Estimating Joint Interventional Distributions from Marginal Interventional Data

Sergio Hernan Garrido Mejia SHGM@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Elke Kirschbaum
Amazon
Tübingen, Germany

Armin Kekić
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
ELLIS Institute
Tübingen, Germany

Atalanti Mastakouri
Amazon
Tübingen, Germany

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

In this paper we show how to exploit *interventional* data to acquire the joint conditional distribution of all the variables using the Maximum Entropy principle. To this end, we extend the Causal Maximum Entropy method to make use of data arising from identifiable interventional distributions in addition to data from the observational distribution. Using Lagrange duality, we prove that the solution to the Causal Maximum Entropy problem with interventional constraints lies in the exponential family, as in the Maximum Entropy solution. Our method allows us to perform two tasks of interest when marginal interventional distributions are provided for any subset of the variables. First, we show how to perform parental discovery from a mixture of observational and single-variable interventional data, and, second, how to infer joint interventional distributions. For the former task, we show on synthetically generated data, that our proposed method outperforms the state-of-the-art method on merging datasets, and yields comparable results to the KCI-test which requires access to joint observations of *all* variables.

Keywords: Causal Maximum Entropy, Causal Marginal Problem, Parental Discovery

1. Introduction

Randomised Controlled Trials (RCTs) are the standard method for identifying the causal effect of a treatment on a target variable. However, their utility is often constrained by cost when studying complex interventions. Determining how a new drug, for instance, interacts with existing medications or medical procedures requires testing a number of combinations that grows exponentially. Furthermore, research goals frequently extend beyond the average treatment effect to understanding its impact under specific combinations of conditions (*e.g.*, medical pre-conditions or risk factors).

While this conditional knowledge is crucial and must be recorded during the trial, designing an RCT capable of considering all potentially relevant factors proves challenging. This difficulty often leads to research where multiple independent studies investigate different aspects or subsets of treatments and conditions. Consequently, researchers rarely have access to the full conditional distribution of the target variable or the joint interventional distribution of multiple treatments. This fundamental lack of joint data complicates the identification of whether a treatment or condition has a direct causal effect or merely an indirect influence through another factor.

As another real-world example beyond the medical domain, consider the problem of finding the effect of different fertilisers and planting methods on a particular crop yield (Hindersah et al., 2022). As the number of fertilisers and planting methods increases, the experimental design becomes prohibitively expensive due to combinatorial explosion. Nevertheless, a researcher interested in the combined effect might have observational data and data from single experiments, such as nitrogen (Qiu et al., 2022) or potassium (Wihardjaka et al., 2022) fertilisers on crop yield. To address this challenge, we propose a method for combining experimental and observational data to infer joint interventional distributions.

To do this, we extend the Causal Maximum Entropy (CMAXENT) principle (Janzing, 2021) to the interventional-CMAXENT (*i-CMAXENT*). We show that the resulting distribution for i-CMAXENT lies in the exponential family, similar to the traditional Maximum Entropy (MAXENT) distribution (Wainwright et al., 2008). This allows us to perform two tasks of interest: (i) We can estimate joint interventional distributions from single variable interventions (Saengkyongam and Silva, 2020; Kekić et al., 2025) in the marginal causal problem setting (Gresele et al., 2022). (ii) We can perform parental discovery (Peters et al., 2016; Heinze-Deml et al., 2018) under the causal marginal setting. That is, under certain graph constraints, it allows us to infer the true causal parents of a variable of interest from a set of potential causes, even when the variables are not jointly observed.

The contributions of this paper are as follows:

- We extend the Causal Maximum Entropy principle (CMAXENT) to use data from experimental conditions (Section 5).
- We prove that given a set of statistical constraints, the solution to the proposed method lies in an exponential family of distributions, extending the scope of these to include causal semantics (Section 5).
- We provide conditions under which this method can be used to combine observational and experimental data (Section 5), which is only possible nonparametrically under strict constraints (Section 3), and test the results empirically against usual CMAXENT (Sections 6 and 7).

