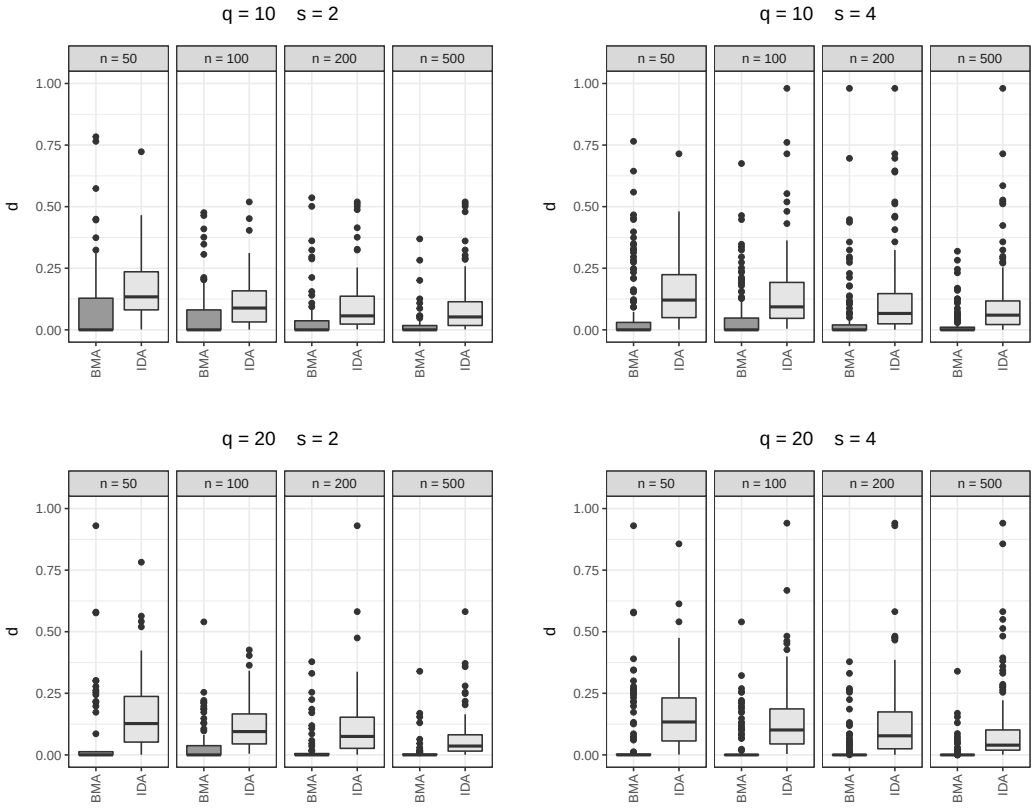


Figure 2.4: *Simulation study. Distribution of the absolute-value distance d between estimated and true causal effects for size of the target* $s \in \{2, 4\}$*, number of variables* $q \in \{10, 20\}$ *and sample size* $n \in \{50, 100, 200, 500\}$*. Methods under comparison are: our BMA-based approach (BMA) and the Joint-IDA method (IDA).*

class, might be "structurally similar" to the true causal DAG and still result in a causal effect which is close to the true one.

2.5.3 Real data analysis

In this section we apply our methodology and joint-IDA to the "Wine quality" dataset of Cortez et al. [2009]; the dataset is publicly available at `https:// archive.ics.uci.edu/`. In our analysis we include observations of seven continuous variables measuring physicochemical properties of a Portoguese wine called Vinho verde, and a response variable representing a sensory score of the wine quality (ranging in $0 - 10$) given by $n = 1593$ independent assessors.

This dataset has been often used for prediction tasks, i.e. to evaluate the quality of wine on the basis of its physicochemical properties only. However, one might be also interested in causal questions, such as whether intervening on one (or more) physicochemical property may change the wine sensory score. As a consequence, this can lead to identify the target of intervention which produces the largest increase in the score.

We run Algorithm 1 to approximate the posterior distribution of DAGs, DAG-parameters and causal effects for any variable in the system and the joint-IDA method based on a CPDAG estimated obtained from PC algorithm. Because one can reasonably assume that the quality score does not affect any of the physicochemical properties (but rather the opposite is argued), we restrict the space of DAGs by imposing that node 1 (the sensory score) cannot have descendant nodes. Such a constraint introduces prior information on the causal structure which is suggested by the problem. In our MCMC algorithm, this is achieved by limiting the set of valid operators of type Insert involving node 1 to those of the form $u \to 1$. In the PC algorithm instead, this background information is included with the following procedure: we first estimate the skeleton between variables $X_1, \dots, X_q$ as in the standard first step of PC. Next, we orient undirected edges between variables Y and covariates $X_2, \dots, X_q$ as $X_j \to Y$, while apply Meek's orientation rules to orient the sub-graph of $X_2, \dots, X_q$; see also Kalish and Buhlmann [2007] for details.

We first assess the convergence of the MCMC algorithm by running two independent chains of length $S = 50000$. Figure 2.5 summarizes the estimated posterior probabilities of edge inclusion (Equation (2.40)) computed from each MCMC chain. The two resulting heatmaps suggest a highly satisfactory agreement between the two chains.

Starting from our MCMC output we consider both the Maximum a Posteriori (MAP) and the Median Probability Model (MPM) as DAG estimates. However, we stress that our final BMA estimate does not rely on a single DAG but rather on a full posterior of DAGs

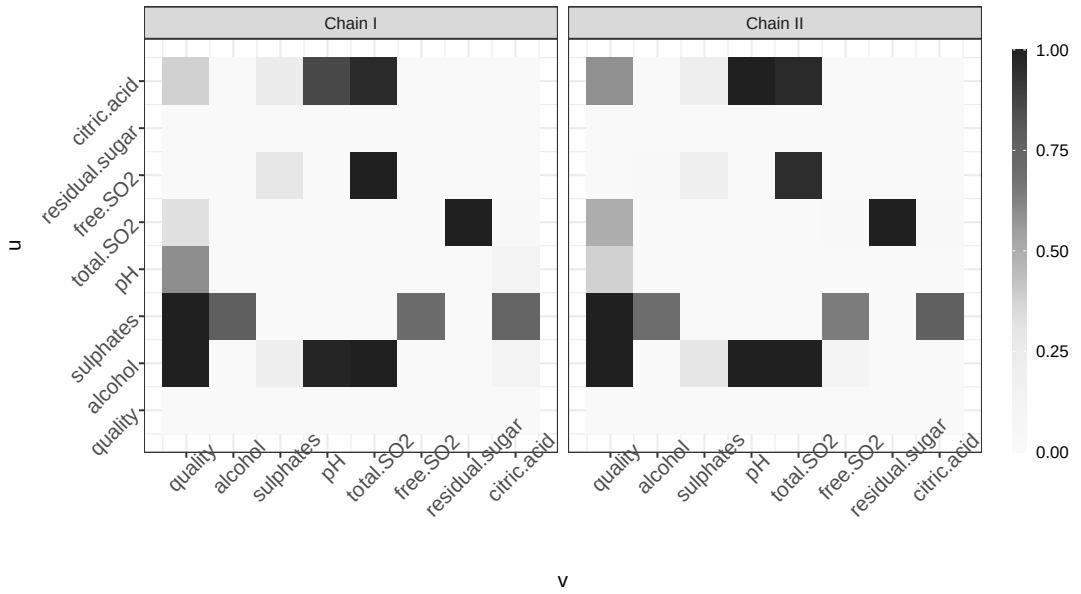


Figure 2.5: *Real data analysis. Heat maps with estimated posterior probabilities of edge inclusion obtained from two independent MCMC chains*

and, accordingly, a single DAG estimate is only constructed as an overall graph summary. The two graphs are reported in Figure 2.6, together with the DAG estimate obtained from the modified version of PC (implemented at significance level $\alpha = 0.01$). There are only a few differences between the three estimates, the most notable being the presence of an additional edge from total.SO2 to quality in the PC estimate.

We now present our results on causal effect estimation. Specifically, we first consider single-node interventions and compute the BMA and joint-IDA estimates of the causal effect on the response for each node (physicochemical property). Moreover, for each pair of nodes, $\{h, k\}$ we obtain the corresponding BMA and joint-IDA causal effect estimates under a joint intervention on $\{X_h, X_k\}$. Results are summarized in the left-side heatmaps of Figure 2.7 Each (h, k)-element $(h \neq k)$ represents the BMA (upper panel) and joint-IDA (lower panel) causal effect estimate of X_k on $Y = X_1$ in a joint intervention on $\{X_h, X_k\}$; main diagonal elements correspond to the causal effects as obtained from single-node interventions. It appears that an increase in variables alcohol and sulphates may result in an increase in wine quality. By converse, a similar effect can be achieved by reducing the level of pH and total.SO2 since the two covariates exhibit negative causal effects.

The right-side heatmaps of Fig. 2.7 reports for each pair (h, k) the sum of the corresponding two (absolute value) BMA (upper panel) and joint-IDA (lower panel) causal effect

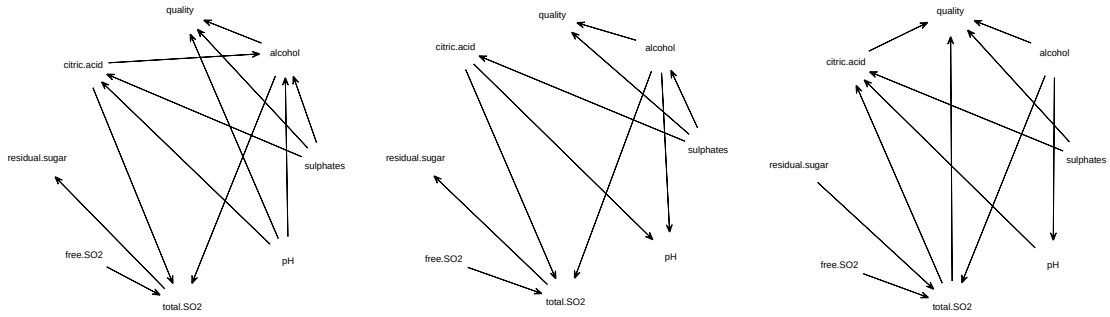


Figure 2.6: *Real data analysis. Comparison between estimated graphs. From left to right: maximum a posteriori, median probability and modified-PC DAG estimates*

estimates obtained under the joint intervention on $\{X_h, X_k\}$, that is $\left|\hat{\theta}^{\{h,k\}}_{h,1}\right| + \left|\hat{\theta}^{\{h,k\}}_{k,1}\right|$. Each of these terms provides an overall measure of the "strength" of the causal effect that a joint intervention on the two variables might produce on the response. As a consequence, this collection of coefficients allows to identify which pair of variables is associated to the largest potential increase in quality sensory score. In particular, it appears that a joint intervention on variables alcohol and sulphates has the largest effect on the response variable. This result is invariant with respect to the method used, as it can be observed by comparing the upper and lower heatmaps of Figure 2.7. Substantial differences between the two methods appear, instead, for variable total.SO2, which under joint-IDA is associated with a (negative) causal effect on quality. In addition, joint-IDA causal effect estimates are somewhat higher than those obtained under our BMA method. We remark that the effect of joint interventions on more than two variables can be evaluated in a similar way. However, for simplicity of exposition, we have limited our analysis to the case of pair-node interventions.

2.6 Discussion

In this chapter, we present a Bayesian methodology for causal structure learning and causal effect estimation. We assume that multivariate observational data have been generated by an unknown Gaussian DAG model. Of special interest is the causal effect of a specific variable on a response arising from a joint intervention on several variables in the system. The latter depends on the underlying causal structure which therefore needs to be estimated. Accordingly, our method combines DAG structural learning and causal effect estimation, leading to a posterior distribution over the space of DAGs, DAG parameters and causal

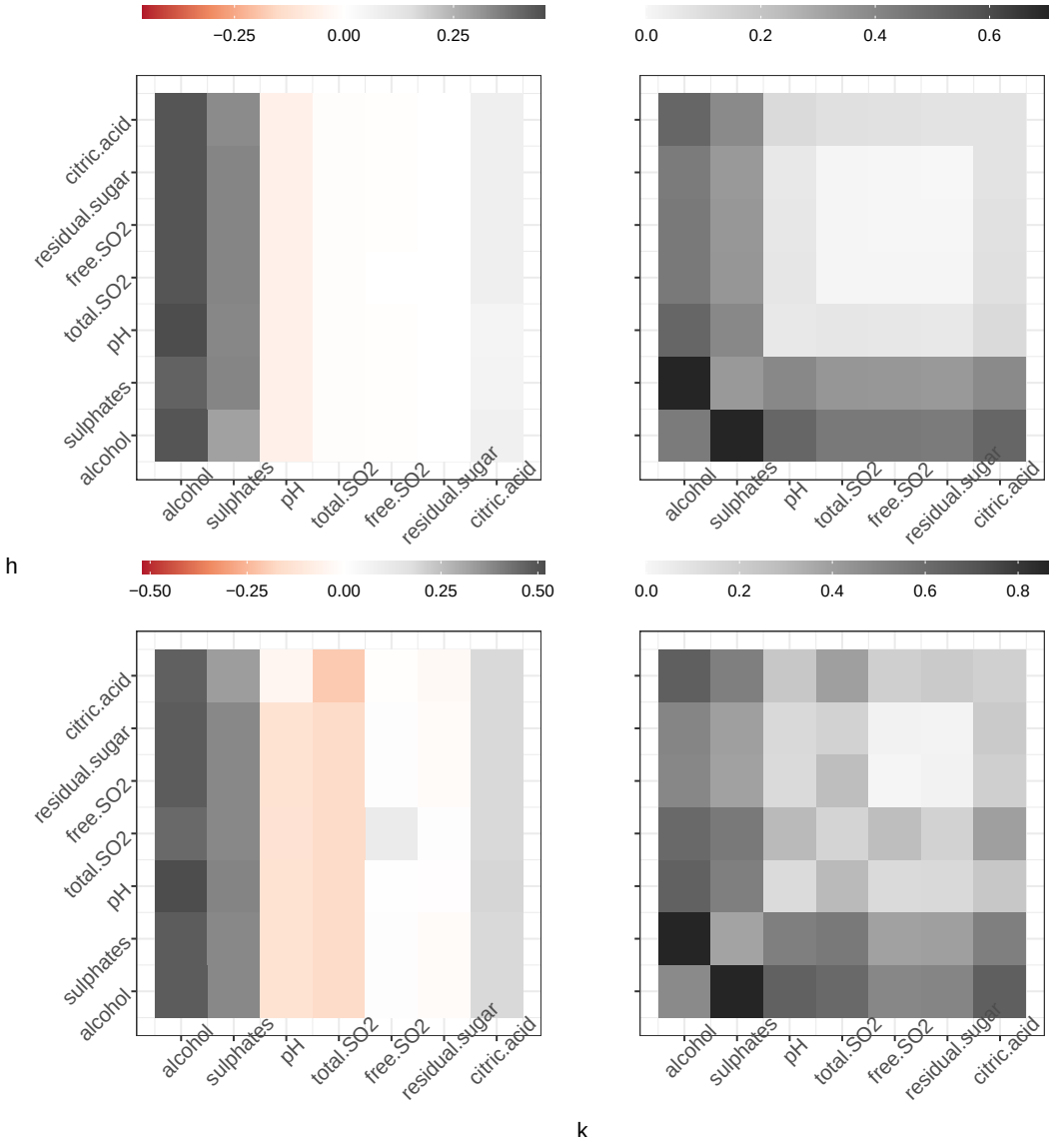


Figure 2.7: *Real data analysis. Left panel: BMA (top) and joint-IDA (bottom) estimates of the causal effect of X_k on Y in a joint intervention on $\{X_h, X_k\}$. Right panel: sum of absolute-value BMA (top) and joint-IDA (bottom) estimates obtained from joint interventions on $\{X_h, X_k\}$.*

effects. Simulation results show that our method outperforms the frequentist benchmark joint-IDA and leads to improved estimates of joint causal effects, especially in scenarios characterized by a moderate sample size. On the other hand, our methodology requires an approximated posterior distribution over the space of DAGs and parameters, which might become computationally expensive as the number of variables increases. Differently, joint-IDA has been specifically developed for high dimensional settings and therefore can efficiently perform even when thousands of variables are involved. However, its output relies on a single estimated equivalence class of DAGs whose identification may affect the causal estimation results.

2.6.1 Future developments

Joint interventions lead to causal effects that can significantly deviate from their single-node counterparts. Accordingly, a desired effect on the response can be obtained through a unique intervention involving several variables simultaneously, rather than a sequence of single-node interventions. Since the number of possible joint interventions grows exponentially in the number of variables, the investigation of an optimization strategy which identifies the optimal intervention targets producing the desired level of response could be of interest.

In addition, in this chapter we considered causal effect estimation from joint *hard* interventions. A more general framework, named soft interventions, assumes that parent-child dependencies are "modified" but yet preserved after intervention. In this setting, Correa and Bareinboim [2020] introduce a set of rules (named σ-calculus) for the identifiability of causal effects arising from soft interventions. They then show how these rules can be applied to identify the causal effect of interventions from a combination of observational and experimental data. A Bayesian framework for the estimation of causal effect arising from soft interventions is, to our knowledge, still lacking and is currently under investigation by the authors.

Finally, a DAG cannot be uniquely identified from observational data and accordingly a possibly large collection of causal effects is estimated. Randomized intervention experiments producing interventional data can be used to improve the identifiability of the data-generating model which consequently reduces the uncertainty around the causal effect estimate; see also Castelletti et al. [2018]. In principle, one could then perform sequential simultaneous intervention leading to the identification of the true causal effect. This issue can be tackled from an optimal design of experiment perspective implementing an objective function whose optimization reduces the uncertainty related to each BMA causal effect estimate of interest.

Bibliography

E. Ben-David, T. Li, H. Massam, and B. Rajaratnam. High Dimensional Bayesian Inference for Gaussian Directed Acyclic Graph Models. *arXiv preprint*, 2015.

K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

F. Castelletti and G. Consonni. Bayesian Inference of Causal Effects from Observational Data in Gaussian Graphical Models. *Biometrics*, 77(1):136–149, 2021a.

F. Castelletti and G. Consonni. Bayesian Causal Inference in Probit Graphical Models. *Bayesian Analysis*, 16(4):1113–1137, 2021b.

F. Castelletti and G. Consonni. Bayesian Graphical Modeling for Heterogeneous Causal Effects. *Statistics in Medicine*, 42(1):15–32, 2023.

F. Castelletti, G. Consonni, M. L. D. Vedova, and S. Peluso. Learning Markov Equivalence Classes of Directed Acyclic Graphs: An Objective Bayes Approach. *Bayesian Analysis*, 13(4):1235–1260, 2018.

D. M. Chickering. Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, 3:507–554, 2002.

J. Correa and E. Bareinboim. A Calculus For Stochastic Interventions: Causal Effect Identification and Surrogate Experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Using Data Mining for Wine Quality Assessment. In J. Gama, V. S. Costa, A. M. Jorge, and P. B. Brazdil, editors, *Discovery Science*, pages 66–79. Springer Berlin Heidelberg, 2009.

G. García-Donato and M. A. Martínez-Beneito. On Sampling Strategies in Bayesian Variable Selection Problems With Large Model Spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013.

D. Geiger and D. Heckerman. Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions. *The Annals of Statistics*, 30 (5):1412–1440, 2002.

S. J. Godsill. On the Relationship between Markov chain Monte Carlo Methods for Model Uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2012.

L. Henckel, E. Perković, and M. H. Maathuis. Graphical Criteria for Efficient Total Effect Estimation Via Adjustment in Causal Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 03 2022.

M. Kalish and P. Buhlmann. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8:613–36, 2007.

M. H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating High-dimensional Intervention Effects from Observational Data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

P. Nandy, M. H. Maathuis, and T. S. Richardson. Estimating the Effect of Joint Interventions from Observational Data in Sparse High-dimensional Settings. *The Annals of Statistics*, 45(2):647–674, 2017.

J. Pearl. *Causality*. Cambridge University Press, 2009.

S. Peluso and G. Consonni. Compatible Priors for Model Selection of High-dimensional Gaussian DAGs. *Electronic Journal of Statistics*, 14(2):4110–4132, 2020.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

A. Roverato and G. Consonni. Compatible Prior Distributions for Directed Acyclic Graph Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1): 47–61, 12 2003.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 2001.

T. Verma and J. Pearl. Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI 90, pages 255–270, New York, NY, USA, 1990. Elsevier Science Inc.

J. Viinikka, A. Hyttinen, J. Pensar, and M. Koivisto. Towards Scalable Bayesian Learning of Causal DAGs. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6584–6594. Curran Associates, Inc., 2020.

S. Wright. The Method of Path Coefficients. *The Annals of Mathematical Statistics*, 5(3): 161–215, 1934.

Chapter 3

Bayesian causal discovery from unknown general interventions

3.1 Introduction

Directed Acyclic Graphs (DAGs) are widely used to represent causal relationships between variables. In this setting, learning the DAG structure from data is referred to as causal discovery. If only observational data are available, a DAG is in general identifiable only up to its Markov equivalence class, which includes all DAGs that imply the same conditional independencies [Verma and Pearl, 1990]. However, if in addition one collects interventional (experimental) data, then it is possible to identify smaller sub-classes of DAGs, known as Interventional-Markov Equivalence Classes (I-MECs) [Hauser and Bühlmann, 2012].

Current methods for causal discovery leveraging experimental data typically assume either hard or soft interventions. In essence, a *hard* intervention consists of fixing the values of certain target variables and graphically corresponds to removing all those edges pointing towards the intervened nodes. On the other hand, a *soft* intervention, or mechanism change [Tian and Pearl, 2001], modifies the relationship between each intervened node and its parents without completely destroying it. However, these two types of interventions do not encompass the full spectrum of manipulations that an experimenter can in practice implement or achieve.