2. Related work

Marginal problem and causality Various methods address the statistical problem of merging information from datasets with overlapping subsets of random variables, called the *Marginal Problem* (Deming and Stephan, 1940; Kellerer, 1964). The problem of combining information from overlapping data in causal structure learning has been studied in (Danks et al., 2008; Tillman and Spirtes, 2011; Dhir and Lee, 2020). However, the structures they are able to learn are based on conditional independence tests within the overlapping datasets; in other words, they have to observe variables jointly. Recently, an extension of this problem was introduced, the *Causal Marginal*

Problem, where information from disjoint datasets is merged into a single *causal* model (Gresele et al., 2022; Sani et al., 2023). Gresele et al. (2022), for example, address whether data from subsets of variables together with a known graph structure can be used to determine a set of joint SCMs that are counterfactually consistent with the marginal data. Their work differs from ours in that we are interested in finding joint interventional distributions, while they focus on bounding counterfactual quantities, which allows them to falsify causal models. Further, they assume the causal graph to be given, while our approach can be used for parental discovery.

Joint interventional distributions A recent object of research interest is the conditions under which joint causal effects can be identified from interventions on variable subsets, single variable interventions being the limiting case. This question has been tackled from both a parametric and a nonparametric point of view. From the nonparametric perspective, causal effect identification from experimental data was studied by Bareinboim and Pearl (2012), where a sound and complete algorithm was introduced for z-identification of causal effects under experimental conditions. This was extended later by Lee et al. (2020), where the g-identification formula was introduced in scenarios where not all experimental data is available. Jung et al. (2023) presented an estimator based on the g-identification formula with properties against bias. Tikka et al. (2021) developed a sound search-based algorithm, where heuristics and search reduction techniques are able to decide (and provide a nonparametric estimator, if possible) for joint effect identification from observational and experimental data.

From the parametric perspective, Saengkyongam and Silva (2020) studies identification in non-linear Gaussian models with confounding between the covariates and the outcomes, while Kekić et al. (2025) provides an unbiased estimator of joint interventional effects from single variable interventions without distributional assumptions. However, they rely on an additive outcome mechanism assumption. Gimenez and Rothenhäusler (2022), explore the identifiability of joint causal effects under different assumptions of the problem such as linear and nonlinear models, the existence of instrumental variables and sparsity. The potential outcome literature has also studied this question from a parametric point of view (see (Shi et al., 2023; Colnet et al., 2024) and references therein).

The complementary question of the conditions under which single variable interventions can be identified from joint causal effects has also been studied. Jeunen et al. (2022), for example, find the conditions under which single variable interventions can be identified from joint interventions in confounded additive noise models. In addition to the previous result, Elahi et al. (2024) study how to obtain all possible causal effects from only some joint interventions in additive noise models with Gaussian noise.

Causal discovery The task of parental discovery has been addressed from various perspectives, with the most prominent ones being those based on invariant causal prediction (Peters et al., 2016; Heinze-Deml et al., 2018). While these methods are powerful and do not assume knowledge about which variables in the system were intervened upon, they cannot operate in the setting of marginally observed sets of variables, such as the causal marginal problem.

Furthermore, other research has focused on causal discovery using observational and experimental data using either Bayesian inference methods (Cooper and Yoo, 1999; Eaton and Murphy, 2007), causal graph conditions (Tian and Pearl, 2001), meta-analysis methods for combining p-values (Tillman, 2009), or constraint-based methods (Triantafillou and Tsamardinos, 2015). A particularly important idea in this line of research is the Joint Causal Inference (JCI) framework (Mooij et al.,

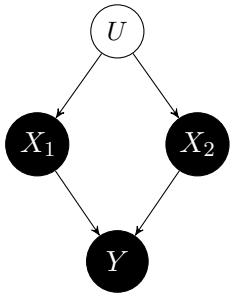


Figure 1: Graph that allows nonparametric marginal data combination

2020), where these ideas were unified. However, none of these methods work under the causal marginal problem, which limits their application in heavy missing data problems.

A method that allows for parental discovery, even under this marginal setting is CMAXENT (Garrido Mejia et al., 2022). However, CMAXENT can only make use of observational data, limiting the amount of information that it can leverage to find the relevant causal parents. An extensive comparison of our approach with CMAXENT can be found in Section 7.

3. Motivation

In this section we will prove that even in cases where the parents of the target variable are confounded we can identify nonparametric joint interventional effects from observational or interventional marginal data. This holds true even in cases where these joint interventional distributions cannot be identified using the rules of do-calculus (Pearl, 2009).

Consider a causal system with binary variables X_1, X_2, Y and U, where U is an unobserved confounder, and the joint distribution $\mathrm{p}(\mathbf{X}, Y) := \mathrm{p}(X_1, X_2, Y)$ is Markov relative to the graph in Figure 1. We are interested in $\mathrm{p}(Y \mid \mathrm{do}(X_1 = x_1, X_2 = x_2))$, which in the case of sufficient causal systems would correspond to $\mathrm{p}(Y \mid X_1 = x_1, X_2 = x_2)$. We have $\mathrm{p}(Y \mid X_1)$, $\mathrm{p}(Y \mid X_2)$ and $\mathrm{p}(\mathbf{X})$ as data. Using a modern interventional distribution identification engine (Tikka et al., 2021), we verify that the joint interventional distribution is not identifiable using the given data. Nonetheless, we can still identify the joint interventional distribution nonparametrically:

Proposition 1 (Nonparametric identification through observational data combination) *Let $\mathbf{X}, Y$ be binary random variables and suppose $\mathrm{p}(\mathbf{x}) > 0$, for all $\mathbf{x}$. Then the joint interventional distribution $\mathrm{p}(Y \mid \mathrm{do}(X_1 = x_1, X_2 = x_2))$ is identifiable using $\mathrm{p}(\mathbf{X})$ and $\mathrm{p}(Y \mid X_i)$ for $i = 1, 2$.*

Remark 2 *In the statement above, it is not necessary that X_i does not cause X_j.*

The proof is based on the fact that there are four equations, namely one for each $\mathrm{p}(Y \mid X_i = x_i)$, and four unknowns $\mathrm{p}(Y \mid X_1 = x_1, X_2 = x_2)$. All proofs can be found in Appendix A. In the particular case of the graph in $Figure\ 1$ the backdoor criterion holds, so that if we had interventional distributions instead of observational ones, the target distribution is also identifiable, giving us the following Corollary:

Corollary 3 (Nonparametric identification through single variable intervention data combination) *Let $\mathbf{X}, Y$ be binary random variables and suppose $\mathrm{p}(\mathbf{x}) > 0$, for all $\mathbf{x}$. Then the joint interventional*

distribution $\mathrm{p}(Y \mid \mathrm{do}(X_1 = x_1, X_2 = x_2))$ *is identifiable using* $\mathrm{p}(\mathbf{X})$ *and the single-variable interventions given by* $\mathrm{p}(Y \mid \mathrm{do}(X_i = x_i))$ *for* $i = 1, 2$*, as long as each* $\mathrm{p}(Y \mid \mathrm{do}(X_i = x_i))$ *is identifiable.*

We see how we cannot apply this idea for more than two variables using either only observational or interventional data. Indeed, the number of unknowns grows exponentially with the number of variables, whereas the number of equations grows linearly. Nevertheless, the same result can be applied once more by combining observational and interventional data in the case of four variables:

Corollary 4 *Let* $\mathbf{X} = \{X_i : i = 1, 2, 3, 4\}, Y$ *be binary random variables and* $\mathrm{p}(\mathbf{x}) > 0$ *for all* $\mathbf{x}$*. If* $\mathrm{p}(Y \mid \mathrm{do}(X_i = x_i))$ *are identifiable using the backdoor criterion for all* i*, we can identify the joint interventional distribution* $\mathrm{p}(Y \mid \mathrm{do}(\mathbf{X}))$ *using* $\mathrm{p}(\mathbf{X})$*,* $\mathrm{p}(Y \mid X_i)$ *and* $\mathrm{p}(Y \mid \mathrm{do}(X_i = x_i))$ *for all* i*.*

This result motivates our search for methods that allow us to combine both observational and interventional data beyond four variables.

4. Causal Maximum Entropy

In this section, we formally state the Maximum Conditional Entropy problem (Koller and Friedman (2009, Chapters 8 and 20), Berger et al. (1996)), and its causal interpretation, Causal Maximum Entropy (CMAXENT) (Sun et al., 2006; Janzing, 2021; Garrido Mejia et al., 2022), which we will use as a basis for our proposed model in Section 5.

Consider a set of D random variables $\mathbf{X} = \{X_1, \ldots, X_D\}$ of *potential causes*, and an *effect* Y, with realisations $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$. We define P to be a probability measure on $\mathcal{Y} \times \mathcal{X}$.