Consider the example in Figure 3.1. DAG *a)* represents a causal structure involving four variables: weekly traffic level ($\mathbf{TR}_t$), weekly average air quality level ($\mathbf{AQ}_t$), weekly initial air quality level ($\mathbf{AQ}_0$), and weekly count of individuals reporting respiratory health issues ($\mathbf{RH}_t$) in a specific urban area. In this context, a hard intervention could consist in prohibiting car access to the area, therefore setting $\mathbf{TR}_t = 0$ for the subsequent weeks. A different policy might impose specific restrictions to vehicles entering the area, such as the adoption of particulate filters. This action would simultaneously reduce traffic levels

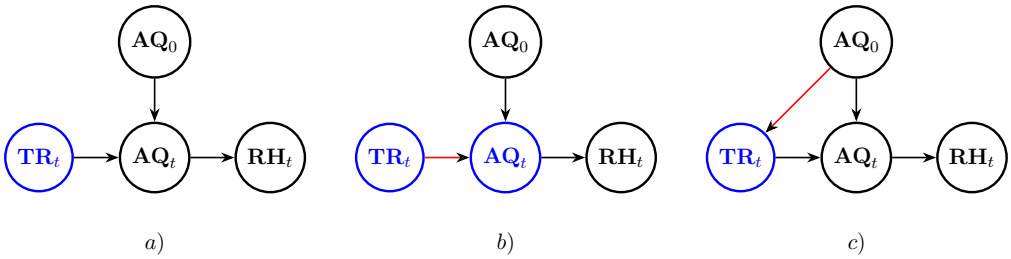


Figure 3.1: *Three DAGs resulting from different types of interventions: a) a hard intervention on* $\boldsymbol{TR}_t$*; b) simultaneous hard (on* $\boldsymbol{TR}_t$*) and soft (on* $\boldsymbol{AQ}_t$*) interventions; c) a general intervention on* $\boldsymbol{TR}_t$*. Target nodes are depicted in blue, while structural modifications induced by the interventions are colored in red.*

and alter the relationship between traffic and air quality, thus resulting in both a hard intervention on $\mathbf{TR}_t$ and a soft intervention on $\mathbf{AQ}_t$. Another possible policy could regulate the number of car accesses on the basis of the initial air quality $\mathbf{AQ}_0$. The resulting post-intervention graph is illustrated in panel *c)* of Figure 3.1, where $\mathbf{AQ}_0$ is now a parent of $\mathbf{TR}_t$. This last type of intervention is commonly referred to in the literature as *dynamic plan* [Pearl and Robins, 1995], although sometimes still labeled as soft intervention [Correa and Bareinboim, 2020]. Throughout the chapter, we use the term *general* for those interventions that modify the parent sets of the target nodes, to emphasize their ability to represent both hard and soft interventions as special cases.

Including general interventions in a causal discovery framework becomes essential in cases where the effect of an intervention is unknown. For instance, in neuroimaging, and specifically in the field of effective connectivity analysis, the objective is to understand how the brain-connectivity network changes in response to external stimuli [Friston, 2011]. In biology, discerning key differences between gene regulatory networks may provide insights into mechanisms of initiation and progression of specific diseases across different groups of patients [Shojaie, 2021].

In this chapter, we develop a Bayesian methodology for causal discovery from unknown general interventions. We set this problem in a Bayesian model selection framework, under which priors on DAG models and associated parameters are combined with a parametric likelihood to obtain a posterior distribution on DAGs and unknown general interventions. This task presents many challenges, primarily the development of *compatible* parameter priors [Roverato and Consonni, 2003] leading to closed-form DAG marginal likelihoods and guaranteeing *score equivalence* of I-Markov equivalent DAGs. We thus provide definitions and graphical characterizations of equivalence classes of DAGs and unknown general

interventions. We then develop a Bayesian framework for data collected under different experimental settings, which applies to parametric models satisfying a set of general assumptions; under the same assumptions, we develop an effective procedure for parameter-prior elicitation which guarantees desirable properties in terms of marginal likelihoods, and in particular score equivalence. Finally, we devise a Markov Chain Monte Carlo (MCMC) scheme to sample from the target distribution, thus allowing for posterior inference of DAG structures and unknown general interventions.

3.1.1 Related work

The first historical work on causal discovery from mixtures of observational and experimental data dates back to Cooper and Yoo [1999], who proposed a Bayesian methodology for data arising from hard interventions with known targets. Issues related to DAG identifiability in this setting were first investigated by Hauser and Bühlmann [2012], who introduced the notion of I-Markov equivalence, provided related graphical characterizations, and developed the Greedy Interventional Equivalence Search (GIES) algorithm for structure learning. Consistency of the underlying BIC score were also established by Hauser and Bühlmann [2015]. An objective Bayesian methodology working on the space of I-Markov equivalence classes in the Gaussian setting was then developed by Castelletti and Consonni [2019]. In the same setting, Wang et al. [2017] developed the Interventional Greedy Sparsest Permutation (IGSP) method, later extended to the case of soft interventions by Yang et al. [2018], who also generalized the identifiability results of Hauser and Bühlmann [2012]. An early methodology dealing with soft interventions was already proposed by Tian and Pearl [2001] who also provided graphical characterizations for Markov equivalence.

The first approach for causal discovery under *uncertain* intervention targets is represented by Eaton and Murphy [2007]. The authors adopted a Bayesian framework for categorical data and allowed the interventions to be soft and unknown, though without addressing identifiability issues. A more recent Bayesian methodology for Gaussian data and accounting for I-Markov equivalence, assuming hard interventions, was instead introduced by Castelletti and Peluso [2023a]. Squires et al. [2020] proposed an extension of IGSP that allows for uncertainty on the targets of intervention and proved its consistency. More recently, Gamella et al. [2022] focused on the case of experimental Gaussian data generated from unknown noise-interventions, providing identifiability results for both DAGs and intervention targets. Similar results, in a non-parametric setting, were provided by Jaber et al. [2020], assuming soft interventions and allowing for the presence of hidden confounders. Finally, Mooij et al. [2020] developed the Joint Causal Inference framework, which encodes

unknown interventions through additional indicator variables in a pooled dataset; they also established under which assumptions constraint-based methods conceived for observational settings can be applied to the pooled dataset to learn the intervention targets.

Learning the effects of unknown general interventions is equivalent to learning structural differences between post-intervention DAGs. Under this perspective, our framework relates to other bodies of literature such as inference of multiple DAGs [Castelletti et al., 2020] as well as to methodologies aiming at directly estimating the structural differences between causal DAGs [Wang et al., 2018].

3.1.2 Outline

In Section 3.2 we introduce the basic notation and background on Structural Causal Models (SCMs) and present our results relative to identifiability of DAGs and general interventions from mixtures of observational and interventional data. In Section 3.3 we develop a Bayesian methodology for causal discovery in this newly defined context, leveraging the results of Section 3.2 in order to provide guidance on model construction and prior elicitation. In Section 3.4, we focus on the construction of a Markov Chain Monte Carlo (MCMC) algorithm for sampling from the posterior distribution over DAGs and interventions. Finally, in Section 3.5, we apply our developed methodology to the Gaussian case and empirically assess the performance of our implementations using both simulated and real-world data. Section 3.6 summarizes our conclusions. All proofs of our main results are provided in the appendix to this chapter.

3.2 Identifiability under general interventions

In this section we discuss identifiability of DAGs and unknown general interventions and provide graphical characterizations of I-Markov equivalence. In Section 3.2.1 we provide some background material on DAGs and Structural Causal Models (SCMs) and we formalize the notion of general intervention. In Section 3.2.2 we define an I-Markov property for this new setting and present our main results on the identifiability of DAGs when interventions are known. Section 3.2.3 extends the results to the case of unknown interventions.

3.2.1 Preliminaries

A Directed Acyclic Graph (DAG) $\mathcal{D} = (V, E)$ with vertex set $V = [q] \coloneqq \{1, \dots, q\}$, and edge set $E \subset V \times V$ is a directed graph with no cycles, i.e. no directed paths starting and ending at the same node. A DAG $\mathcal{D}$ can be represented by a (q, q) adjacency matrix $\boldsymbol{A}$,

such that $\boldsymbol{A}_{ij} = 1$ if $(i, j) \in E$ and 0 otherwise. We let $\mathrm{pa}_{\mathcal{D}}(j)$ be the set of *parents* of node j, that is $\mathrm{pa}_{\mathcal{D}}(j) = \{i \subset V \,|\, \boldsymbol{A}_{ij} = 1\}$, and $\mathrm{fa}_{\mathcal{D}}(j) = j \cup \mathrm{pa}_{\mathcal{D}}(j)$ be the *family* of j in $\mathcal{D}$. Moreover, an edge $i \to j$ is *covered* in $\mathcal{D}$ if $i \cup \mathrm{pa}_{\mathcal{D}}(i) = \mathrm{pa}_{\mathcal{D}}(j)$. We refer to the undirected graph obtained by removing edge directions from a DAG as the *skeleton* of the DAG. Any induced subgraph of the form $i \to j \leftarrow k$, with no edges between i and k, is instead called a *v-structure*. Finally, we say that $\mathcal{D}$ is complete if it has no missing edges.

Under the framework of SCMs, DAGs can be given a causal interpretation by considering each node j as an observable (endogenous) variable X_j and each parent-child relation as a *stable* and *autonomous* mechanism of the form

$$X_j = f_j(X_{\mathrm{pa}_{\mathcal{D}}(j)}, \varepsilon_j), \quad j \in [q], \tag{3.1}$$

where $X_{\mathrm{pa}_{\mathcal{D}}(j)} = \{X_i, i \in \mathrm{pa}_{\mathcal{D}}(j)\}$, f_j is a deterministic function linking X_j to $X_{\mathrm{pa}_{\mathcal{D}}(j)}$ and to an unobserved (exogenous) random variable ε_j [Pearl, 2000]. If $\varepsilon_1, \ldots, \varepsilon_q$ are mutually independent, then the set of structural equations in (3.1) defines a Markovian SCM, and the induced joint density $p(\cdot)$ on $(X_1, \ldots, X_q)$ obeys the Markov property of $\mathcal{D}$, meaning that it factorizes as

$$p(x) = \prod_{j=1}^{q} p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)}), \tag{3.2}$$

where x denotes a realization of the random vector X. The conditional independencies implied by (3.2) can be read-off from the DAG using the notion of *d-separation* [Pearl, 2000]. Let now $\mathcal{M}(\mathcal{D})$ be the set of all positive densities $p(x)$ obeying the Markov property of $\mathcal{D}$. Two DAGs, $\mathcal{D}_1$ and $\mathcal{D}_2$, are called *Markov equivalent* if $\mathcal{M}(\mathcal{D}_1) = \mathcal{M}(\mathcal{D}_2)$. DAGs can be partitioned into *Markov equivalence classes*, each collecting all DAGs that are Markov equivalent. Without specific parametric assumptions, and even under common families of distributions, DAGs can be identified only up to Markov equivalence classes [Pearl, 1988]. The following results provide graphical characterizations of Markov equivalence.

Theorem 1 (Verma and Pearl [1990]). *Two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if they have the same skeleta and the same set of v-structures.*

Theorem 2 (Chickering [1995]). *Two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if there exists a sequence of edge reversals modifying $\mathcal{D}_1$ and such that:*

1. *Each edge reversed is covered;*
2. *After each reversal, $\mathcal{D}_1, \mathcal{D}_2$ belong to the same Markov equivalence class;*
3. *After all reversals $\mathcal{D}_1 = \mathcal{D}_2$.*

Theorem 1 provides a criterion for assessing whether two DAGs belong to the same Markov equivalence class. Theorem 2, instead, is a technical result of great importance to *prove* score equivalence in score-based causal discovery methods.

The mechanisms in Equation (3.1) are stable and autonomous in the sense that it is possible to conceive an external intervention modifying one of the mechanisms (and the corresponding local distribution) without affecting the others. One can envisage different *types* of external interventions [Correa and Bareinboim, 2020]. For any set of *target* variables $T \subset [q]$ and multi-set of *induced parent sets* $P = \{P_1, \dots, P_{|T|}\}$, with $P_j \subset [q]$, we consider interventions producing a mechanism change of the form

$$X_j = \tilde{f}_j(X_{P_j}, \varepsilon_j), \quad \forall\, j \in T. \tag{3.3}$$

We refer to this type of intervention as *general intervention* and, following Correa and Bareinboim [2020], we denote the corresponding operator as $\sigma_{T,P}$. Such intervention induces a new SCM, thus implying a new graphical object.

Definition 3 (Post-intervention graph)**.** *Let $\mathcal{D}$ be a DAG and (T, P) be a pair of intervention targets and induced parent sets defining a general intervention. The post-intervention graph of $\mathcal{D}$ is the graph $\mathcal{D}_{T,P}$ obtained by replacing for each $j \in T$ the new parents P_j induced by the intervention.*

See also Figure 3.2 for an example of DAG and implied intervention graph. Notice that a post-intervention graph need not be a DAG in general. Throughout this chapter we make the following assumption, that we name *validity*.

Definition 4 (Validity)**.** *Let $\mathcal{D}$ be a DAG and (T, P) be a pair of intervention targets and induced parent sets defining a general intervention. The general intervention is* valid *if the post-intervention graph $\mathcal{D}_{T,P}$ is a DAG.*

As a general intervention produces a new Markovian SCM, it also induces a *post-intervention* distribution through the Markov property of $\mathcal{D}_{T,P}$ which can be written as

$$\begin{aligned} p(x \,|\, \sigma_{T,P}) &= \prod_{j=1}^{q} \tilde{p}(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_{T,P}}(j)}) \\ &= \prod_{j \notin T} p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)}) \prod_{j \in T} \tilde{p}(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_{T,P}}(j)}), \end{aligned} \tag{3.4}$$

where the $\tilde{p}(x_j \,|\, \cdot)$'s denote the new local distributions induced by the intervention. For any $j \notin T$, we then have $\tilde{p}(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_{T,P}}(j)}) = p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)})$, so that the local densities of non-intervened nodes are invariant (stable) across pre- and post-intervention distributions. In

the following section we show how these invariances can be leveraged to identify DAGs up to a subset of the original Markov equivalence class (named *I-Markov equivalence class*) and, in the same spirit of Theorem 1 and Theorem 2, we provide a graphical characterization of DAGs belonging to the same I-Markov equivalence class.

3.2.2 DAG identifiability from known general interventions

We consider collections of K experimental settings, or environments, each defined by a general intervention with targets and induced parent sets $T^{(k)}, P^{(k)}$. Let also $\mathcal{T} = \{T^{(k)}\}_{k=1}^K$, $\mathcal{P} = \{P^{(k)}\}_{k=1}^K$ and $\mathcal{I} = (\mathcal{T}, \mathcal{P})$. Each collection of experimental settings entails a family of post-intervention distributions $\big\{p(\cdot \,|\, \sigma_k)\big\}_{k=1}^K$, where to simplify the notation we write $\sigma_k \equiv \sigma_{T^{(k)},P^{(k)}}$ for $k \in [K]$. We assume throughout the chapter that $T^{(1)} = P^{(1)} = \emptyset$, i.e. $k = 1$ corresponds to the observational setting where no intervention has been performed, and $p(\cdot \,|\, \sigma_1) = p(\cdot)$ reduces to the pre-intervention distribution (3.2). Furthermore, we always assume that $\mathcal{I}$ is a collection of targets and induced parent sets defining a *valid* general intervention.

More formally, we can define the possible tuples of joint densities corresponding to K different experimental settings as follows.

Definition 5. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a collection of targets and induced parent sets. Then,*

$$\begin{aligned}\mathcal{M}_{\mathcal{I}}(\mathcal{D}) = \big\{ \{p_k(x)\}_{k=1}^K \mid \forall\, k,l \in [K] : p(x \,|\, \sigma_k) \in \mathcal{M}(\mathcal{D}_k) \text{ and} \\ \forall\, j \notin T^{(k)} \cup T^{(l)}, p_k(x_j \,|\, x_{pa_{\mathcal{D}_k}(j)}) = p_l(x_j \,|\, x_{pa_{\mathcal{D}_l}(j)}) \big\},\end{aligned}$$

where we let for simplicity $p_k(x) = p(x \,|\, \sigma_k)$ and $\mathcal{D}_k = \mathcal{D}_{T^{(k)},P^{(k)}}$. The first condition reflects the fact that, for each experimental setting, the post-intervention distribution obeys the Markov property of the induced post-intervention DAG $\mathcal{D}_k$. The second condition corresponds instead to the local invariances across post-intervention distributions of different experimental settings. Notice that, because of the assumption $T^{(1)} = \emptyset$, $p_1(x) = p(x)$, the observational distribution, and the condition implies that $\forall\, j \notin T^{(k)}$, $p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) = p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)})$. By analogy with the observational case, different DAGs may still imply the same family of pre- and post-intervention distributions, leading to the notion of *I-Markov equivalent* DAGs.

Definition 6 (I-Markov equivalence)**.** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a collection of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1$ and $\mathcal{D}_2$. $\mathcal{D}_1$ and $\mathcal{D}_2$ are I-Markov equivalent (i.e. they belong to the same I-Markov equivalence class) if $\mathcal{M}_{\mathcal{I}}(\mathcal{D}_1) = \mathcal{M}_{\mathcal{I}}(\mathcal{D}_2)$.*

As mentioned, our aim is to develop graphical criteria to establish I-Markov equivalence between DAGs. To this end, we need: i) a graphical object that uniquely represents the DAG $\mathcal{D}$ and the modifications by the valid general interventions; ii) an I-Markov property to read-off the set of conditional independencies and invariances from the graphical object. For the first purpose, we introduce the following construction.

Definition 7. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a collection of targets and induced parents sets. The collection of augmented intervention DAGs ($\mathcal{I}$-DAGs) $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$ is constructed by augmenting each post-intervention DAG $\mathcal{D}_k$ with an $\mathcal{I}$-vertex ζ_k and $\mathcal{I}$-edges $\{\zeta_k \to j, j \in T^{(k)}\}$.*

We provide an example of a collection of $\mathcal{I}$-DAGs in Figure 3.3. The following definition extends the notion of covered edge, originally introduced by Chickering [1995, Definition 2], to our newly defined graphical object.

Definition 8. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a collection of targets and induced parent sets implying a collection of $\mathcal{I}$-DAGs $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$. An edge $i \to j$ in $\mathcal{D}$ is simultaneously covered if:*

1. *$i \to j$ is covered in $\mathcal{D}$;*
2. *For any $k \in [K], k \neq 1$, $i \to j$ is either covered in $\mathcal{D}_k^{\mathcal{I}}$, or $\{i, j\} \subseteq T^{(k)}$;*

For the second purpose instead, we introduce the following definition of I-Markov property.

Definition 9 (I-Markov property)**.** *Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a collection of targets and induced parent sets. Let $\{p_k(x)\}_{k=1}^K$ be a family of strictly positive probability distributions over $(X_1, \dots, X_q)$. Then, $\{p_k(x)\}_{k=1}^K$ satisfies the I-Markov property with respect to $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$ if:*

1. *$p_k(x_A \,|\, x_B, x_C) = p_k(x_A \,|\, x_C)$ for any $k \in [K]$ and any disjoint sets $A, B, C \subset [q]$ such that C d-separates A and B in $\mathcal{D}_k$;*
2. *$p_k(x_A \,|\, x_C) = p_1(x_A \,|\, x_C)$ for any $k \in [K]$ and any disjoint sets A, C such that C d-separates A from ζ_k in $\mathcal{D}_k^{\mathcal{I}}$.*

Point 1. applies the usual Markov property to the pre- and post-intervention graphs $\mathcal{D}_k$, $k \in [K]$. Notice that, because general interventions may induce new parent sets, the set of implied conditional independencies may also change across experimental settings. Point 2. instead imposes a local invariance whenever a d-separation statement involving $\mathcal{I}$-vertices holds in the augmented intervention DAGs. If a tuple of post-intervention distributions $\{p(\cdot \,|\, \sigma_k)\}_{k=1}^K$ is $\mathcal{I}$-Markov w.r.t $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$, then any d-separation statement in $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$ will imply either a conditional independence relationship or an invariance in $\{p(\cdot \,|\, \sigma_k)\}_{k=1}^K$.

Figure 3.2: *A DAG $\mathcal{D}$ and the post-intervention DAG $\mathcal{D}_{T,P}$ for intervention target $T = \{3\}$ and induced parent set $P = \{2\}$.*

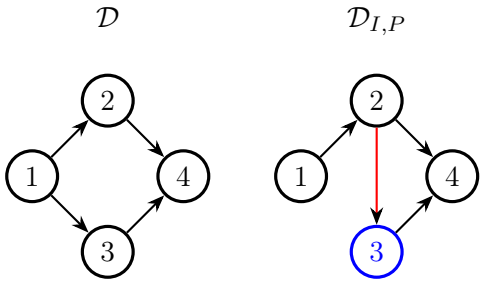
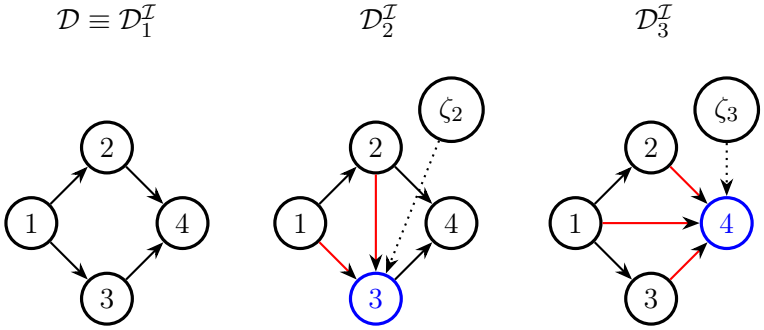


Figure 3.3: *A collection of $\mathcal{I}$-DAGs for DAG $\mathcal{D}$ and a collection of targets and induced parent sets such that $T^{(2)} = \{3\}$, $P^{(2)} = \{1, 2\}$ and $T^{(3)} = \{4\}, P^{(3)} = \{1, 2, 3\}$. Blue nodes represent the intervention targets, while red edges correspond to the induced parent sets.*



Throughout the chapter, we also assume the converse, so that any invariance and any conditional independence relationship in the tuple of distributions implies a d-separation in $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$. Following Squires et al. [2020], we call this assumption $\mathcal{I}$-faithfulness.

Definition 10 (I-Faithfulness)**.** *Let $\mathcal{D}$ be a DAG and a $\mathcal{I}$ a collection of targets and induced parent sets. Let $\{p_k(x)\}_{k=1}^K$ be a set of strictly positive probability distributions over $(X_1, \dots, X_q)$. Then, $\{p_k(x)\}_{k=1}^K$ is said to be I-faithful with respect to $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$ if:*

1. *For any $k \in [K]$ and any disjoint sets $A, B, C \subset [q]$, $p_k(x_A \,|\, x_B, x_C) = p_k(x_A \,|\, x_C)$ if and only if C d-separates A and B in $\mathcal{D}_k$;*
2. *For any $k \in [K]$ and any disjoint sets A, C, $p_k(x_A \,|\, x_C) = p_1(x_A \,|\, x_C)$ if and only if C d-separates A from ζ_k in $\mathcal{D}_k^{\mathcal{I}}$.*

Using the I-Markov property, it is possible to characterize the newly defined I-Markov equivalence class of families of distributions through the $\mathcal{I}$-DAGs, as stated in the following proposition.

Proposition 11. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a collection of targets and induced parent sets. Then $\{p_k(\cdot)\}_{k=1}^K \in \mathcal{M}_{\mathcal{I}}(\mathcal{D})$ if and only if $\{p_k(\cdot)\}_{k=1}^K$ satisfies the I-Markov property with respect to $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$.*

We are finally able to characterize I-Markov equivalence by means of graphical criteria.