Let $\mathbf{F} = \{f_k\}$ be a set of K *marginal* measurable functions with $f_k : \mathcal{Y} \times \mathcal{X}_{S_k} \to \mathbb{R}$, where $S_i \subseteq \{1, \ldots, D\}$ is an index set of potential causes. For these functions, we are given empirical averages denoted by $\tilde{f}_k := \frac{1}{N} \sum_{i=1}^{N} f_k(y^i, \mathbf{x}^i_{S_k})$, where N is the number of observations of y^i and $\mathbf{x}^i_{S_k}$. The expectations of the marginal functions can be computed with respect to any valid joint density of Y and $\mathbf{X}$: $\mathbb{E}[f_k(Y, \mathbf{X}_{S_k})] = \sum_{y,\mathbf{x}} \mathrm{p}(y, \mathbf{x}) f_k(y, \mathbf{x}_{S_k})$.

Furthermore, let J be a set of *conditional* measurable functions $\mathbf{G} = \{g_j\}$ with $g_j : \mathcal{Y} \times \mathcal{X}_{S_j} \to \mathbb{R}^{|\mathbf{X}_{S_j}|}$. For conditional functions, the empirical averages $\tilde{g}_j(\mathbf{x}_{S_j}) := \frac{1}{N} \sum_{i=1}^{N} g_j(y^i, \mathbf{x}_{S_j})$ depend on the value of the conditioning variables X_{S_j}, where the conditioning set, S_j is different for every conditional function g_j. The expectations are computed with respect to a valid conditional distribution: $\mathbb{E}[g_j(Y, \mathbf{x}_{S_j}) \mid \mathbf{X}_{S_j} = \mathbf{x}_{S_j}] = \sum_y \mathrm{p}(y \mid \mathbf{x}_{S_j}) g_j(y, \mathbf{x}_{S_j})$.

Suppose we are given a density of the potential causes $\mathrm{p}(\mathbf{X})$ and the sets of functions $\mathbf{F}$ and $\mathbf{G}$ with their respective empirical averages, but we do not know the conditional density $\mathrm{p}(Y \mid \mathbf{X})$. Then the principle of Maximum Conditional Entropy suggests choosing the conditional density $\mathrm{p}_\lambda(Y \mid \mathbf{X})$ that has expectations consistent with the given empirical averages and also maximises the Shannon conditional entropy $H(Y \mid X)$ (Jaynes, 1957; Berger et al., 1996; Farnia and Tse, 2016):

$$
\begin{aligned}
\max_{\mathrm{p}(y|\mathbf{x})} \quad & H(Y \mid \mathbf{X}) = -\sum_{y,\mathbf{x}} \mathrm{p}(y \mid \mathbf{x}) \mathrm{p}(\mathbf{x}) \log \mathrm{p}(y \mid \mathbf{x}) \\
\text{s.t.} \quad & \mathbb{E}[f_k(Y, \mathbf{X}_{S_k})] = \tilde{f}_k, \ \text{ for all } k = 1, \ldots, K \\
& \mathbb{E}[g_j(Y, \mathbf{x}_{S_j}) \mid \mathbf{X}_{S_j} = \mathbf{x}_{S_j}] = \tilde{g}_j(\mathbf{x}_{S_j}), \ \text{ for all } j = 1, \ldots, J \\
& \sum_y \mathrm{p}(y \mid \mathbf{x}) = 1, \ \text{ for all } \mathbf{x}.
\end{aligned}
\tag{1}
$$

Solving this optimisation problem, using the Lagrange multiplier formalism, yields an exponential family distribution:

$$\mathrm{p}_\lambda(y \mid \mathbf{x}) = \exp\left(\sum_{k=1}^{K} \lambda_k f_k(y, \mathbf{x}_{S_k}) + \sum_{j=1}^{J} \sum_{\mathbf{x}_{S_j}} \lambda_j^{\mathbf{x}_{S_j}} g_j(y, \mathbf{x}) + \alpha(\mathbf{x})\right), \tag{2}$$

where $\alpha(\mathbf{x})$ is the normalising constant and $\lambda = \{\lambda_k, \lambda_j^{\mathbf{x}_{S_j}}\}$ are the Lagrange multipliers.

Intuitively, this density with maximum entropy is as close as possible to the uniform while satisfying the expectation constraints. Jaynes (2003, Chapter 11) interprets MAXENT as a way to find a density without introducing more information than given by the data. Critically, MAXENT, Maximum Conditional Entropy, or CMAXENT do not aim at estimating the "true distribution" of the data, but instead provide a guess given certain statistical properties of the data. The resulting MAXENT distribution posses powerful statistical properties that have been deeply studied in the literature (Grünwald and Dawid, 2004; Wainwright et al., 2008; Farnia and Tse, 2016).

In CMAXENT (Sun et al., 2006; Janzing et al., 2009; Janzing, 2021; Garrido Mejia et al., 2022), causal semantics are introduced via graphical models. In other words, we assume that the variables in our system have some cause-effect relation that can be represented by a causal graph (Pearl, 2009). In CMAXENT, the densities are computed in the causal order given by a (hypothesised) causal graph. For example, if $\mathbf{X}$ are potential causes of Y, we first find the MAXENT density of $\mathbf{X}$ subject to constraints on $\mathbf{X}$, and then the density with Maximum Conditional Entropy of Y given $\mathbf{X}$ using the found $\mathrm{p}(\mathbf{X})$ and subject to the constraints associated with $\mathbf{X}$ and Y. Because of the introduced causal semantics, we can use expectations involving interventions in the estimation of the density $\mathrm{p}_\lambda(y \mid \mathbf{x})$, which corresponds to the so-called Independent Causal Mechanisms (Schölkopf et al., 2012) of Y given its potential causes $\mathbf{X}$.

The use of interventional data would not be possible for non-causal Maximum Conditional Entropy, since the conditional distribution there is devoid of any causal structure and hence cannot be used for operations of the causal hierarchy like interventions or counterfactuals (Pearl and Mackenzie, 2018). In Section 5, we exploit this observation to extend the CMAXENT method to use data from interventions.

5. Interventional CMAXENT (i-CMAXENT)

In this section, we introduce i-CMAXENT , the modification of CMAXENT to include data on interventional distributions, and prove that the solution to the optimisation problem is an exponential family distribution, as in traditional MAXENT.

Before formally introducing i-CMAXENT , we present the necessary assumptions needed to use interventional data in the optimisation problem.

Assumption 5 (Weak causal sufficiency) *We assume there are no unobserved confounders between the effect Y and its potential causes $\mathbf{X}$. However, there* can *exist hidden confounders among the potential causes.*

Assumption 5 is needed to exclude any backdoor paths between any potential cause and the target Y. More specifically, if hidden confounders exist between $\mathbf{X}$ and Y, then we cannot phrase the interventional moments in terms of the distributions of observed nodes alone.

Assumption 6 (Positivity of the potential causes) *We assume* $\mathrm{p}(\mathbf{x}) > 0$ *for all* $\mathbf{x}$.

Assumption 6 is required to ensure well-defined conditional distributions for all values $\mathbf{x}$ of $\mathbf{X}$.

We also assume Faithful f-expectations as in Garrido Mejia et al. (2022, Definition 1), this assumption is used to relate the results of statistical estimation with a causal graph, as in the usual faithfulness assumption.

Assumption 7 (Faithful f-Expectations) *A distribution* P *is said to have faithful* f*-expectations relative to a DAG* $\mathcal{G}$*, if for any distribution* Q *Markov relative to* $\mathcal{G}$ *it holds that whenever* $\lambda^{\mathrm{Q}}_{f_k} \neq 0$*, then* $\lambda^{\mathrm{P}}_{f_k} \neq 0$.

For an intuitive explanation, consider a triplet X, Y, Z drawn from a distribution P Markov relative to a DAG $\mathcal{G}$, for which $X \not\perp\!\!\!\perp_{\mathrm{P}} Y \mid Z$. Then according to Assumption 7, the dependence of X and Y after observing Z, will also hold in the projected space of the MAXENT distribution. By contrapositive, an independence in the space of MAXENT will also hold in the original distribution space. In Appendix F we discuss other definitions of faithfulness in the literature and how Faithful f-expectations relates to these different notions.