Theorem 12. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a collection of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1, \mathcal{D}_2$. $\mathcal{D}_1$ and $\mathcal{D}_2$ belong to the same I-Markov equivalence class if and only if $\mathcal{D}_{1,k}^{\mathcal{I}}$ and $\mathcal{D}_{2,k}^{\mathcal{I}}$ have the same skeleta and v-structures for all $k \in [K]$.*

Theorem 13. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a collection of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1$ and $\mathcal{D}_2$. $\mathcal{D}_1$ and $\mathcal{D}_2$ belong to the same I-Markov equivalence class if and only if there exists a sequence of edge reversals modifying $\mathcal{D}_1$ and such that:*

1. *Each edge reversed is simultaneously covered;*
2. *After each reversal, $\{\mathcal{D}^{\mathcal{I}}_{1,k}\}_{k=1}^{K}$ are DAGs and $\mathcal{D}_1, \mathcal{D}_2$ belong to the same I-Markov equivalence class;*
3. *After all reversals $\mathcal{D}_1 = \mathcal{D}_2$.*

Theorems 12 and 13 resemble Theorems 1 and 2 for the observational case. While Theorem 12 provides a direct graphical tool to assess whether two DAGs are I-Markov equivalent, Theorem 13 is a technical result of key importance for *proving* score-equivalence of DAGs. Moreover, Theorem 12 does not provide a characterization of I-Markov equivalence classes through a single representative graph, as Hauser and Bühlmann [2012] do for the case of hard interventions. Nevertheless, our graphical characterization is similar to the one of perfect I-Markov equivalence offered in the same paper (Theorem 10), and which is based on sequences of post-intervention DAGs. It is thus immediate to prove the following corollary:

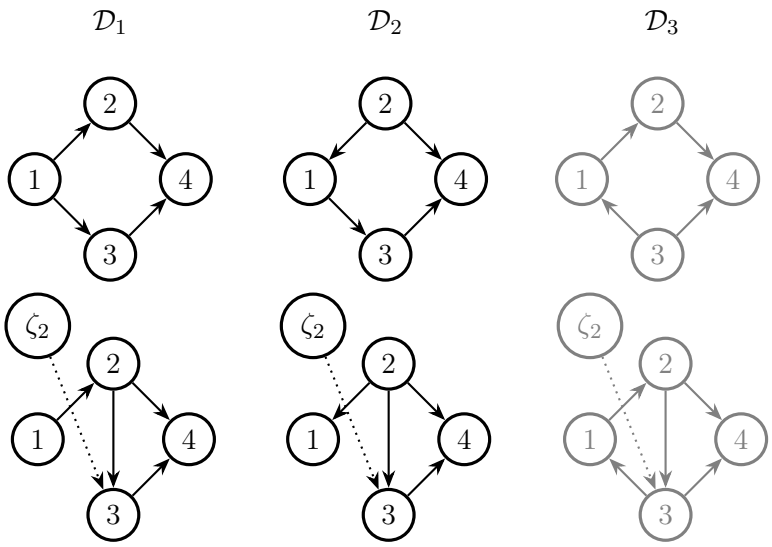
Corollary 14. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a collection of targets and induced parent sets. If $\mathcal{D}_1$ and $\mathcal{D}_2$ are I-Markov equivalent, then they are perfect I-Markov equivalent.*

Because of our validity assumption, for a given (known) $\mathcal{I}$, some DAGs may be excluded from the DAG space. We illustrate this point with an example in Figure 3.4. In such case, the general intervention defined by $T^{(2)} = 3, P^{(2)} = 2$ is valid for $\mathcal{D}_1$ and $\mathcal{D}_2$, but not for $\mathcal{D}_3$, as it would induce a cycle. Accordingly, if we consider the equivalence class defined by this intervention and assume its validity, then node 2 can not be a descendant of node 3. This implies that DAGs for which 2 is instead a descendant of 3 must be excluded from the original DAG space. While this implication may appear undesirable, it is worth noting that it only occurs when the intervention targets are *known*, and the intervention includes the addition of a new parent node. In the next section we instead consider the case of *unknown* interventions, thus avoiding the assumption of known targets and induced parent sets.

3.2.3 DAG identifiability from unknown general interventions

In the previous section we introduced I-Markov equivalence as a limit to DAG identifiability from a collection of experimental settings characterized by *known* targets and induced

Figure 3.4: *Three Markov equivalent DAGs and their post-intervention graphs after a general intervention with $T^{(2)} = \{3\}$, $P^{(2)} = \{2\}$. The intervention is not valid for $\mathcal{D}_3$.*



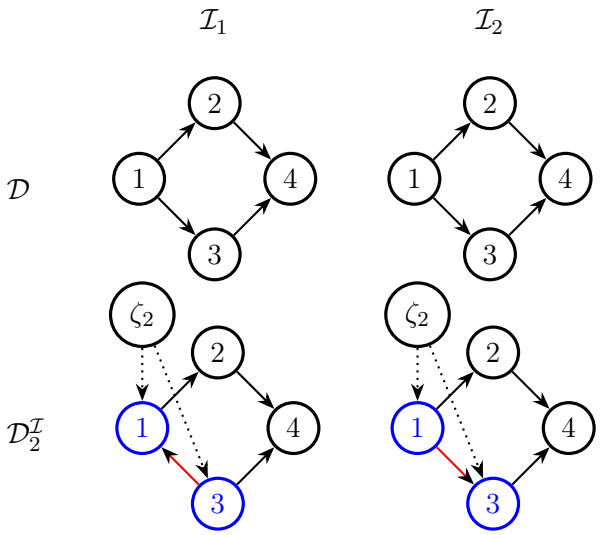
parent-sets $(\mathcal{T}, \mathcal{P})$. In this section, we consider the problem of jointly identifying the the pair $(\mathcal{D}, \mathcal{I})$ from a family of pre- and post-intervention distributions $\{p(\cdot \mid \sigma_k)\}_{k=1}^K$. The same problem has been previously investigated by Squires et al. [2020] in the context of soft interventions. The authors showed that, assuming $\mathcal{I}$-faithfulness, the DAG identifiability limit remains the same even when the targets of intervention are unknown and must be learnt from the data. Their results partially extend to our general intervention setting, but further considerations are required.

We first consider the problem of learning a general intervention from a known DAG $\mathcal{D}$ and a given family of distributions $\{p_k(\cdot)\}_{k=1}^K$. Any general intervention induces a collection of $\mathcal{I}$-DAGs that, through the I-Markov property of Definition 9, implies a set of conditional independencies and invariances. We thus investigate the limits in the identifiability of $(\mathcal{T}, \mathcal{P})$, that is whether different general interventions may imply the same set of conditional independencies and invariances given a DAG $\mathcal{D}$. With a slight abuse of terminology, we will refer to indistinguishable general interventions as I-Markov equivalent.

Definition 15. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}_1, \mathcal{I}_2$ two collections of targets and induced parent sets. $\mathcal{I}_1, \mathcal{I}_2$ are I-Markov equivalent (or, equivalently, belong to the same I-Markov equivalence class) if $\mathcal{M}_{\mathcal{I}_1}(\mathcal{D}) = \mathcal{M}_{\mathcal{I}_2}(\mathcal{D})$.*

Consider for instance the two general interventions depicted in Figure 3.5, where we have $T_1^{(2)} = T_2^{(2)} = \{1, 3\}$, $P_1^{(2)} = \{\{3\}, \emptyset\}$ and $P_2^{(2)} = \{\emptyset, \{1\}\}$. In both cases, the pre-

Figure 3.5: *Two unidintifiable combinations of DAGs and general interventions.*



and post-intervention DAGs have the same skeleta and the same set of v-structures, thus implying the same d-separation statements. As a consequence, also the conditional independencies and invariances are the same and the two general interventions are indistinguishable given data alone, differently from what occurs in the soft-intervention case of Squires et al. [2020]. We then provide the following characterizations of I-Markov equivalence of general interventions.

Theorem 16. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}_1, \mathcal{I}_2$ two collections of targets and induced parent sets. Then, $\mathcal{I}_1, \mathcal{I}_2$ belong to the same I-Markov equivalence class if and only if $\mathcal{D}_k^{\mathcal{I}_1}, \mathcal{D}_k^{\mathcal{I}_2}$ have the same skeleta and v-structures for all $k \in [K]$.*

Theorem 17. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}_1$, $\mathcal{I}_2$ two collection of targets and induced parent sets. Then, $\mathcal{I}_1$, $\mathcal{I}_2$ belong to the same I-Markov equivalence class if and only if for each $\mathcal{I}$-DAG $\mathcal{D}_k^{\mathcal{I}_1}$ there exists a sequence of edge reversals modyfing $\mathcal{D}_k^{\mathcal{I}_1}$ and such that:*

1. *Each edge reversed is covered;*
2. *After each reversal, $\mathcal{D}_k^{\mathcal{I}_1}$ is a DAG and $\mathcal{I}_1$, $\mathcal{I}_2$ belong to the same I-Markov equivalence class;*
3. *After all reversals $\mathcal{D}_k^{\mathcal{I}_1} = \mathcal{D}_k^{\mathcal{I}_2}$.*

I-Markov equivalent general interventions thus imply the same skeleta in $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$, and

in particular, the same sets of $\mathcal{I}$-edges in the augmented DAGs. This implies that the intervention targets are identifiable.

We now consider the problem of *jointly* identifying $(\mathcal{D}, \mathcal{I})$, that is the DAG and the collection of targets and induced parent sets. As before, we will use the term I-Markov equivalent to refer to indistinguishable pairs $(\mathcal{D}_1, \mathcal{I}_1)$ and $(\mathcal{D}_2, \mathcal{I}_2)$.

Definition 18. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAG and $\mathcal{I}_1, \mathcal{I}_2$ two collections of targets and induced parent sets defining a valid general intervention for $\mathcal{D}_1, \mathcal{D}_2$ respectively. $(\mathcal{D}_1, \mathcal{I}_1), (\mathcal{D}_2, \mathcal{I}_2)$ are I-Markov equivalent (or, equivalently, belong to the same I-Markov equivalence class) if $\mathcal{M}_{\mathcal{I}_1}(\mathcal{D}_1) = \mathcal{M}_{\mathcal{I}_2}(\mathcal{D}_2)$.*

As before, we can provide graphical characterizations of I-Markov equivalence for $(\mathcal{D}, \mathcal{I})$.

Theorem 19. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}_1, \mathcal{I}_2$ two collections of targets and induced parent sets defining a valid general intervention for $\mathcal{D}_1, \mathcal{D}_2$ respectively. $(\mathcal{D}_1, \mathcal{I}_1), (\mathcal{D}_2, \mathcal{I}_2)$ belong to the same I-Markov equivalence class if and only if $\mathcal{D}_{1,k}^{\mathcal{I}_1}, \mathcal{D}_{2,k}^{\mathcal{I}_2}$ have the same skeleta and v-structures for all $k \in [K]$.*

Theorem 20. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}_1$, $\mathcal{I}_2$ two collections of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1, \mathcal{D}_2$. $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ belong to the same I-Markov equivalence class if and only if there exists a sequence of edge reversals modifying the collection of $\mathcal{I}$-DAGs $\{\mathcal{D}_k^{\mathcal{I}_1}\}_{k=1}^K$ and such that:*

1. *Each edge reversed in $\mathcal{D}_1$ is simultaneously covered;*
2. *Each edge reverse in $\mathcal{D}_{1,k}^{\mathcal{I}_1}$, for $k \neq 1$, is covered;*
3. *After each reversal, $\{\mathcal{D}_{1,k}^{\mathcal{I}_1}\}_{k=1}^K$ are DAGs and $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ belong to the same I-Markov equivalence class;*
4. *After all reversals $\mathcal{D}_{1,k}^{\mathcal{I}_1} = \mathcal{D}_{2,k}^{\mathcal{I}_2}$ for each $k \in [K]$.*

As before, by Theorem 19, two distinct I-Markov equivalence pairs $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ have the same set of $\mathcal{I}$-edges, meaning that $\mathcal{T}_1 = \mathcal{T}_2$ and the targets are identifiable from the data. $\mathcal{I}_1, \mathcal{I}_2$ thus differ for their induced parent sets, and in particular for the reversal of covered edges connecting two target nodes. Note in addition that the graphical criterion of Theorem 19 is equivalent to the one of Theorem 12. As a consequence, any two unidentifiable pairs $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ imply the same set of conditional independencies and invariances via the I-Markov property and in particular the same as if the general intervention were known. The DAG-identifiability thus remains the same as for the known intervention case.

3.3 Bayesian causal discovery

In this section we introduce a parametric Bayesian framework for the analysis of data collected under general unknown interventions. In Section 3.3.1 we frame the related causal discovery problem under the Bayesian perspective, and specify a likelihood function that integrates data from distinct interventional contexts. In Section 3.3.2 we then introduce a prior elicitation procedure for the collection of model parameters. Finally, in Section 3.3.3 we assign prior distributions to DAGs, intervention targets and parent sets, whose posterior inference represents the ultimate goal of our Bayesian methodology.

3.3.1 Model formulation

Let $\boldsymbol{X} = \big(\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(K)}\big)^\top$ be the (n, q) data matrix, where $\boldsymbol{X}^{(k)}$ is the (n_k, q) dataset containing samples collected under the k-th experimental setting. As in the previous sections, we assume $\boldsymbol{X}^{(1)}$ being an observational dataset, so that $T^{(1)} = P^{(1)} = \emptyset$ and $\mathcal{D}_1 = \mathcal{D}$. Under the Bayesian setting, learning the pair $(\mathcal{D}, \mathcal{I})$ can be framed as a model selection problem which requires the computation of the posterior distribution

$$p\left(\mathcal{D}, \mathcal{I} \,|\, \boldsymbol{X}\right) \propto p\left(\boldsymbol{X} \,|\, \mathcal{D}, \mathcal{I}\right) p(\mathcal{D}, \mathcal{I}). \tag{3.5}$$

We refer to $p(\mathcal{D}, \mathcal{I})$ as the *model prior* and to $p(\boldsymbol{X} \,|\, \mathcal{D}, \mathcal{I})$ as the *model evidence* or *marginal likelihood*. Assuming a parametric family of distributions for the observables, we can write the marginal likelihood as

$$p\big(\boldsymbol{X} \,|\, \mathcal{D}, \mathcal{I}\big) = \int p\big(\boldsymbol{X} \,|\, \Theta^{(\mathcal{K})}, \mathcal{D}, \mathcal{I}\big)\, p\big(\Theta^{(\mathcal{K})} \,|\, \mathcal{D}, \mathcal{I}\big)\, d\Theta^{(\mathcal{K})}, \tag{3.6}$$

where $\Theta^{(\mathcal{K})} = \{\Theta^{(1)}, \dots, \Theta^{(K)}\}$ is the multi-set of parameters associated with the pre- and post-intervention distributions implied by the pair $(\mathcal{D}, \mathcal{I})$. Conditionally on $\Theta^{(\mathcal{K})}$, the observations in $\boldsymbol{X}$ are independent and, within each block $\boldsymbol{X}^{(k)}$, identically distributed, so that the likelihood function can be written as

$$p\big(\boldsymbol{X} \,|\, \Theta^{(\mathcal{K})}, \mathcal{D}, \mathcal{I}\big) = \prod_{k=1}^{K} p\big(\boldsymbol{X}^{(k)} \,|\, \Theta^{(k)}, \mathcal{D}, I^{(k)}\big), \tag{3.7}$$

where $I^{(k)} = (T^{(k)}, P^{(k)})$ and $\Theta^{(k)}$ is the set of parameters of the distribution of the k-th experimental setting. From Definition 9, the I-Markov property implies that: i) the *sampling distribution* of the i-th observation in the k-th block factorises according to the

post-intervention DAG $\mathcal{D}_k$; ii) a set of invariances hold, such that the post-intervention local parameters indexing the non-intervened nodes are equal to the corresponding pre-intervention parameters. From these considerations, it follows that

$$p\big(\boldsymbol{X} \,|\, \Theta^{(\mathcal{K})}, \mathcal{D}, \mathcal{I}\big) = \prod_{j=1}^{q} \Bigg\{ p\big(\boldsymbol{X}_{\cdot j}^{\mathcal{A}(j)} \,|\, \boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)}, \Theta_j^{(1)}, \mathcal{D}\big) \prod_{k : j \in T^{(k)}} p\big(\boldsymbol{X}_{\cdot j}^{(k)} \,|\, \boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(j)}^{(k)}, \Theta_j^{(k)}, \mathcal{D}_k\big) \Bigg\}, \tag{3.8}$$

where $\Theta_j^{(k)}$ is the j-th element of $\Theta^{(k)}$, and we used the equivalent representation of $(\mathcal{D}, \mathcal{I}^{(k)})$ in terms of modified DAG $\mathcal{D}_k$. Moreover, $\mathcal{A}(j) := \{k : j \notin T^{(k)}\}$ is the collection of interventional settings under which node j has not been intervened upon, and $\boldsymbol{X}_{\cdot B}^{\mathcal{A}(j)}$ is the sub-matrix of $\boldsymbol{X}$ with columns indexed by $B \subset [q]$ and blocks corresponding to $\mathcal{A}(j) \subset [K]$. To obtain (3.5) we thus need to specify:

1. A *statistical model* $p\left(\boldsymbol{X} \,|\, \Theta^{(\mathcal{K})}, \mathcal{D}, \mathcal{I}\right)$, in the form of a distribution for the data in Equation 3.8;

2. A *model prior* $p(\mathcal{D}, \mathcal{I})$, describing our prior knowledge on DAG $\mathcal{D}$ and on the effects that the interventions imply on its structure;

3. A *parameter prior* $p(\Theta^{(\mathcal{K})} \,|\, \mathcal{D}, \mathcal{I})$ leading, once combined with the likelihood (3.8), to the marginal likelihood (3.6).

The joint specification of a statistical model and associated parameter prior deserves particular attention and is the main subject of the next section.

3.3.2 Parameter prior elicitation

Under common distributional assumptions (e.g. Gaussian), it is not possible to distinguish between DAGs belonging to the same I-Markov equivalence class [Hauser and Bühlmann, 2012]. In a Bayesian model-selection framework, this feature translates into the compatibility requirement that I-Markov equivalent DAGs are assigned equal marginal likelihoods, a property usually referred to as *score equivalence*. In this section we show how the procedure proposed by Geiger and Heckerman [2002] for DAG model selection from observational data can be extended to our interventional setting. Their methodology relies on a set of assumptions (Assumptions 1-5 in the original paper) that translate into our setting as follows:

A1 *(Complete model equivalence and regularity)*: Let $\mathcal{C}$ be the collection of complete DAGs on the set of nodes V, each implying a statistical model $p(x \mid \Theta_C, C)$, for $C \in \mathcal{C}$. For any two complete DAGs $C_i, C_j \in \mathcal{C}, i \neq j$, we have that $p(x \mid \Theta_{C_i}, C_i) = p(x \mid \Theta_{C_j}, C_j)$. Moreover, there exists a one-to-one mapping $\kappa_{i,j}$ between the DAG-parameters $\Theta_{C_i}, \Theta_{C_j}$ such that $\Theta_{C_j} = \kappa_{i,j}(\Theta_{C_i})$ and the Jacobian $|\partial\Theta_{C_i}/\partial\Theta_{C_j}|$ exists and is nonzero for all values of Θ_{C_i};

A2 *(Likelihood and prior modularity)*: For any two DAGs $\mathcal{D}_i, \mathcal{D}_j$ and any node $l \in V$ such that $\mathrm{pa}_{\mathcal{D}_i}(l) = \mathrm{pa}_{\mathcal{D}_j}(l)$, we have that, for any collection of targets and induced parent sets $\mathcal{I}$,

$$p\big(x_l^{(k)} \mid x^{(k)}_{\mathrm{pa}_{\mathcal{D}_{i,k}}(l)}, \Theta_l^{(k)}, \mathcal{D}_{i,k}\big) = p\big(x_l^{(k)} \mid x^{(k)}_{\mathrm{pa}_{\mathcal{D}_{j,k}}(l)}, \Theta_l^{(k)}, \mathcal{D}_{j,k}\big),$$

$$p\big(\Theta_l^{(k)} \mid \mathcal{D}_{i,k}\big) = p\big(\Theta_l^{(k)} \mid \mathcal{D}_{j,k}\big);$$

A3 *(Global parameter independence)*: For every DAG $\mathcal{D}$ and any collection of targets and induced parent sets $\mathcal{I}$,

$$p\big(\Theta^{(\mathcal{K})} \mid \mathcal{D}, \mathcal{I}\big) = \prod_{j=1}^{q} \left\{ p\big(\Theta_j^{(1)} \mid \mathcal{D}\big) \prod_{k : j \in I^{(k)}}^{K} p\big(\Theta_j^{(k)} \mid \mathcal{D}_k\big) \right\}.$$

We refer the reader to Geiger and Heckerman [2002] for a detailed discussion of these assumptions in the observational setting. Most importantly for our purposes, given Assumption **A3**, we can specify priors for the parameters indexing each term in (3.8) independently. The following procedure is therefore applied to each node $j \in V$ and experimental context $k \in [K]$:

(i) Identify a complete DAG $C_{j,k}$ such that $\mathrm{pa}_{C_{j,k}}(j) = \mathrm{pa}_{\mathcal{D}_k}(j)$;

(ii) Assign a prior to $\Theta_{C_{j,k}}$, the parameter of the selected complete DAG model $C_{j,k}$;

(iii) Assign to $\Theta_j^{(k)}$ the same prior assigned to $\Theta_{j,C_{j,k}}$ in step **(ii)**, where $\Theta_{j,C_{j,k}} \in \Theta_{C_{j,k}}$ is the parameter indexing the j-th node.

Accordingly, because of Assumption **A1**, the proposed procedure allows to specify a parameter prior for any pair $(\mathcal{D}, \mathcal{I})$ from a single parameter prior on a complete DAG model C. Therefore, the marginal likelihood $p(\boldsymbol{X} \mid \mathcal{D}, \mathcal{I})$ can be computed as in the following proposition.

Proposition 21. *Given any complete DAG C and a data matrix $\boldsymbol{X}$ collecting observations from K different experimental settings, for any valid pair $(\mathcal{D}, \mathcal{I})$ Assumptions **A1-A3** imply*

$$p\left(\boldsymbol{X} \mid \mathcal{D}, \mathcal{I}\right) = \prod_{j=1}^{q} \left\{ \frac{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot fa_{\mathcal{D}}(j)} \mid C\right)}{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot pa_{\mathcal{D}}(j)} \mid C\right)} \prod_{k: j \in I^{(k)}} \frac{p\left(\boldsymbol{X}^{(k)}_{\cdot fa_{\mathcal{D}_k}(j)} \mid C\right)}{p\left(\boldsymbol{X}^{(k)}_{\cdot pa_{\mathcal{D}_k}(j)} \mid C\right)} \right\}, \tag{3.9}$$

where $p\big(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot B} \mid C\big)$ is the marginal data distribution computed under any complete DAG C.

Notice that the resulting marginal likelihood provides a *decomposable* score for the pair $(\mathcal{D}, \mathcal{I})$, since it corresponds to a product of q terms each involving a node j and its parents $\mathrm{pa}_{\mathcal{D}_k}(j)$ in each DAG $\mathcal{D}_k$ only. Importantly, it also guarantees score equivalence for I-Markov equivalent pairs $(\mathcal{D}, \mathcal{I})$.

Theorem 22 (Score equivalence)**.** *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}_1, \mathcal{I}_2$ two collections of targets and induced parent sets defining a valid general intervention for $\mathcal{D}_1, \mathcal{D}_2$ respectively. If $(\mathcal{D}_1, \mathcal{I}_1)$ and $(\mathcal{D}_2, \mathcal{I}_2)$ are I-Markov equivalent, then Assumptions A1-A3 imply*

$$p(\boldsymbol{X} \mid \mathcal{D}_1, \mathcal{I}_1) = p(\boldsymbol{X} \mid \mathcal{D}_2, \mathcal{I}_2). \tag{3.10}$$

3.3.3 Prior on $(\mathcal{D}, \mathcal{I})$

Recall that $\mathcal{I} = (\mathcal{T}, \mathcal{P})$, where $\mathcal{T} = \{T^{(k)}\}_{k=1}^{K}$ and $\mathcal{P} = \{P^{(k)}\}_{k=1}^{K}$. For convenience, we represent the (possibly) different parent sets induced by the K experimental settings, $\mathcal{P}$, through K (q, q) matrices $\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(K)}$ such that for any (l, j)-element $\boldsymbol{P}^{(k)}_{lj}$ we have $\boldsymbol{P}^{(k)}_{lj} = 1$ if $l \to j \in \mathcal{D}_k$ and $j \in T^{(k)}$, 0 otherwise. Conditionally on DAG $\mathcal{D}$ and target $T^{(k)}$, we assume independently across $k \in \{2, \ldots, K\}$,

$$\begin{aligned} p\big(\boldsymbol{P}^{(k)} \mid \boldsymbol{\phi}^{(k)}, T^{(k)}, \mathcal{D}\big) \;&=\; \left\{ \prod_{j=1}^{q} \prod_{j \in T^{(k)}} \mathrm{pBern}\big(\boldsymbol{P}^{(k)}_{lj} \mid \phi^{(k)}_j\big) \right\} \mathbb{1}\left\{\mathcal{D}_k \text{ is a DAG}\right\} \\ \phi^{(k)}_j \;&\overset{\text{iid}}{\sim}\; \mathrm{Beta}\big(a_\phi, b_\phi\big), \quad j \in T^{(k)}, \end{aligned} \tag{3.11}$$

where $\boldsymbol{\phi}^{(k)} = \big\{\phi^{(k)}_j\big\}_{j \in T^{(k)}}$. The hierarchical prior (3.11) leads to the marginal (integrated w.r.t. $\boldsymbol{\phi}^{(k)}$) prior on $\boldsymbol{P}^{(k)}$

$$p\big(\boldsymbol{P}^{(k)} \mid T^{(k)}, \mathcal{D}\big) \;=\; \left\{ \prod_{j \in T^{(k)}} \frac{\mathcal{B}\left(a_\phi + |\boldsymbol{P}^{(k)}_{\cdot j}|, b_\phi + q - |\boldsymbol{P}^{(k)}_{\cdot j}|\right)}{\mathcal{B}\big(a_\phi, b_\phi\big)} \right\} \mathbb{1}\left\{\mathcal{D}_k \text{ is a DAG}\right\},$$

where $|\boldsymbol{P}_{\cdot j}^{(k)}| = \sum_{l=1}^{q} \boldsymbol{P}_{lj}^{(k)}$ and $\mathcal{B}(\cdot)$ denotes the Beta function.
Now consider $T^{(k)}$, the intervention target associated with the experimental setting k. We represent $T^{(k)} \subseteq [q]$ through a $(q, 1)$ vector $\boldsymbol{h}_k$ whose j-th element $h_k(j)$ is equal to 1 if $j \in T^{(k)}$, 0 otherwise. We assume, independently across $k \in \{2, \dots, K\}$,

$$\begin{aligned} p\left(\boldsymbol{h}_k \,|\, \eta_k\right) &= \prod_{j=1}^{q} \text{pBern}\left(h_k(j) \,|\, \eta_k\right) \\ \eta_k &\sim \text{Beta}\left(a_\eta, b_\eta\right). \end{aligned} \tag{3.12}$$

Equation (3.12) leads to the integrated prior on $T^{(k)}$

$$p\big(T^{(k)}\big) = p(\boldsymbol{h}_k) = \frac{\mathcal{B}\big(a_\eta + |T^{(k)}|, b_\eta + q - |T^{(k)}|\big)}{\mathcal{B}\big(a_\eta, b_\eta\big)},$$

where $|T^{(k)}| = \sum_{j=1}^{q} h_k(j)$ is the number of intervened nodes in context k.
Finally, let $\mathcal{S}_q$ be the set of all DAGs with q nodes. We assign a prior to $\mathcal{D} \in \mathcal{S}_q$ through a Beta-Binomial distribution on the number of edges in the graph. Specifically, let $\boldsymbol{S}^{\mathcal{D}}$ be the adjacency matrix of the skeleton of $\mathcal{D}$, and $\boldsymbol{S}_{lj}^{\mathcal{D}}$ its (l, j)-element. We assign

$$\begin{aligned} p\big(\boldsymbol{S}^{\mathcal{D}} \,|\, \omega\big) &= \prod_{l<j} \text{pBern}\big(\boldsymbol{S}_{lj}^{\mathcal{D}} \,|\, \omega\big) \\ \omega &\sim \text{Beta}(a_{\mathcal{D}}, b_{\mathcal{D}}), \end{aligned} \tag{3.13}$$

leading to

$$p(\boldsymbol{S}^{\mathcal{D}}) = \frac{\mathcal{B}\big(a_{\mathcal{D}} + |\boldsymbol{S}^{\mathcal{D}}|, b_{\mathcal{D}} + q(q-1)/2 - |\boldsymbol{S}^{\mathcal{D}}|\big)}{\mathcal{B}\big(a_{\mathcal{D}}, b_{\mathcal{D}}\big)},$$

where $|\boldsymbol{S}^{\mathcal{D}}|$ is the number of edges in $\mathcal{D}$ (equivalently in its skeleton) and $q(q-1)/2$ is the maximum number of edges in a DAG on q nodes. Finally, we set $p(\mathcal{D}) \propto p(\boldsymbol{S}^{\mathcal{D}})$ for each $\mathcal{D} \in \mathcal{S}_q$.

3.4 MCMC scheme and posterior inference

In this section we describe the Markov Chain Monte Carlo (MCMC) strategy that we adopt to approximate the posterior distribution (3.5). Specifically, Section 3.4.1 introduces the random scan Metropolis-Hastings algorithm which is at the basis of our sampler, while Section 3.4.2 illustrates how the MCMC output can be used to provide estimates of the

underlying causal DAG structure and the effects of the general interventions.

3.4.1 Sampling scheme

Our MCMC algorithm has the structure of a random-scan component-wise Metropolis-Hastings [Brooks et al., 2011, Chapter 1], in which the parameter of interest is partitioned into K components, each indexing one of the K experimental settings. Specifically, the first component corresponds to the DAG $\mathcal{D}$, while the remaining ones to the collection of unknown targets and induced parent sets $I^{(k)} = (T^{(k)}, P^{(k)})$ for $k \in \{2, \dots, K\}$. Sampling from each component occurs in a random order through standard proposal and acceptance/rejection steps as in a Metropolis-Hastings sampler. A high-level illustration of the scheme is provided in Algorithm 3.5.

Our main algorithm adopts the equivalent representation of $(\mathcal{D}, \mathcal{I})$ in terms of $\mathcal{I}$-DAGs $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$. In this way, it is possible to explore the space of possible pairs $(\mathcal{D}, \mathcal{I})$ using a set of simple operators inducing local modifications on DAGs. Specifically, we consider three types of operators: $Insert(u, v), Delete(u, v)$, and $Reverse(u, v)$, corresponding respectively to the insertion, deletion, and reversal of the edge (u, v). Also notice that the modified graph obtained by applying any of these operators may not be a DAG. Accordingly, we impose to the operators above the following *validity* requirement (**vr**).

Definition 23. *Let $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$ be a sequence of $\mathcal{I}$-DAGs. An operator inducing a sequence of modified $\mathcal{I}$-DAGs $\{\widetilde{\mathcal{D}}_k^{\mathcal{I}}\}_{k=1}^K$ is* valid *if every graph in $\{\widetilde{\mathcal{D}}_k^{\mathcal{I}}\}_{k=1}^K$ is a DAG.*

Let now $\mathcal{O}_{\mathcal{D}}$ be the set of all valid operators on DAG $\mathcal{D}$. Our proposal distribution draws randomly an operator in $\mathcal{O}_{\mathcal{D}}$, and then apply it to $\mathcal{D}$ to obtain $\widetilde{\mathcal{D}}$. Accordingly, the (proposal) probability of a transition from $\mathcal{D}$ to $\widetilde{\mathcal{D}}$ is $q(\widetilde{\mathcal{D}} \,|\, \mathcal{D}) = 1/|\mathcal{O}_{\mathcal{D}}|$, where $|\mathcal{O}_{\mathcal{D}}|$ is the number of elements in $\mathcal{O}_{\mathcal{D}}$. We use the same proposal scheme for the update of $\mathcal{D}_k^{\mathcal{I}}$. Notice however that the same operator may imply different modifications when applied to the observational DAG $\mathcal{D}$ or to an $\mathcal{I}$-DAG $\mathcal{D}_k^{\mathcal{I}}$. In the former case, the implied modification also affects all the $\mathcal{I}$-DAGs; in the latter case, the effect is local and refers corresponding to the $\mathcal{I}$-DAG indexing the k-th experimental setting. Accordingly, we need a different construction for the set of operators relative to the observational and experimental components. For the former case, Algorithm 4 constructs the set $\mathcal{O}_{\mathcal{D}}$ simply by considering all possible valid insertions, deletions, and reversals of the edges of the observational DAG. Differently, for the latter case, Algorithm 5 includes in $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$ all the operators implying: i) the insertion of an intervention target, ii) the modification of the parent sets of a target node and iii) the deletion of an intervention target (provided that the parents of the target in the DAG and in the $\mathcal{I}$-DAG are the same).

Algorithm 3: Random-scan MH for posterior inference

Input: Data matrix $\boldsymbol{X}$, number of MCMC iterations S, initial values for DAG, targets and induced parent sets $\mathcal{D}^0, \mathcal{T}^0, \mathcal{P}^0$

Output: S samples from $p(\mathcal{D}, \mathcal{T}, \mathcal{P} \mid \boldsymbol{X})$

Construct $\{\mathcal{D}_k^{0\mathcal{I}}\}_{k=1}^K$;
Set $\mathcal{I}^0 = (\mathcal{T}^0, \mathcal{P}^0)$;
for *s in 1:S* **do**
  Sample $\boldsymbol{\pi}$, a permutation vector of length K;
  Set $\{\mathcal{D}^s, \mathcal{I}^s\} = \{\mathcal{D}^{s-1}, \mathcal{I}^{s-1}\}$;
  for *k in 1:K* **do**
    if $\boldsymbol{\pi}_k = 1$ **then**
      Construct $\mathcal{O}_{\mathcal{D}^s}$ using Algorithm 4;
      Propose $\widetilde{\mathcal{D}}$ by sampling uniformly at random from $\mathcal{O}_{\mathcal{D}^s}$;
      Set $\mathcal{D}^s = \widetilde{\mathcal{D}}$ with probability

$$\alpha_{\widetilde{\mathcal{D}}} = \min\left\{1; \frac{p\big(\boldsymbol{X} \mid \widetilde{\mathcal{D}}, \{I_s^{(j)}\}_{j \neq \boldsymbol{\pi}_k}\big)}{p\big(\boldsymbol{X} \mid \mathcal{D}^s, \{I_s^{(j)}\}_{j \neq \boldsymbol{\pi}_k}\big)} \cdot \frac{p(\widetilde{\mathcal{D}})}{p(\mathcal{D}^s)} \cdot \frac{q(\mathcal{D}^s \mid \widetilde{\mathcal{D}})}{q(\widetilde{\mathcal{D}} \mid \mathcal{D}^s)}\right\}$$

    end
    else
      Construct $\mathcal{O}_{\mathcal{D}^{s\mathcal{I}}_{\boldsymbol{\pi}_k}}$ using Algorithm 5;
      Propose $\widetilde{\mathcal{D}}^{\mathcal{I}}_{\boldsymbol{\pi}_k}$ by sampling uniformly at random from $\mathcal{O}_{\mathcal{D}^{s\mathcal{I}}_{\boldsymbol{\pi}_k}}$;
      Recover $\widetilde{I}^{(\boldsymbol{\pi}_k)} = (\widetilde{T}^{(\boldsymbol{\pi}_k)}, \widetilde{P}^{(\boldsymbol{\pi}_k)})$ from $(\widetilde{\mathcal{D}}^{\mathcal{I}}_{\boldsymbol{\pi}_k}, \mathcal{D}^s)$;
      Set $I_s^{(\boldsymbol{\pi}_k)} = \widetilde{I}^{(\boldsymbol{\pi}_k)}$ with probability

$$\alpha_{\widetilde{e}_{\boldsymbol{\pi}_k}} = \min\left\{1; \frac{p\big(\boldsymbol{X} \mid \mathcal{D}^s, \{I_s^{(j)}\}_{j \neq \boldsymbol{\pi}_k}, \widetilde{I}^{(\boldsymbol{\pi}_k)}\big)}{p\big(\boldsymbol{X} \mid \mathcal{D}^s, \{I_s^{(j)}\}_{j \neq \boldsymbol{\pi}_k}, I_s^{(\boldsymbol{\pi}_k)}\big)} \cdot \frac{p\big(\widetilde{I}^{(\boldsymbol{\pi}_k)}\big)}{p\big(I_s^{(\boldsymbol{\pi}_k)}\big)} \cdot \frac{q\big(\mathcal{D}_k^{s\mathcal{I}} \mid \widetilde{\mathcal{D}}_k^{\mathcal{I}}\big)}{q\big(\widetilde{\mathcal{D}}_k^{\mathcal{I}} \mid \mathcal{D}_k^{s\mathcal{I}}\big)}\right\}$$

    end
  end
end
Recover $\{\mathcal{T}^s, \mathcal{P}^s\}_{s=1}^S$ from $\{\mathcal{I}^s\}_{s=1}^S$;
return $\{\mathcal{D}^s, \mathcal{T}^s, \mathcal{P}^s\}_{s=1}^S$;

Algorithm 4: Random-scan MH: Construction of $\mathcal{O}_{\mathcal{D}}$

Input: A collection of $\mathcal{I}$-DAGs $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$
Output: A set of valid operators $\mathcal{O}_{\mathcal{D}}$

Set $\mathcal{O}_{\mathcal{D}} = \emptyset$;
Construct $E_I = \{(u, v) :_{uv} =_{vu} = 0\}$;
Construct $E_D = \{(u, v) :_{uv} = 1\}$;
for $e \in E_D$ **do**
  Add $Delete(e)$ to $\mathcal{O}_{\mathcal{D}}$;
  if $Reverse(e)$ *satisfies* ***vr*** **then** add it to $\mathcal{O}_{\mathcal{D}}$;
end
for $e \in E_I$ **do**
  if $Insert(e)$ *satisfies* ***vr*** **then** add it to $\mathcal{O}_{\mathcal{D}}$;
end
return $\mathcal{O}_{\mathcal{D}}$;

Algorithm 5: Random-scan MH: Construction of $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$

Input: A collection of $\mathcal{I}$-DAGs $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$
Output: A set of valid operators $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$

Set $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}} = \emptyset$;
Recover $(T^{(k)}, P^{(k)})$ from $(\mathcal{D}, \mathcal{D}_k^{\mathcal{I}})$;
for $v \notin T^{(k)}$ **do**
  Add $Insert(\zeta_k, v)$ to $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$;
end
for $v \in T^{(k)}$ **do**
  for $u \in nd_{\mathcal{D}_k}(v)$ **do**
    if $u \in pa_{\mathcal{D}_k}(v)$ **then**
      Add $Delete(u, v)$ to $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$;
      if $Reverse(u, v)$ *satisfies* ***vr*** *and* $u \in T^{(k)}$ **then**
        Add $Reverse(u, v)$ to $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$;
      end
    end
    else
      Add $Insert(u, v)$ to $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$;
    end
    if $pa_{\mathcal{D}_k}(v) = pa_{\mathcal{D}}(v)$ **then** add $Delete(\zeta_k, v)$ to $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$;
  end
end
return $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$;

The proposal distributions defined above are of key importance to ensure that the Markov chain implied by the Metropolis-Hastings is reversible, aperiodic and irreducible, so that the MCMC scheme provides an approximation of the posterior distribution, as stated in the following proposition.

Proposition 24. *The finite Markov chain defined by Algorithm 3, 4, and 5 is reversible, aperiodic, and irreducible. Accordingly, it has $p(\mathcal{D}, \mathcal{T}, \mathcal{P} \mid \boldsymbol{X})$ as its unique stationary distribution.*

3.4.2 Posterior inference

Output of Algorithm 3 consists of a sample of size S from the posterior distribution $p(\mathcal{D}, \mathcal{T}, \mathcal{P} \mid \boldsymbol{X})$. This MCMC output can be used to obtain summaries of specific features of the posterior distribution, such as DAG structures, both corresponding to the observational distribution of the variables, or a post-intervention distribution (represented by a modified DAG), as well as identifying the targets and parent sets induced by the interventions.

Point estimates of a DAG structure can be recovered through a Maximum A Posteriori (MAP) DAG estimate, corresponding to the DAG with the highest posterior probability, or based on the so-called Median Probability Model (MPM) originally introduced by Barbieri and Berger [2004] in the linear regression setting. To obtain an MPM-based estimate of a DAG we need to compute first a collection of marginal Posterior Probabilities of edge Inclusion (PPIs) for each possible directed link (u, v) in any DAG $\mathcal{D}_k$. Each corresponds to the (u, v)-element of a (q, q) matrix $\boldsymbol{J}^{(k)}$,

$$\boldsymbol{J}^{(k)}_{uv} = \widehat{p}(u \to v \in \mathcal{D}_k \mid \boldsymbol{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\{u \to v \in \mathcal{D}_k^s\}, \tag{3.14}$$

where $\mathcal{D}_k^s$ is the modified DAG of context k visited at iteration s. When $k = 1$ the above matrix collects the PPIs relative to $\mathcal{D}$, the DAG indexing the observational distribution of the q variables. An MPM DAG estimate, $\widehat{\mathcal{D}}_k$, for each $k = [K]$, is finally obtained by including those edges whose PPIs is greater than 0.5.

Now consider the intervention targets $T^{(1)}, \ldots, T^{(K)}$. We can recover a marginal posterior probability of inclusion for a node $j \in [q]$ in the target $T^{(k)}$, $k \in \{2, \ldots, K\}$, as

$$\boldsymbol{T}^{(k)}_j = \widehat{p}(j \in T^{(k)}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\{j \in T^{(k)}_s\}, \tag{3.15}$$

while by definition $\boldsymbol{T}^{(1)}_j = 0$ for each j. The resulting collection of probabilities is organized

in a (q, K) matrix $\boldsymbol{T}$ with (k, j)-element corresponding to $\boldsymbol{T}_j^{(k)}$. As a point summary of the posterior distribution of $T^{(k)}$, we again consider a median-probability based estimate $\widehat{\boldsymbol{T}}^{(k)}$ such that, for each $j \in [q]$, $\widehat{\boldsymbol{T}}^{(k)} = 1$ if $\boldsymbol{T}_j^{(k)} \geq 0.5$, 0 otherwise.

A useful feature of our method is that it can be adopted to detect differences between experimental contexts that are reflected into modifications of the DAG structure, as induced by the interventions. These can be represented by means of a *difference-graph* [Wang et al., 2018] which is constructed as follows. Consider the two DAGs $\mathcal{D}_1$ and $\mathcal{D}_k$, for $k \in \{2, \ldots, K\}$ Let also $T^{(k)}$ be the intervention target associated with $\mathcal{D}_k$. The difference-graph of $(\mathcal{D}_1, \mathcal{D}_k)$, denoted as $\mathcal{G}^{(k)}$, is the graph whose adjacency matrix $\boldsymbol{G}^{(k)}$ has (u, v)-element

$$\boldsymbol{G}_{uv}^{(k)} = \begin{cases} 1 & \text{if } v \in T^{(k)} \text{ and } u \in \{\text{pa}_{\mathcal{D}_1}(v) \cup \text{pa}_{\mathcal{D}_k}(v)\}, \\ 0 & \text{otherwise.} \end{cases}$$

In other terms, an edge $u \to v$ is included in $\mathcal{G}^{(k)}$ whenever v is an intervention target and u is a parent of v in at least on of the two DAGs, implying that the local distribution of node v has been modified as the effect of a (soft or general) intervention. For any $\mathcal{G}^{(k)}$ we can provide an MCMC-based estimate, $\widehat{\mathcal{G}}^{(k)}$ by following the same rationale leading to the MPM DAG and based on the collection of estimated PPIs.

3.5 Simulations and real data analysis

In this section we apply our methodology for structure learning under general interventions to simulated and real data. To this end, in Section 3.5.1 we first specialize our framework to Gaussian DAG models. In Section 3.5.2 we thus evaluate its performance on simulated Gaussian data and compare it with alternative benchmark approaches. Finally, in Section 3.5.3 we present an application to biological protein expression data.

3.5.1 Gaussian DAGs

For the random vector $X = (X_1, \ldots, X_q)^\top$, we consider a linear Gaussian Structural Equation Model (SEM) of the form

$$X = \boldsymbol{B}^\top X + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D}), \tag{3.16}$$

where $\boldsymbol{B}$ is a (q, q) matrix of regression coefficients with (l, j)-element $\boldsymbol{B}_{lj} \neq 0$ if and only if $l \in \text{pa}_{\mathcal{D}}(j)$, and $\boldsymbol{D} = \text{diag}(\boldsymbol{D}_{11}, \ldots, \boldsymbol{D}_{qq})$ is a (q, q) matrix collecting the conditional

variances of the q variables. Equivalently, we can write for each $j \in [q]$

$$X_j = \sum_{l \in \mathrm{pa}_{\mathcal{D}}(j)} \boldsymbol{B}_{lj} X_l + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \boldsymbol{D}_{jj}). \tag{3.17}$$

Equation (3.16) implies $X \,|\, \boldsymbol{\Sigma}, \mathcal{D} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{B})^{-\top} \boldsymbol{D} (\boldsymbol{I} - \boldsymbol{B})^{-1}$, the right-hand side corresponding to the modified Cholesky decomposition of the covariance matrix. Consider now a family of experimental settings with intervention targets $T^{(1)}, \ldots, T^{(K)}$ and implied modified DAGs $\mathcal{D}_1, \ldots, \mathcal{D}_K$. For each $k = [K]$ we have

$$X_j = \sum_{l \in \mathrm{pa}_{\mathcal{D}_k}(j)} \boldsymbol{B}^{(k)}_{lj} X_l + \varepsilon^{(k)}_j, \quad \varepsilon^{(k)}_j \sim \mathcal{N}\big(0, \boldsymbol{D}^{(k)}_{jj}\big), \quad j \in T^{(k)}, \tag{3.18}$$

where $(\boldsymbol{B}^{(k)}, \boldsymbol{D}^{(k)})$ are the DAG-parameters induced by the general intervention. Notice that all the (l, j)-elements of $(\boldsymbol{B}^{(k)}, \boldsymbol{D}^{(k)})$ not involved in (3.18) are exactly those in $(\boldsymbol{B}, \boldsymbol{D})$ because of the assumed invariances between pre- and post-intervention distributions (see Equations (3.4) and (3.8)). For each experimental setting $k \in [K]$, the post-intervention joint distribution of X is then $X \,|\, \boldsymbol{\Sigma}_k, \mathcal{D}_k \sim \mathcal{N}_q\big(\mathbf{0}, \boldsymbol{\Sigma}_k\big)$, where $\boldsymbol{\Sigma}_k = \big(\boldsymbol{I} - \boldsymbol{B}^{(k)}\big)^{-\top} \boldsymbol{D}^{(k)} \big(\boldsymbol{I} - \boldsymbol{B}^{(k)}\big)^{-1}$. Because of the prior elicitation procedure introduced in Section 3.3.2, to compute the DAG marginal likelihood (3.9) we only need to specify a prior on the parameter of a complete (unconstrained) Gaussian DAG model. It is immediate to show that assumptions **A1-A3** of Section 3.3.2 are satisfied in the Gaussian setting by $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{U})$, namely a Wishart distribution on $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ having expectation $a\boldsymbol{U}^{-1}$ with $a > q - 1$ and $\boldsymbol{U}$ a (q, q) s.p.d. matrix. By combining such prior with the likelihood of n i.i.d. samples from $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, we obtain the following formula for the marginal data distribution relative to any subset of the q variables $B \subset [q]$:

$$p(\boldsymbol{X}_{\cdot B}) = \pi^{-\frac{n|B|}{2}} \frac{|\boldsymbol{U}_{BB}|^{\frac{a-|\bar{B}|}{2}}}{|\widetilde{\boldsymbol{U}}_{BB}|^{\frac{a-|\bar{B}|+n}{2}}} \frac{\Gamma_{|B|}\left(\frac{a-|\bar{B}|+n}{2}\right)}{\Gamma_{|B|}\left(\frac{a-|\bar{B}|}{2}\right)}, \tag{3.19}$$

where $\bar{B} = [q] \backslash B$ and $\widetilde{\boldsymbol{U}} = \boldsymbol{U} + \boldsymbol{X}^\top \boldsymbol{X}$; see for instance Press [2012]. This formula, implemented in Equation (3.9) for suitable elements (rows and columns) of the data matrix $\boldsymbol{X} = \big(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}\big)^\top$, specializes the DAG marginal likelihood to the Gaussian setting. Note that the resulting marginal likelihood provides an adaptation to our interventional setting of the popular Bayesian Gaussian equivalent (BGe) score, originally introduced by Heckerman and Geiger [1995] for the case of i.i.d. observational data; see also Geiger and Heckerman [2002]. When coupled with the model prior introduced in Section 3.3.3, this

result fully specializes our general methodology to the Gaussian setting.

3.5.2 Simulation studies

We evaluate the performance of our method under several simulated scenarios where we vary i) the number of experimental settings $K \in \{2, 4\}$, ii) the number of variables $q \in \{10, 20\}$ and iii) the sample size $n_k \in \{100, 500, 1000\}$ that we assume equal across $k \in [K]$.

For each combination of K and q, 40 true DAGs, intervention targets and induced parent sets are generated as follows. We first draw a sparse DAG $\mathcal{D}$ with a probability of edge inclusion $3/(2q-2)$, so that the expected number of edges in the DAG grows linearly with the number of variables [Peters and Bühlmann, 2014]. Each target $T^{(k)}$, $k \in \{2, \dots, K\}$, is then generated by randomly including each node $j \in [q]$ in $T^{(k)}$ with probability $\theta_k = 0.2$. For each node $j \in T^{(k)}$, consider now matrix $\boldsymbol{P}^{(k)}$ which represents the (possibly different) parent sets induced by the intervention; the latter is constructed by randomly generating a new DAG with same topological ordering as $\mathcal{D}$, and replacing the original parent set of j with that of the new DAG. Finally, conditionally on DAG $\mathcal{D}$ and the so-obtained modified DAGs $\mathcal{D}_2, \dots, \mathcal{D}_K$, we draw the set of distinct parameters $\boldsymbol{B}_{lj}^{(k)}$ uniformly in $[-1, -0.1] \cup [0.1, 1]$, while we fix $\boldsymbol{D}_{jj}^{(k)} = 1$ for each $j \in [q]$ and $k \in [K]$. Finally, by recovering $\boldsymbol{\Sigma}_k$ from $\big(\boldsymbol{B}^{(k)}, \boldsymbol{D}^{(k)}\big)$, n_k observations are generated from $\mathcal{N}_q\big(\boldsymbol{0}, \boldsymbol{\Sigma}_k\big)$, for $k \in [K]$. Output is finally a collection of simulated datasets $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(K)}$.

We implement our method by running Algorithm 3 for number of MCMC iterations $S = 3000q$, discarding the initial $1000q$ draws that are used as a burn-in period. We set $a_\phi = b_\phi = 1$, $a_\eta = b_\eta = 1$ and $a_{\mathcal{D}} = a_{\mathcal{D}} = 1$ in the hierarchical model priors of Section 3.3.3. These specific choices result in uniform priors for the inclusion of a node in an intervention target (3.12), as a new parent (3.11) as well as for the probability of edge inclusion in $\mathcal{D}$ (3.13). Finally, we set $a = q$ and $\boldsymbol{U} = \boldsymbol{I}_q$ in the Wishart prior on $\boldsymbol{\Omega}$, leading to a weakly informative prior whose weight corresponds to a sample of size one.

We evaluate the performance of our method in the tasks of DAG learning and target identification. To this end, we consider as point estimates of DAGs and targets the Median Probability DAG model and Median Probability Targets as introduced in Section 3.4.2. Since there are no existing methods for causal discovery that align precisely with our framework of general interventions, providing a fully equitable comparison is not straightforward. To address this issue, we benchmark our approach against alternative methodologies designed for slightly different contexts. Specifically, we consider three methods: GIES [Hauser and Bühlmann, 2012], its recent extension GnIES [Gamella et al., 2022], and UT-IGSP [Squires et al., 2020].

GIES, which requires exact knowledge of the intervention targets, serves as a reference for the DAG structure learning task. In contrast, both GnIES and UT-IGSP learn the intervention targets from the data, but assume slightly different definitions of interventions. Specifically, GniES considers *noise-interventions*, which only modify the error-term distribution of the interventioned nodes in (3.1). Differently, UT-IGSP works under the framework of *soft interventions*.

Although the interventions considered by the methods above produce different post-intervention distributions, the implied invariances coincide, thus making our comparison sensible. In addition, all benchmarks provide an I-Essential Graph (I-EG) estimate which represents an I-Markov equivalence class of DAGs. We therefore adapt the MPM DAG estimate provided by our method by constructing the representative I-EG. Figure 3.6 summarizes the Structural Hamming Distance (SHD) between each I-EG estimate and true I-EG, for all methods under comparison; SHD is defined as the number of insertions, deletions or reversals needed to transform the estimated graph into the true DAG; accordingly lower values of SHD imply better performances.

Figure 3.7 instead reports the number of errors (both false positives and false negatives) relative to target identification for our method, GnIES and UT-IGSP. Our method exhibits a superior performance in comparison with the benchmarks, as also expected because of deviations of the simulated data from the assumptions underlying their methods. Therefore, the two benchmarks reveal difficulties in recovering a causal DAG structure from interventional data whose generating mechanism is consistent with a broader framework of interventions.

As described in Section 3.4.2, the output provided by our method can be also adapted to learn differences between DAGs corresponding to different experimental settings. For this specific goal, Wang et al. [2018] developed the Difference Causal Inference (DCI) algorithm. To assess the performance of our method in this context and compare it with DCI, we consider the same simulation scenarios for $K = 2$ defined before. With regard to DCI, we consider two implementations. In the first one, following [Belyaeva et al., 2021], we set $\alpha_{ug} = 0.001$, $\alpha_{sk} = 0.5$ and $\alpha_{dd} = 0.001$ as confidence levels for the tests used in the first, second and third step of the algorithm, respectively. In the second one, we use DCI with stability selection and we consider the grid of possible parameterizations constructed by considering $\alpha_{ug} \in \{0.001, 0.01\}$, $\alpha_{sk} \in \{0.1, 0.5\}$ and $\alpha_{dd} \in \{0.001, 0.01\}$. Figure 3.8 summarizes the sum of falsely identified and non-identified edges in the estimated difference-graph of $(\mathcal{D}_1, \mathcal{D}_2)$. Both methods improve their ability in recovering structural differences between the two DAGs as the sample size increases. Moreover, the performance of our method is slightly better than DCI, expecially under the $q = 20$ scenario.

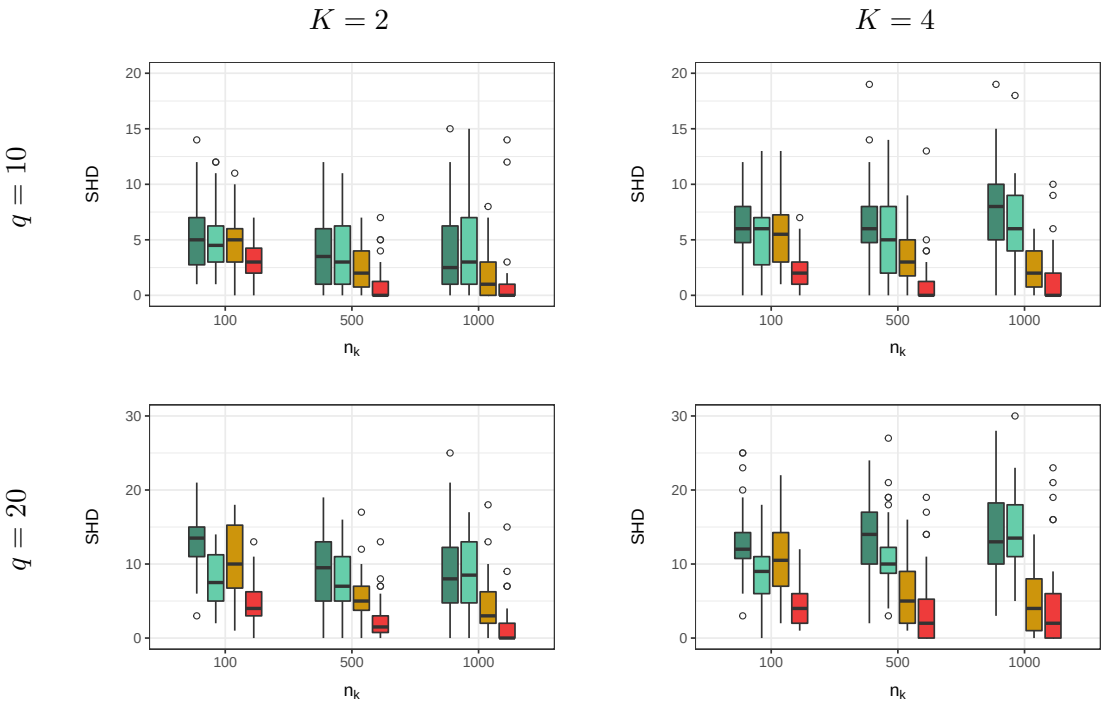


Figure 3.6: *Simulations. Distribution (across* 40 *simulations) of the Structural Hamming Distance (SHD) between true DAG and graph estimate, under scenarios* $q \in \{10, 20\}$ *(number of variables),* $K \in \{2, 4\}$ *(number of experimental contexts), and for increasing samples sizes* $n_k \in \{100, 500, 1000\}$*. Methods under comparison are: GIES and GnIES (dark and light blue), UT-IGSP (yellow) and our Bayesian approach (red).*

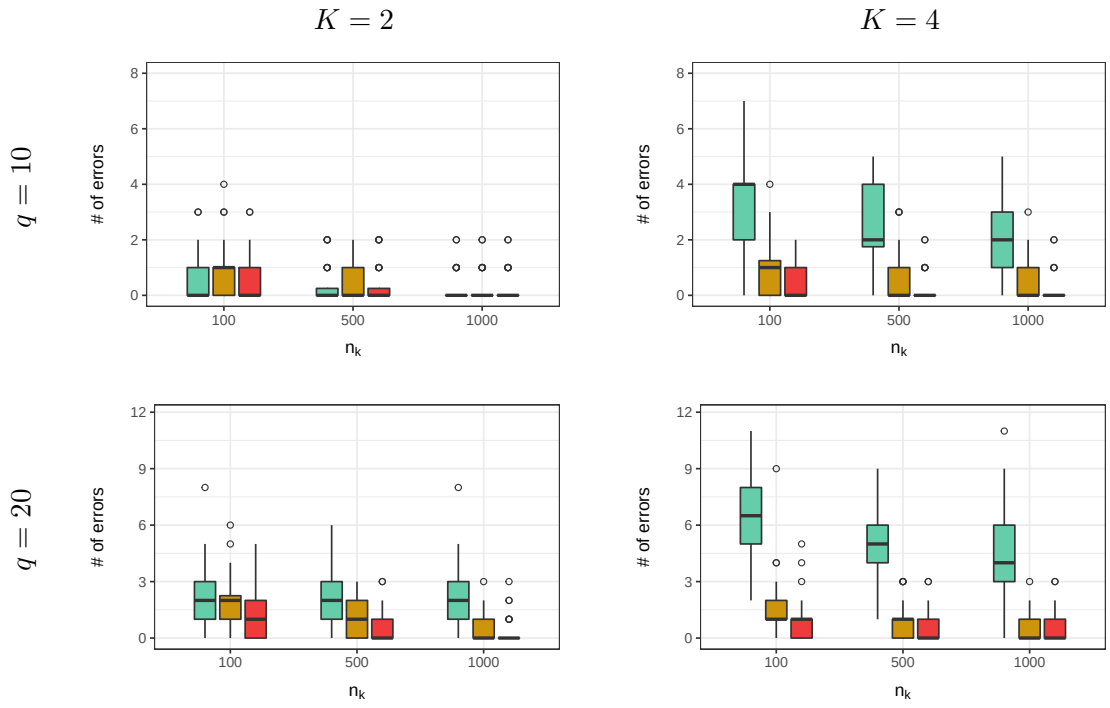


Figure 3.7: *Simulations. Distribution (across* 40 *simulations) of the number of false positives and false negatives (# of errors) between true and estimated targets, under scenarios* $q \in \{10, 20\}$ *(number of variables),* $K \in \{2, 4\}$ *(number of experimental contexts), and for increasing samples sizes* $n_k \in \{100, 500, 1000\}$*. Methods under comparison are: GnIES (light blue), UT-IGSP (yellow) and our Bayesian approach (red).*

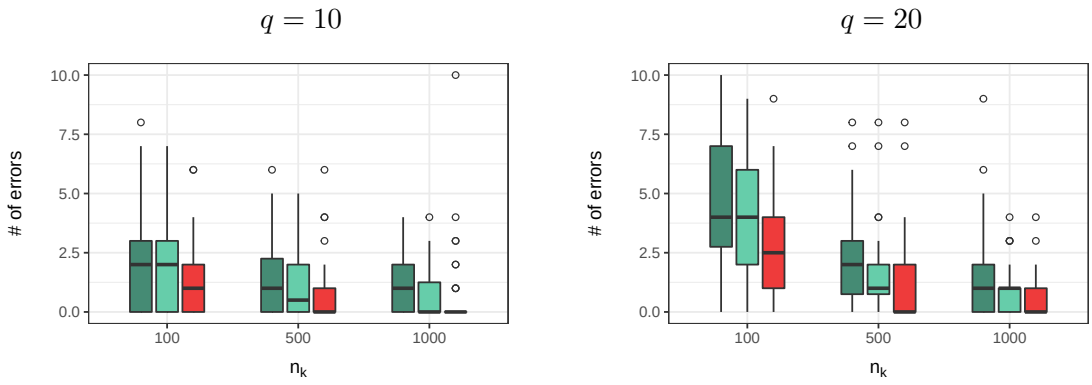


Figure 3.8: *Simulations. Distribution (across* 40 *simulations) of the sum of falsely identified and non-identified varying edges between context* $k = 1$ *and* $k = 2$*, under scenarios* $q \in \{10, 20\}$ *(number of variables) and for increasing samples sizes* $n_k \in \{100, 500, 1000\}$*. Method under comparison are: DCI and DCI with stability selection (dark and light blue) and our Bayesian approach (red).*

3.5.3 Real data analysis

We apply our methodology to a dataset of protein expression measurements from patients affected by Acute Myeloid Leukemia (AML). Subjects are classified into groups corresponding to distinct AML subtypes which were identified according to the French-American-British (FAB) system based on morphological features, cytogenetics, and assessment of recurrent molecular abnormalities. The complete dataset is provided as a supplement to Kornblau et al. [2009] and was previously analyzed from a multiple graphical modelling perspective by Peterson et al. [2015] and Castelletti et al. [2020]. Specifically, the authors developed Bayesian methodologies to infer a distinct graphical structure for each group (subtype), and simultaneously allowing for similar features across groups through a hierarchical prior on graphs favoring network relatedness. Given the distinct prognosis associated with each AML subtype, it is reasonable to expect variations in protein interactions among groups, as revealed by the analysis of Castelletti et al. [2020]. The investigation of such variations is of great interest from a therapeutic perspective, since it can provide valuable insights on the efficacy of a treatment capable of protein regulation depending on the specific patient's subtype; see also Castelletti and Consonni [2023].

Similarly to Peterson et al. [2015], we consider the level of $q = 18$ proteins and phosphoproteins involved in apoptosis and cell cycle regulation according to the KEGG database, relative to $n = 178$ diagnosed AML patients corresponding to the following $K = 4$ subtypes:

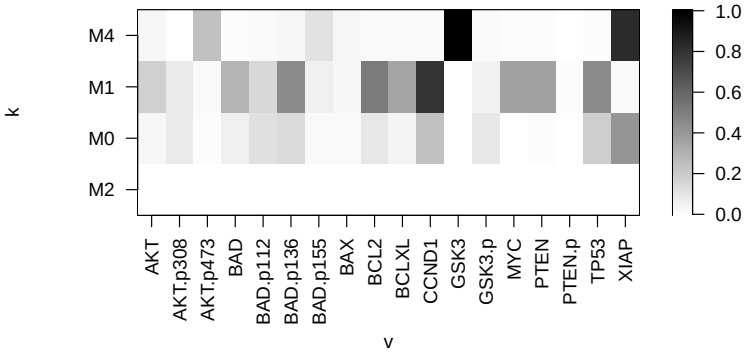


Figure 3.9: *AML data. Estimated marginal posterior probabilities of target inclusion, computed for each node $v \in [q]$ across AML subtypes, each corresponding to an experimental context k. Subtype M2 corresponds to the reference (observational) context.*

M0 (17 subjects), M1 (34 subjects), M2 (68 subjects) and M4 (59 subjects). We designate the largest group, M2, as the observational reference group, and attribute differences among subtypes to unspecified general interventions that may have altered the reference network structure. We implement our methodology by running Algorithm 3 for a number of MCMC iterations $S = 250000$, and discarding the initial 50000 draws which are used as a burn-in period. We consider for all priors the same weakly informative hyperparameter choices employed in the simulation study of Section 3.5.2.

As a summary of the MCMC output we first compute the marginal probability of target inclusion according to Equation (3.15) for each node $v \in [q]$ and AML subtype (experimental context k). The resulting collection of probabilities is summarized in the heat map of Figure 3.9. Results show that a few proteins are with high probability targeted as the result of unknown interventions that affect the network of protein interactions under any of the subtypes. Specifically, only four proteins, namely BCL2 and CCND1 under Subtype M1 and GSK3 and XIAP under Subtype M4, are identified as intervention targets with a posterior probability exceeding 0.5. Differences in the implied set of parent-child relations involving such nodes are therefore expected in the implied post-intervention graphs. By converse, there are no proteins whose probabilities of intervention are higher than the 0.5 threshold under Subtype M0.

According to Equation (3.14), we then compute the Posterior Probability of Inclusion (PPI) for each possible directed edge (u, v) and each group-specific post-intervention DAG, corresponding to one of the four subtypes. Results for each subtype M0, M1, M2, M4

are reported in the (q, q) heat maps of Figure 3.10, where any (u, v)-element in the plots corresponds to the marginal probability of inclusion of $u \to v$ in one of the four DAGs.

Finally, as single graphs summarizing the entire MCMC output, we provide a collection of context-specific MPM DAG estimates, $\widehat{\mathcal{D}}_k, k = 1, \ldots, 4$. These are reported in Figure 3.11, where for ease of interpretation the graph indexing the observational context (Subtype M2) corresponds to the I-EG representing the equivalence class of the estimated DAG. As expected from the previous results, the four graphs exhibit several similarities. An instance is the path involving the PTEN, PTEN.p and BAD.p136, BAD.p155 proteins. Such associations are consistent with findings in Peterson et al. [2015] who also identified (undirected) links between these proteins under all groups. In addition, our method detects a direct effect of BAD.p136 on PTEN.p, as well as of PTEN on BAD.p155 for all leukemia patients. A notable difference across groups is instead represented by the absence of the directed link AKT $\to$ GSK3 in group M4 as the effect of a (hard) intervention targeting GSK3 and which removes its parents. Notably, the correlation of GSK3 with a number of proteins involved in AML, and primarly AKT, was established in the medical literature; see for instance Ruvolo et al. [2015] and Ricciardi et al. [2017]. In particular, the AKT/GSK3 path was shown to represent a critical axis in AML, which may be a therapeutic target in AML patients with intermediate cytogenetics (M2 subtype). Out results show that an *intervention* on AKT aimed at regulating the GSK3 protein may be beneficial for patients characterized by AML subtypes M0, M1, M2, while uneffective whenever applied to M4 patients since there are no paths from AKT downstreaming to GSK3.

3.6 Discussion

In this chapter we introduce a statistical framework for causal discovery from multivariate interventional data. The notion of general intervention that we implement allows for structural modifications in the parent-child relations involving the intervened nodes, where the latter can be both known in advance or completely uncertain. Under both contexts, we first establish DAG identifiability and provide graphical criteria to characterize interventional Markov equivalence of DAGs. We then develop a Bayesian methodology for structure learning, by introducing an effective procedure which dramatically simplifies parameter prior elicitation. In addition, it provides a closed-form expression for the DAG marginal likelihood which guarantees score equivalence among I-Markov equivalent DAGs. We complete our Bayesian model formulation by assigning priors to model parameters corresponding to DAGs, intervention targets, and modified parent sets. Finally, to approximate the corresponding posterior distribution, we develop a Markov Chain Monte Carlo (MCMC) sampler

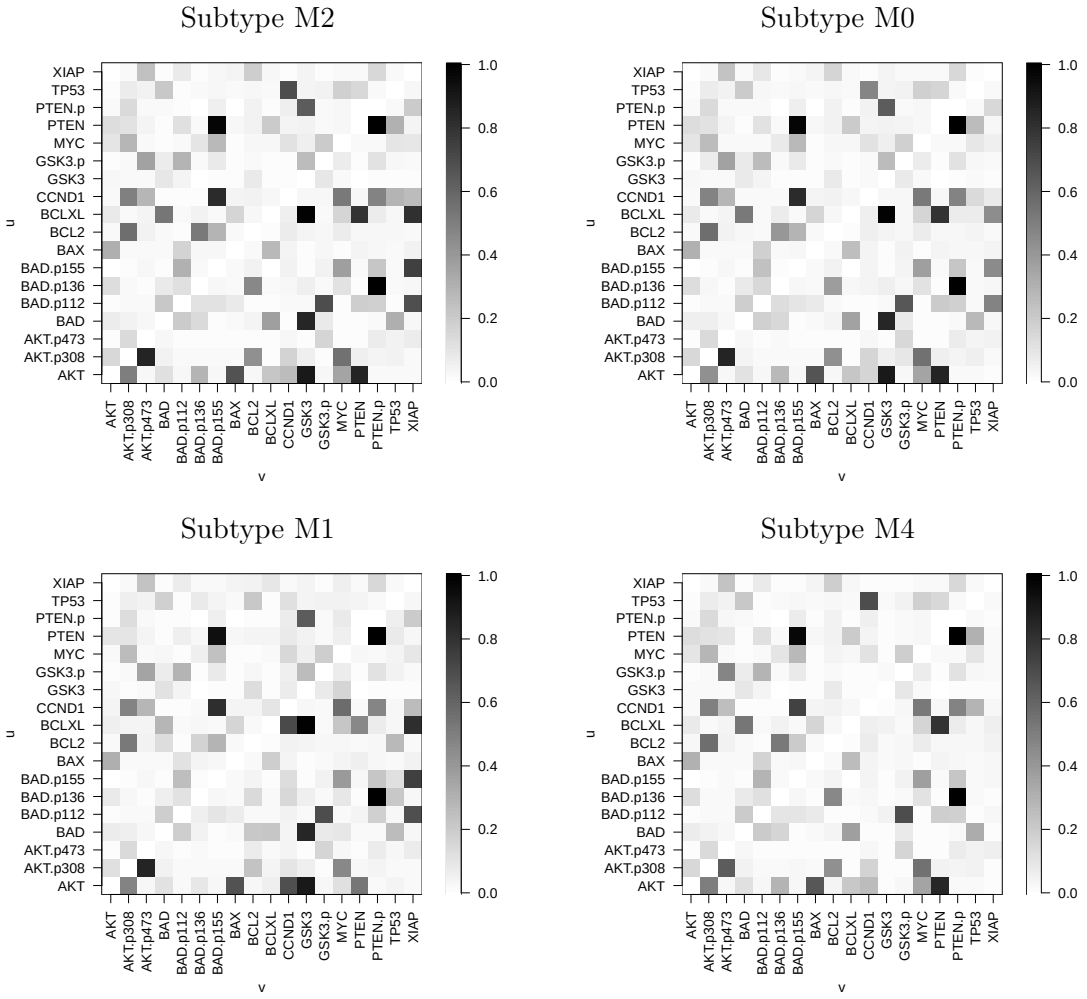


Figure 3.10: *AML data. Estimated marginal posterior probabilities of edge inclusion, computed for each possible directed edge* (u, v)*,* $u, v \in [q]$ *and group-specific post-intervention DAG, each corresponding to one of the four AML subtypes.*

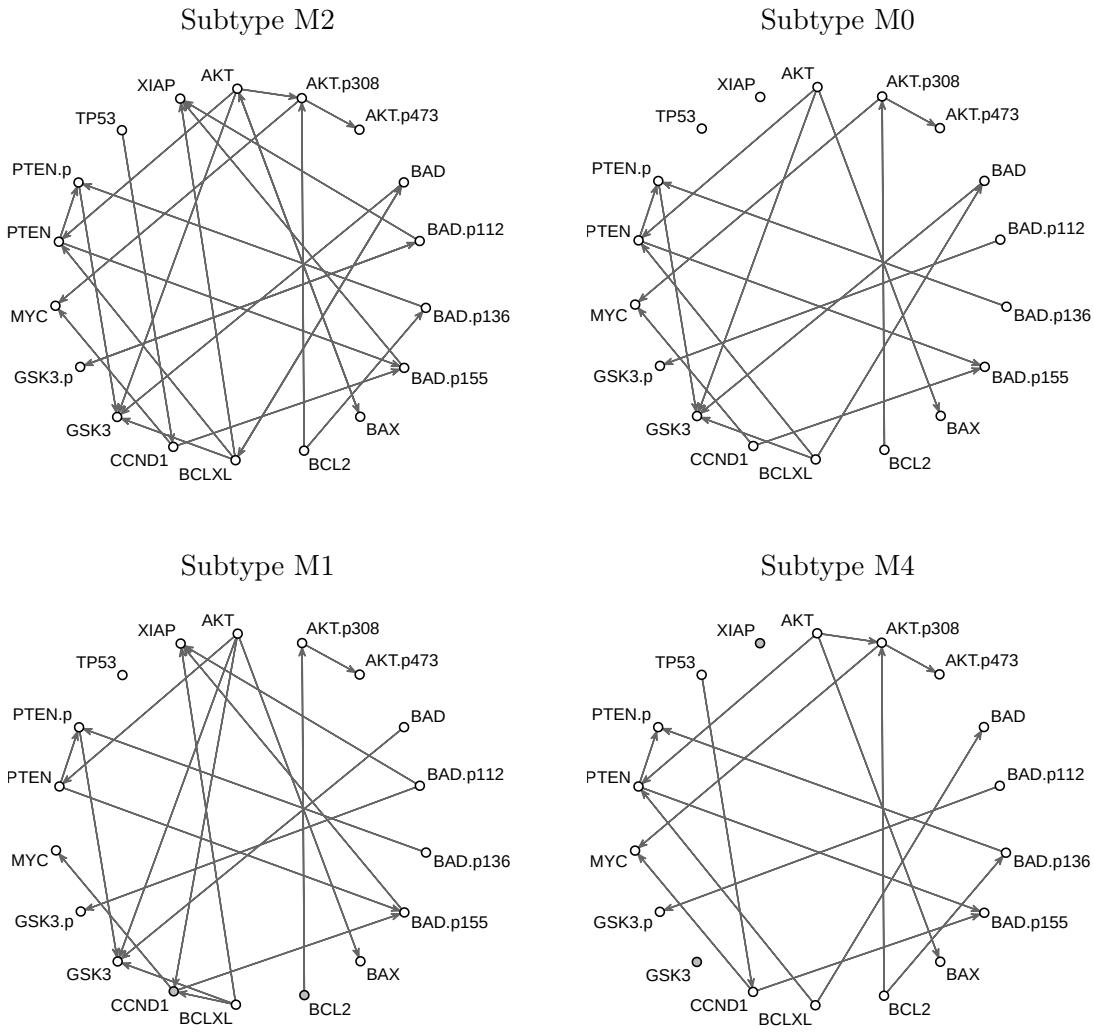


Figure 3.11: *AML data. Median Probability graph Model (MPM) estimates obtained under each AML subtype. Graph corresponding to Subtype M2 is the representative I-EG.*

based on a random scan Metropolis Hastings scheme.

3.6.1 Future developments

Our Bayesian framework for causal discovery relies on a set of general assumptions on the likelihood and prior that are satisfied under various parametric families, and notably zero-mean Gaussian models, when equipped with a Wishart prior on the precision matrix. Within such context, the full development of a methodology for structure learning and target identification is possible, and asymptotic properties relative to posterior ratio consistency could be established along the lines of Castelletti and Peluso [2023b] and Castelletti and Peluso [2023a] for the case of known and unknown hard interventions respectively. Similarly, our framework can be implemented for the analysis of categorical DAGs, under a multinomial-Dirichlet model. The resulting method would extend the original methodology of Heckerman et al. [1995], developed for i.i.d. observational samples and leading to their BDeu score, to an experimental setting of general (unknown) interventions.

Our approach for causal discovery is based on the assumption that the data are generated according to a Markovian Structural Causal Model (SCM) with no cycles, and which can be thus represented by a Directed *Acyclic* Graph. Besides the absence of cycles, our SCM representation assumes that there are no latent (unmeasured) confounders. Recently, Bongers et al. [2021] proposed a general theory for causal discovery which allows for the presence of both latent confounders and cycles, establishing identifiability conditions of SCMs as well as several statistical properties of their methodology. An extension of our method for causal discovery under general interventions towards this direction can be also of interest.

Appendix

Appendix A: Proofs of Section 3.2

This section contains all the proofs of the main results presented in Sections 3.2.2 and 3.2.3 of the manuscript. The numbering of such propositions and theorems in this section is the same as in the main text. Additional auxiliary lemmas and propositions that are newly introduced within this appendix follow instead the sequential numbering in line with the main text.

A.1 Proofs of Section 3.2.2

The definition of the I-Markov property and the subsequent graphical characterization of I-Markov equivalence in terms of skeleta and v-structures that we propose in Section 3.2.2 is similar to the one provided by Yang et al. [2018] for the case of soft interventions. As a consequence, our approach to proving Proposition 11 and Theorem 12 closely aligns with their strategy.

We first characterize I-Markov equivalence in our setting in terms of the ensued factorization:

Lemma 25. *$\{p_k(\cdot)\}_{k=1}^K \in \mathcal{M}_{\mathcal{I}}(\mathcal{D})$ if and only if there exists $p(\cdot) \in \mathcal{M}(\mathcal{D})$ such that, for each $k \in [K]$, $p_k(\cdot)$ factorizes as $\prod_{j \notin T^{(k)}} p(x_j \,|\, x_{pa_{\mathcal{D}}(j)}) \prod_{j \in T^{(k)}} p_k(x_j \,|\, x_{pa_{\mathcal{D}_k}(j)})$.*

Proof. If - Suppose there exists $p(\cdot) \in \mathcal{M}(\mathcal{D})$ such that the factorization above holds. The first condition from the definition of the I-Markov equivalence class, namely that $p_k(x) \in \mathcal{M}(\mathcal{D}_k)$ is trivially satisfied for all $k \in [K]$. As for the second condition, note that for all $j \notin T^{(k)}$ we have $p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) = p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)}) = p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)})$. As a consequence, $p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) = p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)}) = p_{k'}(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_{k'}}(j)})$, $\forall\, j \notin T^{(k)} \cup T^{(k')}$ and $T^{(k)}, T^{(k')} \in \mathcal{T}$. Hence $\{p_k(x)\}_{k=1}^K \in \mathcal{M}_{\mathcal{I}}(\mathcal{D})$.

Only if - Suppose that $\{p_k(x)\}_{k=1}^K \in \mathcal{M}_{\mathcal{I}}(\mathcal{D})$. To prove that there exists $p(x) \in \mathcal{M}(\mathcal{D})$ such that the factorization in the lemma holds, take any $p(x) \in \mathcal{M}(\mathcal{D})$. By definition, it

holds that $p_k(x) = \prod_{j=1}^{q} p_k(x_j \mid x_{\mathrm{pa}_{\mathcal{D}_k}(j)})$. From the second condition, we have that for any $k \in [K]$ and $j \notin T^{(k)}$, $p_k(x_j \mid x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) = p(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)})$, where $p(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)})$ is an arbitrary strictly positive density, so that the factorization in the lemma holds for all $T \in \mathcal{T}$. □

Proposition 11. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a collection of targets and induced parent sets. Then $\{p_k(\cdot)\}_{k=1}^K \in \mathcal{M}_{\mathcal{I}}(\mathcal{D})$ if and only if $\{p_k(\cdot)\}_{k=1}^K$ satisfies the I-Markov property with respect to $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$.*

Proof. If - Choose any $k \in [K]$ and use the chain rule to factorize $p_k(\cdot)$ according to the topological ordering of $\mathcal{D}_k$, so that

$$p_k(x) = \prod_{j=1}^{q} p_k(x_j \mid x_{a_j(\pi_{\mathcal{D}_k})}),$$

where $a_j(\pi_{\mathcal{D}_k})$ represents all the nodes that precede j in the topological ordering implied by $\mathcal{D}_k$. As each node is d-separated from its non-descendants given its parents, from the first condition of the general I-Markov property we obtain

$$p_k(x) = \prod_{j=1}^{q} p_k(x_j \mid x_{\mathrm{pa}_{\mathcal{D}_k}(j)}).$$

Moreover, each node $j \notin T^{(k)}$ is d-separated from ζ_k given its parents in $\mathcal{D}_k^{\mathcal{I}}$. Hence, from the second condition of the general I-Markov property we have $p_k(x_j \mid x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) = p(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)})$, so that

$$p_k(x) = \prod_{j \notin T^{(k)}} p(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}) \prod_{j \in T^{(k)}} p_k(x_j \mid x_{\mathrm{pa}_{\mathcal{D}_k}(j)}).$$

Hence the result follows from the Lemma above.

Only if - We want to prove that if $p_k(\cdot)$ factorizes according to

$$p_k(x) = \prod_{j \notin T^{(k)}} p(x_j \mid x_{\mathrm{pa}_{\mathcal{D}}(j)}) \prod_{j \in T^{(k)}} \tilde{p}(x_j \mid x_{\mathrm{pa}_{\mathcal{D}_k}(j)})$$

for all $k \in [K]$, then the general I-Markov property holds, namely the collection of $\mathcal{I}$-DAGs $\{\mathcal{D}_k^{\mathcal{I}}\}_{k=1}^K$ can be used to recover all the conditional independencies and invariances through d-separation criteria.

As for the conditional independencies, note that by Lemma 25 we have that $p_k(\cdot)$ factorizes according to $\mathcal{D}_k$ for all $k \in [K]$. Hence, for each $k \in [K]$ the Markov property defined on d-separation criteria must hold with respect to $\mathcal{D}_k$. Hence the first condition of the I-Markov property must hold.

For the second condition, instead, we want to show that the invariant components of the distribution are exactly those whose nodes j's are d-separated from ζ_I given a set C in $\mathcal{D}_k^{\mathcal{I}}$, for all $k \in [K]$. Consider any two disjoint sets $A, C \subset [q]$ and $k \in [K]$ and suppose that C d-separates A from ζ_k in $\mathcal{D}_k^{\mathcal{I}}$. Now, let V_{An} be the ancestral set of A and C in $\mathcal{D}_k$. Denote with $B' \subset V_{An}$ those nodes that are also d-connected to ζ_k in $\mathcal{D}_k^{\mathcal{I}}$ given C and with $A' = V_{An} \backslash \{B' \cup C\}$ the sets of ancestors of A and C that are not d-connected to ζ_k and that are not in the conditioning set C. Note that $V_{An} = A' \cup B' \cup C$. From the factorization, we have that

$$
\begin{aligned}
p_k(x) &= p_k(x_{A'}, x_{B'}, x_C, x_{V \backslash V_{An}}) \\
&= \prod_{j \in A'} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \prod_{j \in B'} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \\
&\quad \prod_{j \in C} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \prod_{j \in V \backslash V_{An}} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \\
&= \prod_{j \in A'} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \prod_{j \in B'} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \prod_{j \in C, \mathrm{pa}_{\mathcal{D}_k}(j) \cap A' = \emptyset} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \\
&\quad \prod_{j \in C, \mathrm{pa}_{\mathcal{D}_k}(j) \cap A' \neq \emptyset} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \prod_{j \in V \backslash V_{An}} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \\
&= \prod_{j \in A'} p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)}) \prod_{j \in B'} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \prod_{j \in C, \mathrm{pa}_{\mathcal{D}_k}(j) \cap A' = \emptyset} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}) \\
&\quad \prod_{j \in C, \mathrm{pa}_{\mathcal{D}_k}(j) \cap A' \neq \emptyset} p(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}}(j)}) \prod_{j \in V \backslash V_{An}} p_k(x_j \,|\, x_{\mathrm{pa}_{\mathcal{D}_k}(j)}),
\end{aligned}
$$

where the last equality follows from the fact that

- if $j \in A'$, then j is d-separated from ζ_k in $\mathcal{D}_k^{\mathcal{I}}$ given C and thus j can not be a child of ζ_k;
- if $j \in C$ and there exists at least one $h \in \mathrm{pa}_{\mathcal{D}_k}(j)$ such that $h \in A'$, then j can not be a child of ζ_k: if it were, then conditioning on j its parents would be d-connected to ζ_k given C;

and recalling that $j \in \mathrm{ch}_{\zeta_k}(\mathcal{D}_k^{\mathcal{I}})$ if and only if $j \in T^{(k)}$. Similarly, the (union of) parents of nodes in A' and $\{j \in C \,|\, \mathrm{pa}_{\mathcal{D}_k}(j) \cap A' \neq \emptyset\}$ are subsets of $A' \cup C$, while the parents of B' and $\{j \in C \,|\, \mathrm{pa}_{\mathcal{D}_k}(j) \cap A' = \emptyset\}$ are subsets of $B' \cup C$. We can thus write

$$
p_k(x) = g(x_{A'}, x_C) g_k(x_{B'}, x_C) g_k(x_{V \backslash V_{An}}),
$$

to highlight the observational and interventional blocks in the factorization above and their arguments. We can thus marginalize out $A'\backslash A$, B' and $V\backslash V_{An}$, thus obtaining

$$\begin{aligned} p_k(x_A, x_C) &= \int_{X_{(A'\backslash A)\cup B'\cup(V\backslash V_{An})}} g(x_{A'}, x_C) g_k(x_{B'}, x_C) g_k(x_{V\backslash V_{An}}) \\ &= \int_{X_{(A'\backslash A)\cup B'}} g(x_{A'}, x_C) g_k(x_{B'}, x_C) \\ &= \int_{X_{(A'\backslash A)}} g(x_{A'}, x_C) \int_{X_{B'}} g_k(x_{B'}, x_C) \\ &= \tilde{g}(x_A, x_C)\tilde{g}_k(x_C). \end{aligned}$$

Using the latter expression we can write

$$\begin{aligned} p_k(x_A \,|\, x_C) = \frac{p_k(x_A, x_C)}{p_k(x_C)} &= \frac{\tilde{g}(x_A, x_C)\tilde{g}_k(x_C)}{\int_{X_A} \tilde{g}(x_A, x_C)\tilde{g}_k(x_C)} \\ &= \frac{\tilde{g}(x_A, x_C)\tilde{g}_k(x_C)}{\tilde{g}_k(x_C)\int_{X_A} \tilde{g}(x_A, x_C)} \\ &= \frac{\tilde{g}(x_A, x_C)}{\int_{X_A} \tilde{g}(x_A, x_C)}, \end{aligned}$$

which does not depend on $T^{(k)}$ and is thus invariant as required by the Markov property. □

Theorem 12. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a collection of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1$ and $\mathcal{D}_2$. $\mathcal{D}_1$ and $\mathcal{D}_2$ belong to the same I-Markov equivalence class if and only if $\mathcal{D}_{1,k}^{\mathcal{I}}$ and $\mathcal{D}_{2,k}^{\mathcal{I}}$ have the same skeleta and v-structures for all $k \in [K]$.*

Proof. If: Because $\mathcal{D}_{1,k}^{\mathcal{I}}$ and $\mathcal{D}_{2,k}^{\mathcal{I}}$ have the same sleketon and set of v-structures for each $k \in [K]$, the two collections of $\mathcal{I}$-DAGs $\{\mathcal{D}_{1,k}^{\mathcal{I}}\}_{k=1}^K, \{\mathcal{D}_{2,k}^{\mathcal{I}}\}_{k=1}^K$ satisfy the same d-separation statements, thus implying the same sets of conditional independencies and invariances through the I-Markov property, so that $\mathcal{M}_{\mathcal{I}}(\mathcal{D}_1) = \mathcal{M}_{\mathcal{I}}(\mathcal{D}_2)$.

Only if: Suppose there exists a $k^* \in [K]$ such that $\mathcal{D}_{1,k^*}^{\mathcal{I}}$ and $\mathcal{D}_{2,k^*}^{\mathcal{I}}$ do not have the same skeleton and set of v-structures. Denote with $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ the post intervention DAGs corresponding to the k^*th experimental setting. Note that $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ have the same skeleta and sets of v-structures, otherwise $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ wouldn't be Markov equivalent and consequently $(\mathcal{D}_1, \mathcal{D}_2)$ wouldn't be I-Markov equivalent given $\mathcal{I}$. Moreover, $\mathcal{D}_{1,k^*}^{\mathcal{I}}$ and $\mathcal{D}_{2,k^*}^{\mathcal{I}}$

have the same $\mathcal{I}$-edges, as these are determined by $T^{(k^*)}$. They thus differ for the sets of v-structures involving $\mathcal{I}$-edges. Suppose that $\zeta_{k^*} \to v \leftarrow w$ is a v-structure in $\mathcal{D}^{\mathcal{I}}_{1,k^*}$ which is not present in $\mathcal{D}^{\mathcal{I}}_{2,k^*}$, meaning that $w \notin T^{(k^*)}$ and $w \in P_v^{(k^*)}$. As the modified DAGs $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ have the same skeleton, then $\zeta_{k^*} \to v \to w \in \mathcal{D}^{\mathcal{I}}_{2,k^*}$. As the parent set of v is fixed by the intervention, we would have that both $v \leftarrow w \in \mathcal{D}^{\mathcal{I}}_{2,k}$ and $v \to w \in \mathcal{D}^{\mathcal{I}}_{2,k}$, which implies a cycle and thus a contradiction with the validity assumption. □

We now shift our focus on the transformational characterization of Theorem 13

Lemma 26. *Let $\mathcal{D}_1$ be a DAG containing the edge $u \to v$ and $\mathcal{I}$ a collection of targets and induced parent sets defining a general intervention. Let $\mathcal{D}_2$ be a graph identical to $\mathcal{D}_1$ except for the reversal of $u \to v$. $\mathcal{D}_1$ and $\mathcal{D}_2$ belong to the same I-Markov Equivalence class if and only if $u \to v$ is simultaneously covered;*

Proof. If: Suppose $u \to v$ is simultaneously covered. Then, $u \to v$ is covered in $\mathcal{D}_1$ and, for any $k \neq 1$, $u \to v$ is either (i) covered in $\mathcal{D}^{\mathcal{I}}_{1,k}$ or (ii) $\{u, v\} \subseteq T^{(k)}$. In case (i), we cannot have $u \in T^{(k)}$ and $v \notin T^{(k)}$ (or viceversa) by the definition of covered edge in the $\mathcal{I}$-DAG. The parent sets of the two nodes in the $\mathcal{I}$-DAGs are thus the same as in the observational DAG $\mathcal{D}$ and the proof follows from Chickering [1995, Lemma 1]. In case (ii), both u and v are targets of intervention and reversing $u \to v$ in $\mathcal{D}_1$ does not cause any change in the parent sets of the nodes in the $\mathcal{I}$-DAGs. $u \to v$ thus has to be covered only in $\mathcal{D}$ and the proof follows again from Chickering [1995, Lemma 1].

Only if: Suppose that $u \to v$ is not simultaneously covered. Then, at least one of the following statements is true: (i) $u \to v$ is not covered in $\mathcal{D}_1$; (ii) there exists $k^* \in [K]$ such that $u \to v$ is not covered in $\mathcal{D}^{\mathcal{I}}_{1,k^*}$ and $\{u, v\} \not\subseteq T^{(k^*)}$. In case (i) the proof follows from Chickering [1995, Lemma 1]. In case (ii), we have that, by the definition of a covered edge, $\text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(u) \cup u \neq \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(v)$. In particular, either there exists at least one z such that $z \in \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(u), z \notin \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(v)$, or there exists at least one node w such that $w \in \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(v), w \notin \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(u)$. Consider the first case. Then, either (a) $z = \zeta_{k^*}$ or (b) $z \neq \zeta_{k^*}$. In case (a), note that $v \notin T^{(k^*)}$, by definition of z, so that $u \to v \in \mathcal{D}^{\mathcal{I}}_{1,k^*}$. As the intervention is defining the parent set of node u, we have that $\text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(u) = \text{pa}_{\mathcal{D}^{\mathcal{I}}_{2,k^*}}(u)$. Moreover, the intervention is supposed to be valid, so that $v \notin \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(u)$. We thus have that $u \to v \in \mathcal{D}^{\mathcal{I}}_{1,k^*}$, while both $u \to v, v \to u \notin \mathcal{D}^{\mathcal{I}}_{2,k^*}$. As $\mathcal{D}^{\mathcal{I}}_{1,k^*}, \mathcal{D}^{\mathcal{I}}_{2,k^*}$ differ for their skeleta, they can not be I-Markov equivalent. In case (b), instead, by the definition of a not simultaneously-covered edge, we have that ζ_{k^*} does not belong to the common parents of $\{u, v\}$. Hence, $\{u, v\} \not\subseteq T^{(k)}$ and $u \to v$ is covered in $\mathcal{D}^{\mathcal{I}}_{1,k^*}$ if and only if it is covered in

$\mathcal{D}_1$ (and the same holds for $\mathcal{D}_2$). The proof thus follows from Chickering [1995, Lemma 1]. The proof for case $w \in \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(v), w \notin \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(u)$ follows by a similar reasoning. □

Let $\Delta(\mathcal{D}_1, \mathcal{D}_2)$ denote the set of edges in $\mathcal{D}_1$ that have opposite orientation in $\mathcal{D}_2$ and $\Psi_v = \{u \,|\, u \to v \in \Delta(\mathcal{D}_1, \mathcal{D}_2)\}$, the set of nodes that are parents of v in $\mathcal{D}_1$ and children of v in $\mathcal{D}_2$. Algorithm 6 was first presented in Chickering [1995] to find a covered edge belonging to $\Delta(\mathcal{D}_1, \mathcal{D}_2)$ for two Markov Equivalent DAGs and it can be also adopted in our setting.

Algorithm 6: Find-Edge (Chickering, 1995)

Input: DAGs $\mathcal{D}_1, \mathcal{D}_2$
Output: Edge from $\Delta(\mathcal{D}_1, \mathcal{D}_2)$
1 Perform a topological sort on the nodes in $\mathcal{D}_1$;
2 Let v be the minimal node with respect to the sort for which $\Psi_v \neq \emptyset$;
3 Let u be the maximal node with respect to the sort for which $u \in \Psi_v$;
4 **return** $u \to v$

Lemma 27. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two I-Markov equivalent DAGs for $\mathcal{I}$, a collection of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1, \mathcal{D}_2$. The edge $u \to v$ output from Algorithm 6 with input two $\mathcal{D}_1, \mathcal{D}_2$ is simultaneously covered.*

Proof. We know from Lemma 2 in Chickering [1995] that $u \to v$ is covered in $\mathcal{D}_1$. Suppose now that $u \to v$ is not simultaneously covered. Hence, there must exist at least one $k^* \neq 1$ such that $u \to v$ is not covered in $\mathcal{D}^{\mathcal{I}}_{1,k^*}$ and $\{u, v\} \not\subseteq T^{(k^*)}$. In particular, either (i) $u \in T^{(k^*)}, v \notin T^{(k^*)}$ or (ii) $v \in T^{(k^*)}, u \notin T^{(k^*)}$. Suppose (i). Note that $v \notin \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(u)$ as the intervention is supposed to be valid. Hence, we have that $\zeta_{k^*} \to u \to v$ in $\mathcal{D}^{\mathcal{I}}_{1,k^*}$ and $\zeta_{k^*} \to u \not\leftarrow v$ in $\mathcal{D}^{\mathcal{I}}_{2,k^*}$. Because $\mathcal{D}_1, \mathcal{D}_2$ now differ for their skeleton in one of the $\mathcal{I}$-DAGs, they can not be I-Markov equivalent. Suppose (ii). In this case, we have that either (a) $u \notin \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(v)$ or (b) $u \in \text{pa}_{\mathcal{D}^{\mathcal{I}}_{1,k^*}}(v)$. In case (a), we have that $u \not\to v \leftarrow \zeta_{k^*}$ in $\mathcal{D}^{\mathcal{I}}_{1,k^*}$ and $u \leftarrow v \leftarrow \zeta_k$ in $\mathcal{D}^{\mathcal{I}}_{2,k^*}$, as the parents of v remain invariant between $\mathcal{D}_2$ and $\mathcal{D}^{\mathcal{I}}_{2,k^*}$. The difference in skeleton implies that $\mathcal{D}_1, \mathcal{D}_2$ are not I-Markov equivalent, a contradiction. In case (b), for the same reason we would have $u \to v \leftarrow \zeta_{k^*}$ in $\mathcal{D}^{\mathcal{I}}_{1,k^*}$ and $u \leftrightarrow v \leftarrow \zeta_{k^*}$ in $\mathcal{D}^{\mathcal{I}}_{2,k^*}$ thus contradicting the fact that $\mathcal{I}$ is a valid collection of targets and induced parent sets. □

Theorem 13. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a collection of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1$ and $\mathcal{D}_2$. $\mathcal{D}_1$ and $\mathcal{D}_2$ belong to the same*

I-Markov equivalence class if and only if there exists a sequence of edge reversals modifying $\mathcal{D}_1$ *and such that:*

1. *Each edge reversed is simultaneously covered;*
2. *After each reversal,* $\{\mathcal{D}^{\mathcal{I}}_{1,k}\}_{k=1}^{K}$ *are DAGs and* $\mathcal{D}_1, \mathcal{D}_2$ *belong to the same I-Markov equivalence class;*
3. *After all reversals* $\mathcal{D}_1 = \mathcal{D}_2$.

Proof. If: The proof follows immediately from the definition of the sequence.
Only if: We show that all the conditions are satisfied if we apply the procedure Find-Edge to $\mathcal{D}_1, \mathcal{D}_2$ to identify the next edge to reverse in $\mathcal{D}_1$. We know that $u \to v$, the output of Find-Edge, is a simultaneously covered edge (Lemma 27). As it is simultaneously covered, the DAG obtained by reversing the edge still belongs to the same I-Markov equivalence class by Lemma 26. Moreover, $|\Delta(\mathcal{D}, \mathcal{D}')|$ decreases by one at each step. All the three conditions are thus satisfied. □

A.2 Proofs of Section 3.2.3

We here report the proofs of the results presented in Section 3.2.3, concerning the identifiability of i) unknown general interventions and ii) unknown DAGs and general interventions.

Theorem 16. *Let* $\mathcal{D}$ *be a DAG and* $\mathcal{I}_1, \mathcal{I}_2$ *two collections of targets and induced parent sets defining a general intervention. Then,* $\mathcal{I}_1, \mathcal{I}_2$ *belong to the same I-Markov equivalence class if and only if* $\mathcal{D}^{\mathcal{I}_1}_k, \mathcal{D}^{\mathcal{I}_2}_k$ *have the same skeleta and v-structures for all* $k \in [K]$.

Proof. If: As $\mathcal{D}^{\mathcal{I}_1}_k$ and $\mathcal{D}^{\mathcal{I}_2}_k$ have the same skeleton and same set of v-structures for all $k \in [K]$, they imply the same d-separation statements, thus implying the same sets of conditional independencies and invariances through the I-Markov property, so that $\mathcal{M}_{\mathcal{I}_1}(\mathcal{D}) = \mathcal{M}_{\mathcal{I}_2}(\mathcal{D})$.
Only if: Suppose there exists a $k^* \in [K]$ such that $\mathcal{D}^{\mathcal{I}_1}_{k^*}$ and $\mathcal{D}^{\mathcal{I}_2}_{k^*}$ do not have the same skeleton and set of v-structures. Denote with $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ the post intervention DAGs corresponding to the k^*th experimental setting. Note that $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ have the same skeleta and sets of v-structures, otherwise $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ wouldn't be Markov equivalent and consequently $(\mathcal{I}_1, \mathcal{P}_1)$, $(\mathcal{I}_2, \mathcal{P}_2)$ wouldn't be I-Markov equivalent. $\mathcal{D}^{\mathcal{I}_1}_{k^*}$ and $\mathcal{D}^{\mathcal{I}_2}_{k^*}$ thus differ for their sets of $\mathcal{I}$-edges and for v-structures involving the $\mathcal{I}$-edges. In case of a difference in skeleton, suppose without loss of generality that $\mathcal{D}^{\mathcal{I}_1}_{k^*}$ has an additional $\mathcal{I}$-edge $\zeta_{k^*} \to v$ which is not in $\mathcal{D}^{\mathcal{I}_2}_{k^*}$. Then, we have that $p_{k^*}(x_v \,|\, x_{\mathrm{pa}_{\mathcal{D}_{1,k^*}}(v)}) \neq p_1(x_v \,|\, x_{\mathrm{pa}_{\mathcal{D}_{1,k^*}}(v)})$, while

$p_{k^*}(x_v \mid x_{\mathrm{pa}_{\mathcal{D}_{2,k^*}}(v)}) = p_1(x_v \mid x_{\mathrm{pa}_{\mathcal{D}_{1,k^*}}(v)})$ and $\mathcal{I}_1$, $\mathcal{I}_2$ can't be I-Markov equivalent. In case of a difference in the sets of v-structures, suppose that $\zeta_{k^*} \to v \leftarrow w$ is a v-structure in $\mathcal{D}_{k^*}^{\mathcal{I}_1}$ which is not present in $\mathcal{D}_{k^*}^{\mathcal{I}_2}$. Accordingly $w \notin T_1^{(k^*)}$. As the modified DAGs $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ have the same skeleton, then $\zeta_{k^*} \to v \to w \in \mathcal{D}_{k^*}^{\mathcal{I}_2}$. However, because the parent set of w is changing between the two DAGs and $w \notin T_1^{(k^*)}$, it means that $w \in T_2^{(k^*)}$, so that $\zeta_{k^*} \to w \in \mathcal{D}_{k^*}^{\mathcal{I}_2}$, inducing a difference in skeleton. □

We now shift our focus on the transformational characterization of Theorem 17.

Lemma 28. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}_1, \mathcal{I}_2$ be two collections of targets and induced parent sets defining general interventions and such that, for some $k \in [K]$, $\mathcal{D}_k^{\mathcal{I}_1}, \mathcal{D}_k^{\mathcal{I}_2}$ differ only for the reversal of $u \to v \in \mathcal{D}_k^{\mathcal{I}_1}$ becoming $v \to u \in \mathcal{D}_k^{\mathcal{I}_2}$. $\mathcal{I}_1$ and $\mathcal{I}_2$ belong to the same I-Markov equivalence class if and only if $u \to v$ is covered in $\mathcal{D}_k^{\mathcal{I}_1}$.*

Proof. If: The proof is identical to Chickering [1995, Lemma 1].

Only if: Notice that, by construction, $\mathcal{D}_k^{\mathcal{I}_1}, \mathcal{D}_k^{\mathcal{I}_2}$ have the same skeleta and the same $\mathcal{I}$-edges in particular, so that $T_1^{(k)} = T_2^{(k)}$. Suppose now that $u \to v$ is not covered in $\mathcal{D}_k^{\mathcal{I}_1}$. Then $\mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(u) \cup u \neq \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(v)$. In particular, either (i) there exists some $z \in \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(u), z \notin \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(v)$ or (ii) there exists some $w \in \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(v), w \notin \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(u)$. In case (i), suppose that $z = \zeta_k$. In this case, $u \in T_1^{(k)}$ and $v \notin T_1^{(k)}$, so that $\mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(v) = \mathrm{pa}_{\mathcal{D}}(v)$. Because of the edge reversal, $\mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_1}}(v) \neq \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_2}}(v)$, implying that $\mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}_2}}(v) \neq \mathrm{pa}_{\mathcal{D}}(v)$ and $v \in T_2^{(k)}$, which is a contradiction as $T_1^{(k)} = T_2^{(k)}$ by construction. Hence, $z \neq \zeta_k$ and the proof follows from Chickering [1995, Lemma 1]. The proof for case (ii) follows by a similar reasoning. □

Lemma 29. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}_1, \mathcal{I}_2$ be two collections of targets and induced parent sets defining a general intervention and belonging to the same I-Markov equivalence class. The edge $u \to v$ output from Algorithm 6 with input $\mathcal{D}_k^{\mathcal{I}_1}, \mathcal{D}_k^{\mathcal{I}_2}$ is covered.*

Proof. The proof is identical to the one of Lemma 2 in Chickering [1995]. □

Theorem 17. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}_1$, $\mathcal{I}_2$ two collections of targets and induced parent sets. Then, $\mathcal{I}_1$, $\mathcal{I}_2$ belong to the same I-Markov equivalence class if and only if for each $\mathcal{I}$-DAG $\mathcal{D}_k^{\mathcal{I}_1}$ there exists a sequence of edge reversals modifying $\mathcal{D}_k^{\mathcal{I}_1}$ and such that:*

1. *Each edge reversed is covered;*
2. *After each reversal, $\mathcal{D}_k^{\mathcal{I}_1}$ is a DAG and $\mathcal{I}_1$, $\mathcal{I}_2$ belong to the same I-Markov equivalence class;*

3. After all reversals $\mathcal{D}_k^{\mathcal{I}_1} = \mathcal{D}_k^{\mathcal{I}_2}$.

Proof. If: It follows immediately from the definition of the sequence.
Only if: We show that all the conditions are satisfied if we apply the procedure Find-Edge with input $\mathcal{D}_k^{\mathcal{I}_1}$ and $\mathcal{D}_k^{\mathcal{I}_2}$, for all $k \neq 1$. We know that $u \to v$, output of Find-Edge is covered (Lemma 29) and that the $\mathcal{I}$-DAG obtained by reversing $u \to v$ corresponds to a collection of targets and induced parent sets which is I-Markov equivalent to the initial one (Lemma 28). At each step, $\Delta(\mathcal{D}_k^{\mathcal{I}_1}, \mathcal{D}_k^{\mathcal{I}_2})$ decreases by one. All the three conditions are thus satisfied. □

We now consider the set of results concerning the joint identifiability of a pair $(\mathcal{D}, \mathcal{I})$.

Theorem 19. *Let* $\mathcal{D}_1, \mathcal{D}_2$ *be two DAGs and* $\mathcal{I}_1, \mathcal{I}_2$ *two collections of targets and induced parent sets defining a valid general intervention for* $\mathcal{D}_1, \mathcal{D}_2$ *respectively.* $(\mathcal{D}_1, \mathcal{I}_1), (\mathcal{D}_2, \mathcal{I}_2)$ *belong to the same I-Markov equivalence class if and only if* $\mathcal{D}_{1,k}^{\mathcal{I}_1}, \mathcal{D}_{2,k}^{\mathcal{I}_2}$ *have the same skeleta and v-structures for all* $k \in [K]$.

Proof. If: As $\mathcal{D}_k^{\mathcal{I}_1}$ and $\mathcal{D}_k^{\mathcal{I}_2}$ have the same skeleta and set of v-structures for all $k \in [K]$, they imply the same d-separation statements, thus implying the same sets of conditional independencies and invariances through the I-Markov property, so that $\mathcal{M}_{\mathcal{I}_1}(\mathcal{D}) = \mathcal{M}_{\mathcal{I}_2}(\mathcal{D})$.
Only if: Suppose there exists a $k^* \in [K]$ such that $\mathcal{D}_{1,k^*}^{\mathcal{I}_1}$ and $\mathcal{D}_{2,k^*}^{\mathcal{I}_2}$ do not have the same skeleton and set of v-structures. Denote with $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ the post intervention DAGs corresponding to the k^*th experimental setting. Note that $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ have the same skeleta and sets of v-structures, otherwise $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ would not be Markov equivalent and consequently $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ wouldn't be I-Markov equivalent. $\mathcal{D}_{1,k^*}^{\mathcal{I}_1}$ and $\mathcal{D}_{2,k^*}^{\mathcal{I}_2}$ thus differ for their sets of $\mathcal{I}$-edges or for v-structures involving the $\mathcal{I}$-edges. In case of a difference in the $\mathcal{I}$-edges, suppose without loss of generality that $\mathcal{D}_{1,k^*}^{\mathcal{I}_1}$ has an additional $\mathcal{I}$-edge $\zeta_{k^*} \to v$ which is not in $\mathcal{D}_{2,k^*}^{\mathcal{I}_2}$. Then, we have that $p_{k^*}(x_v \,|\, x_{\mathrm{pa}_{\mathcal{D}_{1,k^*}}(v)}) \neq p_1(x_v \,|\, x_{\mathrm{pa}_{\mathcal{D}_{1,k^*}}(v)})$, while $p_{k^*}(x_v \,|\, x_{\mathrm{pa}_{\mathcal{D}_{2,k^*}}(v)}) = p_1(x_v \,|\, x_{\mathrm{pa}_{\mathcal{D}_{1,k^*}}(v)})$ and $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ can not be I-Markov equivalent. In case of a difference in the sets of v-structures, suppose that $\zeta_{k^*} \to v \leftarrow w$ is a v-structure in $\mathcal{D}_{1,k^*}^{\mathcal{I}_1}$ which is not present in $\mathcal{D}_{2,k^*}^{\mathcal{I}_2}$. As the modified DAGs $\mathcal{D}_{1,k^*}, \mathcal{D}_{2,k^*}$ have the same skeleton, then $\zeta_{k^*} \to v \to w \in \mathcal{D}_{2,k^*}^{\mathcal{I}_2}$. We thus have that w is d-separated from ζ_{k^*} in $\mathcal{D}_{1,k}^{\mathcal{I}_1}$, but not in $\mathcal{D}_{2,k}^{\mathcal{I}_2}$. By the I-Markov property, it follows that $p_{k^*}(x_w) = p_1(x_w)$, while $p_{k^*}(x_w) \neq p_1(x_w)$ and $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ can not be I-Markov equivalent. □

Lemma 30. *Let* $\mathcal{D}_1, \mathcal{D}_2$ *be two DAGs and* $\mathcal{I}_1, \mathcal{I}_2$ *two collections of targets and induced parent sets defining a valid general intervention for both* $\mathcal{D}_1, \mathcal{D}_2$. *Suppose in addition that* $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ *differ only for the reversal of* $u \to v \in \mathcal{D}_1$ *becoming* $v \to u \in \mathcal{D}_2$.

$(\mathcal{D}_1, \mathcal{I}_1), (\mathcal{D}_2, \mathcal{I}_2)$ belong to the same I-Markov equivalence class if and only if $u \to v$ is simultaneously covered in $\mathcal{D}_1$.

Proof. By construction, we have that $\mathcal{I}_1 = \mathcal{I}_2$. Consequently, the proof is identical to the one of Lemma 26. □

Lemma 31. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}_1, \mathcal{I}_2$ two collections of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1, \mathcal{D}_2$. Suppose in addition that $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ differ only for the reversal of $u \to v \in \mathcal{D}_{1,k^*}^{\mathcal{I}_1}$ becoming $v \to u \in \mathcal{D}_{2,k^*}^{\mathcal{I}_2}$, for some $k^* \neq 1$. $(\mathcal{D}_1, \mathcal{I}_1), (\mathcal{D}_2, \mathcal{I}_2)$ belong to the same I-Markov equivalence class if and only if $u \to v$ is covered in $\mathcal{D}_{1,k}^{\mathcal{I}_1}$.*

Proof. By construction, $\mathcal{D}_1 = \mathcal{D}_2$. Consequently, the proof is identical to the one of Lemma 28. □

Theorem 20. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}_1$, $\mathcal{I}_2$ two collections of targets and induced parent sets defining a valid general intervention for both $\mathcal{D}_1, \mathcal{D}_2$. $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ belong to the same I-Markov equivalence class if and only if there exists a sequence of edge reversals modifying the collection of $\mathcal{I}$-DAGs $\{\mathcal{D}_k^{\mathcal{I}_1}\}_{k=1}^K$ and such that:*

1. *Each edge reversed in $\mathcal{D}_1$ is simultaneously covered;*
2. *Each edge reverse in $\mathcal{D}_{1,k}^{\mathcal{I}_1}$, for $k \neq 1$, is covered;*
3. *After each reversal, $\{\mathcal{D}_{1,k}^{\mathcal{I}_1}\}_{k=1}^K$ are DAGs and $(\mathcal{D}_1, \mathcal{I}_1)$, $(\mathcal{D}_2, \mathcal{I}_2)$ belong to the same I-Markov equivalence class;*
4. *After all reversals $\mathcal{D}_{1,k}^{\mathcal{I}_1} = \mathcal{D}_{2,k}^{\mathcal{I}_2}$ for each $k \in [K]$.*

Proof. One can construct a sequence of edge reversals satisfying all the conditions by first using Algorithm 6 with inputs $\mathcal{D}_{1,k}^{\mathcal{I}_1}, \mathcal{D}_{1,k}^{\mathcal{I}_2}$ for $k \in [K], k \neq 1$, and then using the same Algorithm with inputs $\mathcal{D}_1, \mathcal{D}_2$. For each of these two steps, the proofs follow the ones of the corresponding Theorems 13 and 17. □

Appendix B: Proofs of Section 3.3

This section contains the proofs of the main results presented in Section 3.3 of the manuscript. The numbering of such propositions and theorems in this section is the same as in the main text.

Proposition 21. *Given any complete DAG C and a data matrix $\boldsymbol{X}$ collecting observations from K different experimental settings, for any valid pair $(\mathcal{D}, \mathcal{I})$ Assumptions **A1**-**A3** imply*

$$p\left(\boldsymbol{X} \mid \mathcal{D}, \mathcal{I}\right) = \prod_{j=1}^{q} \left\{ \frac{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot fa_{\mathcal{D}}(j)} \mid C\right)}{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot pa_{\mathcal{D}}(j)} \mid C\right)} \prod_{k: j \in T^{(k)}} \frac{p\left(\boldsymbol{X}^{(k)}_{\cdot fa_{\mathcal{D}_k}(j)} \mid C\right)}{p\left(\boldsymbol{X}^{(k)}_{\cdot pa_{\mathcal{D}_k}(j)} \mid C\right)} \right\},$$

where $p\big(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot B} \mid C\big)$ is the marginal data distribution computed under any complete DAG C.

Proof. Using Equations (3.6) and (3.8), together with Assumption **A3**, we can write

$$\begin{aligned}
p\big(\boldsymbol{X} \mid \mathcal{D}, \mathcal{I}\big) &= \int p\big(\boldsymbol{X} \mid \Theta^{(\mathcal{K})}, \mathcal{D}, \mathcal{I}\big)\, p\big(\Theta^{(\mathcal{K})} \mid \mathcal{D}, \mathcal{I}\big)\; d\Theta^{(\mathcal{K})} \\
&= \int \prod_{j=1}^{q} \Bigg\{ p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot j} \mid \boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)}, \Theta^{(1)}_j, \mathcal{D}\right) \prod_{k: j \in T^{(k)}} p\left(\boldsymbol{X}^{(k)}_{\cdot j} \mid \boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(j)}, \Theta^{(k)}_j, \mathcal{D}_k\right) \\
&\qquad\qquad p\left(\Theta^{(1)}_j \mid \mathcal{D}\right) \prod_{k: j \in T^{(k)}} p\left(\Theta^{(k)}_j \mid \mathcal{D}_k\right) \Bigg\}\; d\Theta^{(\mathcal{K})} \\
&= \prod_{j=1}^{q} \Bigg\{ \int p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot j} \mid \boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)}, \Theta^{(1)}_j, \mathcal{D}\right) p\left(\Theta^{(1)}_j \mid \mathcal{D}\right) d\Theta^{(1)}_j \\
&\qquad\qquad \prod_{k: j \in T^{(k)}} \int p\left(\boldsymbol{X}^{(k)}_{\cdot j} \mid \boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(j)}, \Theta^{(k)}_j, \mathcal{D}_k\right) p\left(\Theta^{(k)}_j \mid \mathcal{D}_k\right)\; d\Theta^{(\mathcal{K})} \Bigg\}.
\end{aligned}$$

By Assumption **A2** (likelihood and prior modularity), it follows that

$$\begin{aligned}
p\big(\boldsymbol{X} \mid \mathcal{D}, \mathcal{I}\big) &= \prod_{j=1}^{q} \Bigg\{ \int p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot j} \mid \boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{pa}_{C_j}(j)}, \Theta^{(1)}_j, C_j\right) p\left(\Theta^{(1)}_j \mid C_j\right)\; d\Theta^{(1)} \\
&\qquad\qquad \prod_{k: j \in T^{(k)}} \int p\left(\boldsymbol{X}^{(k)}_{\cdot j} \mid \boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{C_{j,k}}(j)}, \Theta^{(k)}_j, C_{j,k}\right) p\left(\Theta^{(k)}_j \mid C_{j,k}\right)\; d\Theta^{(k)} \Bigg\} \\
&= \prod_{j=1}^{q} \Bigg\{ p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot j} \mid \boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{pa}_{C_j}(j)}, C_j\right) \prod_{k: j \in T^{(k)}} p\left(\boldsymbol{X}^{(k)}_{\cdot j} \mid \boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{C_{j,k}}(j)}, C_{j,k}\right) \Bigg\}.
\end{aligned}$$

Now by Assumption **A1** (complete model equivalence) and recalling that $\mathrm{pa}_{C_j}(j) = \mathrm{pa}_{\mathcal{D}}(j)$

and $\mathrm{pa}_{C_{j,k}}(j) = \mathrm{pa}_{\mathcal{D}_k}(j)$, we obtain

$$p\big(\boldsymbol{X} \,|\, \mathcal{D}, \mathcal{I}\big) = \prod_{j=1}^{q} \left\{ p\left(\boldsymbol{X}_{\cdot j}^{\mathcal{A}(j)} \,|\, \boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)}, C\right) \prod_{k: j \in T^{(k)}} p\left(\boldsymbol{X}_{\cdot j}^{(k)} \,|\, \boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(j)}^{(k)}, C\right) \right\}$$
$$= \prod_{j=1}^{q} \left\{ \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \,|\, C\right)} \prod_{k: j \in T^{(k)}} \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\mathcal{D}_k}(j)}^{(k)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(j)}^{(k)} \,|\, C\right)} \right\},$$

which completes the proof. □

Theorem 22 (Score equivalence)**.** *Let $\mathcal{D}_1, \mathcal{D}_2$ be two DAGs and $\mathcal{I}_1, \mathcal{I}_2$ two collections of targets and induced parent sets defining a valid general intervention for $\mathcal{D}_1, \mathcal{D}_2$ respectively. If $(\mathcal{D}_1, \mathcal{I}_1)$ and $(\mathcal{D}_2, \mathcal{I}_2)$ are I-Markov equivalent, then Assumptions A1-A3 imply*

$$p(\boldsymbol{X} \,|\, \mathcal{D}_1, \mathcal{I}_1) = p(\boldsymbol{X} \,|\, \mathcal{D}_2, \mathcal{I}_2).$$

Proof. By Theorem 20, there exists a sequence of edge reversals applied to either $\mathcal{D}_1$ or $\mathcal{D}_{1,k}^{I}, k \neq 1$ and such that, at the end of the sequence $(\mathcal{D}_1, \mathcal{I}_1) = (\mathcal{D}_2, \mathcal{I}_2)$. Let for simplicity $(\mathcal{D}, \mathcal{I})$ be the pair of DAG and collection of targets and induced parent sets obtained at a given step of the sequence. We can consider the Bayes factor between $(\mathcal{D}, \mathcal{I})$ and $(\widetilde{\mathcal{D}}, \widetilde{\mathcal{I}})$, the corresponding pair obtained at the subsequent step. These two pairs differ for either (i) a simultaneously covered edge reversal or (ii) a covered edge reversal in one of the $\mathcal{I}$-DAGs $\mathcal{D}_k^{\mathcal{I}}, k \neq 1$. In case (i), suppose that $\mathcal{D}, \widetilde{\mathcal{D}}$ differ for the simultaneously covered edge $u \to v \in \mathcal{D}$, which is reversed in $\widetilde{\mathcal{D}}$, while $\mathcal{I} = \widetilde{\mathcal{I}}$. Then

$$\frac{p\big(\boldsymbol{X} \,|\, \mathcal{D}, \mathcal{I}\big)}{p\big(\boldsymbol{X} \,|\, \widetilde{\mathcal{D}}, \widetilde{\mathcal{I}}\big)} = \left(\prod_{j=1}^{q} \left\{ \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \,|\, C\right)} \prod_{k: j \in T^{(k)}} \frac{p\Big(\boldsymbol{X}_{\cdot \mathrm{fa}_{\mathcal{D}_{1,k}}(j)}^{(k)} \,|\, C\Big)}{p\Big(\boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}_{1,k}}(j)}^{(k)} \,|\, C\Big)} \right\} \right) \cdot$$
$$\cdot \left(\prod_{j=1}^{q} \left\{ \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}}(j)}^{\mathcal{A}(j)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}}(j)}^{\mathcal{A}(j)} \,|\, C\right)} \prod_{k: j \in \widetilde{T}^{(k)}} \frac{p\Big(\boldsymbol{X}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}_k}(j)}^{(k)} \,|\, C\Big)}{p\Big(\boldsymbol{X}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}_k}(j)}^{(k)} \,|\, C\Big)} \right\} \right)^{-1}$$
$$= \left(\prod_{j=1}^{q} \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)}^{\mathcal{A}(j)} \,|\, C\right)} \right) \cdot \left(\prod_{j=1}^{q} \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}}(j)}^{\mathcal{A}(j)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}}(j)}^{\mathcal{A}(j)} \,|\, C\right)} \right)^{-1}$$
$$= \left(\frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\mathcal{D}}(u)}^{\mathcal{A}(u)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}}(u)}^{\mathcal{A}(u)} \,|\, C\right)} \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\mathcal{D}}(v)}^{\mathcal{A}(v)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\mathcal{D}}(v)}^{\mathcal{A}(v)} \,|\, C\right)} \right) \cdot \left(\frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}}(u)}^{\mathcal{A}(v)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}}(u)}^{\mathcal{A}(v)} \,|\, C\right)} \frac{p\left(\boldsymbol{X}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}}(v)}^{\mathcal{A}(v)} \,|\, C\right)}{p\left(\boldsymbol{X}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}}(v)}^{\mathcal{A}(v)} \,|\, C\right)} \right)^{-1}.$$

Because $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ differ for the reversal of the simultaneously covered edge $u \to v$, then the following equalities holds:

$$\mathrm{pa}_{\mathcal{D}}(u) = \mathrm{pa}_{\widetilde{\mathcal{D}}}(v), \quad \mathrm{fa}_{\mathcal{D}}(v) = \mathrm{fa}_{\widetilde{\mathcal{D}}}(u), \quad \mathrm{fa}_{\mathcal{D}}(u) = \mathrm{pa}_{\mathcal{D}}(v), \quad \mathrm{fa}_{\widetilde{\mathcal{D}}}(v) = \mathrm{pa}_{\widetilde{\mathcal{D}}}(u). \tag{3.20}$$

Therefore, the ratio simplifies to 1 if $A(u) = A(v)$. To prove this, notice that for any $j \in [q]$

$$\begin{aligned}\mathcal{A}(j) &\coloneqq \big\{k \in [K] : j \notin T^{(k)}\big\} \\ &= \big\{k \in [K] : \zeta_k \notin \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}}}(j)\big\}.\end{aligned}$$

Suppose now $\mathcal{A}(u) \neq \mathcal{A}(v)$. As a consequence, there exists $k \in [K]$ such that $\zeta_k \in \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}}}(u)$, while $\zeta_k \notin \mathrm{pa}_{\mathcal{D}_k^{\mathcal{I}}}(v)$, or viceversa. In both cases, this however would imply that $u \to v$ is not simultaneously covered, which is a contradiction, and therefore $\mathcal{A}(u) = \mathcal{A}(v)$. In case (ii), suppose that, for some $k \in [K]$, $\mathcal{D}_k^{\mathcal{I}}, \widetilde{\mathcal{D}}_k^{\mathcal{I}}$ differ for the covered edge $u \to v \in \mathcal{D}_k^{\mathcal{I}}$, which is reversed in $\widetilde{\mathcal{D}}_k^{\mathcal{I}}$. Then $\mathcal{D} = \widetilde{\mathcal{D}}$ and

$$\begin{aligned}\frac{p\big(\boldsymbol{X} \,|\, \mathcal{D}, \mathcal{I}\big)}{p\big(\boldsymbol{X} \,|\, \widetilde{\mathcal{D}}, \widetilde{\mathcal{I}}\big)} &= \left(\prod_{j=1}^{q} \left\{ \frac{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{fa}_{\mathcal{D}}(j)} \,|\, C\right)}{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{pa}_{\mathcal{D}}(j)} \,|\, C\right)} \prod_{k : j \in T^{(k)}} \frac{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{fa}_{\mathcal{D}_k}(j)} \,|\, C\Big)}{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(j)} \,|\, C\Big)} \right\}\right) \cdot \\ &\quad \cdot \left(\prod_{j=1}^{q} \left\{ \frac{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}}(j)} \,|\, C\right)}{p\left(\boldsymbol{X}^{\mathcal{A}(j)}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}}(j)} \,|\, C\right)} \prod_{k : j \in \widetilde{T}^{(k)}} \frac{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}_k}(j)} \,|\, C\Big)}{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}_k}(j)} \,|\, C\Big)} \right\}\right)^{-1} \\ &= \left(\frac{p\left(\boldsymbol{X}^{(k)}_{\cdot \mathrm{fa}_{\mathcal{D}_k}(u)} \,|\, C\right)}{p\left(\boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(u)} \,|\, C\right)} \frac{p\left(\boldsymbol{X}^{(k)}_{\cdot \mathrm{fa}_{\mathcal{D}_k}(v)} \,|\, C\right)}{p\left(\boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\mathcal{D}_k}(v)} \,|\, C\right)}\right) \cdot \left(\frac{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}_k}(u)} \,|\, C\Big)}{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}_k}(u)} \,|\, C\Big)} \frac{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{fa}_{\widetilde{\mathcal{D}}_k}(v)} \,|\, C\Big)}{p\Big(\boldsymbol{X}^{(k)}_{\cdot \mathrm{pa}_{\widetilde{\mathcal{D}}_k}(v)} \,|\, C\Big)}\right)^{-1},\end{aligned}$$

where the second equality follows from the fact that by the I-Markov equivalence of $\mathcal{I}$ and $\widetilde{\mathcal{I}}$, $\mathcal{T} = \widetilde{\mathcal{T}}$. Since $u \to v$ is covered in the two DAGs, the equalities in (3.20) still hold and the ratio simplifies to 1. □

Appendix C: Proofs of Section 3.4

This section contains the proof of Proposition 24 which establishes the convergence of Algorithms 3 to the posterior distribution $p(\mathcal{D}, \mathcal{T}, \mathcal{P} \,|\, \boldsymbol{X})$.

Proposition 24. *The finite Markov chain defined by Algorithm 3, 4, and 5 is reversible, aperiodic, and irreducible. Accordingly, it has $p(\mathcal{D}, \mathcal{T}, \mathcal{P} \,|\, \boldsymbol{X})$ as its unique stationary dis-*

tribution.

Proof. The reversibility and aperiodicity of Algorithm 3 follows immediately from the properties of the Metropolis-Hastings algorithm [Craiu and Rosenthal, 2014]. To prove irreducibility, notice that if, at each step of the Markov chain, both (i) $p(\tilde{\mathcal{D}}, \mathcal{I} \mid \boldsymbol{X})$ and (ii) the proposal ratio are strictly greater than zero, then evaluating the irreducibility of Algorithm 1 reduces to evaluating the irreducibility of the Markov chain defined by the proposal distribution, illustrated in Algorithm 7. Requirement (i) is trivially satisfied in the case of

Algorithm 7: Markov chain defined by the proposal distribution of Algorithm 3

Input: Number of iterations S, initial values for DAG, targets and induced parent sets $\mathcal{D}^0, \mathcal{T}^0, \mathcal{P}^0$
Output: A sample from a Markov chain over $(\mathcal{D}, \mathcal{T}, \mathcal{P})$

Construct $\left\{\mathcal{D}_k^{s\mathcal{I}}\right\}_{k=1}^K$;
Set $\mathcal{I}^0 = (\mathcal{T}^0, \mathcal{P}^0)$;
for *s in 1:S* **do**
  Sample $\boldsymbol{\pi}$, a permutation vector of length K;
  Set $\{\mathcal{D}^s, \mathcal{I}^s\} = \{\mathcal{D}^{s-1}, \mathcal{I}^{s-1}\}$;
  for *k in 1:K* **do**
    if $\boldsymbol{\pi}_k = 1$ **then**
      Construct $\mathcal{O}_{\mathcal{D}^s}$ using Algorithm 4;
      Sample $\widetilde{\mathcal{D}}$ uniformly at random from $\mathcal{O}_{\mathcal{D}^s}$;
      Set $\mathcal{D}^s = \widetilde{\mathcal{D}}$
    end
  end
  else
    Construct $\mathcal{O}_{\mathcal{D}^{s\mathcal{I}}_{\boldsymbol{\pi}_k}}$ using Algorithm 5;
    Sample $\widetilde{\mathcal{D}}^{\mathcal{I}}_{\boldsymbol{\pi}_k}$ uniformly at random from $\mathcal{O}_{\mathcal{D}^{s\mathcal{I}}_{\boldsymbol{\pi}_k}}$;
    Recover $\widetilde{I}^{(\boldsymbol{\pi}_k)} = (\widetilde{T}^{(\boldsymbol{\pi}_k)}, \widetilde{P}^{(\boldsymbol{\pi}_k)})$ from $(\widetilde{\mathcal{D}}^{\mathcal{I}}_{\boldsymbol{\pi}_k}, \mathcal{D}^s)$;
    Set $I_s^{(\boldsymbol{\pi}_k)} = \widetilde{I}^{(\boldsymbol{\pi}_k)}$
  end
end
Recover $\{\mathcal{T}^s, \mathcal{P}^s\}_{s=1}^S$ from $\{\mathcal{I}^s\}_{s=1}^S$;
return $\{\mathcal{D}^s, \mathcal{T}^s, \mathcal{P}^s\}_{s=1}^S$;

priors on $(\mathcal{D}, \mathcal{I})$ with full support, as both the proposal distributions defined by Algorithm 4 and 5 explicitly take into account the validity requirement while defining the set of possible operators. Condition (ii) is satisfied if each move in the Markov chain is invertible, that is $q(\tilde{\mathcal{D}} \mid \mathcal{D}) > 0$ if and only if $q(\mathcal{D} \mid \tilde{\mathcal{D}}) > 0$. Because of the structure of our proposal

distributions in Algorithms 4 (5) this is equivalent to establish for each type of operator the existence of an *inverse* operator; specifically, we need to prove that if an operator belongs to $\mathcal{O}_{\mathcal{D}}$ ($\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$), then its inverse operator belongs to $\mathcal{O}_{\tilde{\mathcal{D}}}$ ($\mathcal{O}_{\tilde{\mathcal{D}}_k^{\mathcal{I}}}$) too. For $\mathcal{O}_{\mathcal{D}}$, whose construction is based on operators $Insert(u, v)$, $Delete(u, v)$ and $Reverse(u, v)$ applied to $u, v \in [q], u \neq v$, the proof is immediate: $Insert(u, v)$ is the inverse operator of $Delete(u, v)$ and viceversa, while $Reverse(u, v)$ is the inverse operator of $Reverse(v, u)$. The same holds when the three operators are applied to $u, v \in [q]$ for the construction of $\mathcal{O}_{\mathcal{D}_k^{\mathcal{I}}}$. In addition, when operators $Insert$ and $Delete$ involve ζ_k, we have $Insert(\zeta_k, v)$ as the inverse operator of $Delete(\zeta_k, v)$ and viceversa.

We can thus prove the irreducibility of the chain defined by Algorithm 3 by proving the irreducibility of the Markov chain defined by Algorithm 7. At each step s of the algorithm, the proposed value is accepted and the new sequence of $\mathcal{I}$-DAGs $\{\mathcal{D}_{s,k}^{\mathcal{I}}\}_{k=1}^{K}$ is obtained by sequentially updating each $\mathcal{I}$-DAG in a random order defined by the random permutation π_s. Notice that each component-wise update is reversible as shown before. Moreover, any permutation vector π admits an inverse permutation vector. Therefore, to prove the irreducibility of 7, it is sufficient to note that starting from any DAG $\{\tilde{\mathcal{D}}_k^{\mathcal{I}}\}$, it is always possible to reach the sequence of empty $\mathcal{I}$-DAGs $\{\bar{\mathcal{D}}_k^{\mathcal{I}}\}_{k=1}^{K}$ by repeated edge deletions. By reversibility, this implies that it is always possible to reach any DAG starting from any other DAG. As the irreducibility of 7 implies the irreducibility of 3, the result follows. □

Bibliography

M. M. Barbieri and J. O. Berger. Optimal Predictive Model Selection. *The Annals of Statistics*, 32(3):870–897, 2004.

A. Belyaeva, C. Squires, and C. Uhler. DCI: Learning Causal Differences between Gene Eegulatory Networks. *Bioinformatics*, 37(18):3067–3069, 2021.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of Structural Causal Models with Cycles and Latent Variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

S. Brooks, A. Gelman, G. Jones, and X. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

F. Castelletti and G. Consonni. Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. *The Annals of Applied Statistics*, 13(4):2289–2311, 2019.

F. Castelletti and G. Consonni. Bayesian Graphical Modeling for Heterogeneous Causal Effects. *Statistics in Medicine*, 42(1):15–32, 2023.

F. Castelletti and S. Peluso. Network Structure Learning Under Uncertain Interventions. *Journal of the American Statistical Association*, 118(543):2117–2128, 2023a.

F. Castelletti and S. Peluso. Bayesian Learning of Network Structures from Interventional Experimental Data. *Biometrika*, 05 2023b.

F. Castelletti, L. La Rocca, S. Peluso, F. C. Stingo, and G. Consonni. Bayesian Learning of Multiple Directed Networks from Observational Data. *Statistics in Medicine*, 39(30): 4745–4766, 2020.

D. M. Chickering. A transformational characterization of equivalent bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 87–98, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

G. F. Cooper and C. Yoo. Causal Discovery from a Mixture of Experimental and Observational Data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 116–125, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

J. Correa and E. Bareinboim. A Calculus For Stochastic Interventions: Causal Effect Identification and Surrogate Experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

R. Craiu and J. S. Rosenthal. Bayesian Computation Via Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, 1(1):179–201, 2014.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 107–114, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

K. J. Friston. Functional and Effective Connectivity: A Review. *Brain Connectivity*, 1(1): 13–36, 2011.

J. L. Gamella, A. Taeb, C. Heinze-Deml, and P. Bühlmann. Characterization and Greedy Learning of Gaussian Structural Causal Models under Unknown Interventions, 2022.

D. Geiger and D. Heckerman. Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions. *The Annals of Statistics*, 30 (5):1412–1440, 2002.

A. Hauser and P. Bühlmann. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *Journal of Machine Learning Research*, 13(79):2409–2464, 2012.

A. Hauser and P. Bühlmann. Jointly Interventional and Observational Data: Estimation of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):291–318, 05 2015.

D. Heckerman and D. Geiger. Learning Bayesian Networks: A Unification for Discrete and Gaussian Domains. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 274–284, 01 1995.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20:197–243, 1995.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561. Curran Associates, Inc., 2020.

S. M. Kornblau, R. Tibes, Y. H. Qiu, W. Chen, H. M. Kantarjian, M. Andreeff, K. R. Coombes, and G. B. Mills. Functional Proteomic Profiling of AML Predicts Response and Survival. *Blood*, 1(113):154–164, 2009.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

J. Pearl and J. Robins. Probabilistic Evaluation of Sequential Plans from Causal Models with Hidden Variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 444–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

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

C. Peterson, F. C. Stingo, and M. Vannucci. Bayesian Inference of Multiple Gaussian Graphical Models, journal = Journal of the American Statistical Association. 110(509): 159–174, 2015.

S. J. Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference, Second Edition*. Dover Publications, 2012.

M. R. Ricciardi, S. Mirabilii, R. Licchetta, M. Piedimonte, and A. Tafuri. Targeting the Akt, GSK-3, Bcl-2 Axis in Acute Myeloid Leukemia. *Advances in biological regulation*, 65:36–58, 2017.

A. Roverato and G. Consonni. Compatible Prior Distributions for Directed Acyclic Graph Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1): 47–61, 12 2003.

P. P. Ruvolo, Y. Qiu, K. R. Coombes, N. Zhang, E. S. Neeley, V. R. Ruvolo, N. J. Hail, G. Borthakur, M. Konopleva, M. Andreeff, and S. M. Kornblau. Phosphorylation of GSK3α/β Correlates with Activation of AKT and is Prognostic for Poor Overall Survival in Acute Myeloid Leukemia Patients. *BBA Clinical*, 4:59–68, 2015.

A. Shojaie. Differential Network Analysis: A Statistical Perspective. *WIREs Computational Statistics*, 13(2):1508, 2021.

C. Squires, Y. Wang, and C. Uhler. Permutation-Based Causal Structure Learning with Unknown Intervention Targets. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pages 1039–1048. PMLR, 03–06 Aug 2020.

J. Tian and J. Pearl. Causal Discovery from Changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 512–521, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

T. Verma and J. Pearl. Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI 90, pages 255–270, New York, NY, USA, 1990. Elsevier Science Inc.

Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation-based Causal Inference Algorithms with Interventions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,

S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5822–5831. Curran Associates, Inc., 2017.

Y. Wang, C. Squires, A. Belyaeva, and C. Uhler. Direct Estimation of Differences in Causal Graphs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3770–3781. Curran Associates, Inc., 2018.

K. Yang, A. Katcoff, and C. Uhler. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5541–5550. PMLR, 10–15 Jul 2018.