Next, let us introduce the required additional notation. Let $\mathbf{H} = \{h_l\}$ be a set of L *interventional functions* for which the empirical averages are denoted by $\tilde{\mathbf{H}}$. Further, suppose that, in addition to a conditioning set S^C_l, we also have a set S^I_l of variables that were intervened upon. For these functions, the expectations are computed with respect to the interventional density $\mathrm{p}(Y \mid \mathrm{do}(\mathbf{X}_{S^I_l}), \mathbf{X}_{S^C_l})$:

$$\mathbb{E}[h_l(Y, \mathbf{x}_{S^I_l}, \mathbf{x}_{S^C_l}) \mid \mathrm{do}(\mathbf{X}_{S^I_l} = \mathbf{x}_{S^I_l}), \mathbf{X}_{S^C_l} = \mathbf{x}_{S^C_l}] = \sum_y \mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S^I_l}), \mathbf{x}_{S^C_l}) h_l(y, \mathbf{x}_{S^I_l}, \mathbf{x}_{S^C_l}). \tag{3}$$

We assume $S^C_l \cap S^I_l = \emptyset$ for all indices l, as otherwise it would imply that the variables in the distributions we used to compute the empirical averages are both conditioned and intervened upon.

The reasoning behind the extension to i-CMAXENT is the following: for a given interventional function h_l, if the corresponding interventional density $\mathrm{p}(Y \mid \mathrm{do}(\mathbf{X}_{S^I_l}), \mathbf{X}_{S^C_l})$ is identifiable, we can express the expectation in Equation (3) using the observational densities $\mathrm{p}(Y \mid \mathbf{X})$ and $\mathrm{p}(\mathbf{X})$. In this case, we can estimate $\mathrm{p}_\lambda(Y \mid \mathbf{X})$ by fitting the parameters λ such that the interventional expectations under $\mathrm{p}_\lambda(Y \mid \mathbf{X})$ are as close as possible to those given as constraints. In other words, we leverage graphical nonparametric identification to write the interventional expression in terms of observational probabilities, allowing us to introduce interventional data as constraints for the joint density $\mathrm{p}_\lambda(Y, \mathbf{X})$. As above, we find the distribution that maximises the conditional entropy such that its interventional expectations are close to the observed empirical averages.

We obtain the optimisation problem for i-CMAXENT by adding the following constraint to Equation (1):

$$\mathbb{E}[h_l(Y, \mathbf{x}_{S^I_l}, \mathbf{x}_{S^C_l}) \mid \mathrm{do}(\mathbf{X}_{S^I_l} = \mathbf{x}_{S^I_l}), \mathbf{X}_{S^C_l} = \mathbf{x}_{S^C_l}] = \tilde{h}_l(\mathbf{x}_{S^I_l}, \mathbf{x}_{S^C_l}), \text{ for all } l = 1, \ldots, L. \tag{4}$$

Note that even though $\mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S^I_l}), \mathbf{x}_{S^C_l})$ appears as part of the constraints, we consider it to be identifiable and hence can compute this interventional distribution as a functional of observational distributions.

By including empirical averages that come from interventions on $\mathbf{X}$, we can use a richer class of data sources in comparison to CMAXENT, which could only use observational data. In the following theorem, we study the resulting distribution of the i-CMAXENT optimisation problem.

Theorem 8 (Exponential family of i-CMAXENT) *Using the Lagrange multiplier formalism, the solution of Equation (1) with the additional constraint from Equation (4) is given by the following exponential family:*

$$\begin{aligned} \mathrm{p}_{\lambda}(y \mid \mathbf{x}) = \exp\Bigg(& \sum_{k=1}^{K} \lambda_k f_k(y, \mathbf{x}) + \sum_{j=1}^{J} \sum_{\mathbf{x}_{S_j}} \lambda_j^{\mathbf{x}_{S_j}} g_j(y, \mathbf{x}) \\ & + \sum_{l=1}^{L} \sum_{\mathbf{x}_{S_l^I}, \mathbf{x}_{S_l^C}} \lambda_l^{\mathbf{x}_{S_l^I}, \mathbf{x}_{S_l^C}} h_l(y, \mathbf{x}_{S_l^I}, \mathbf{x}_{S_l^C}) + \beta(\mathbf{x}) \Bigg), \end{aligned} \tag{5}$$

where $\beta(\mathbf{x})$ is the normalising constant, as in the conditional case.

In Theorem 8, we show that the well-known exponential family solutions of MAXENT and Maximum Conditional Entropy (Wainwright et al., 2008; Farnia and Tse, 2016) can be generalised to the i-CMAXENT solution in an intuitive way. The proof is shown in Appendix A.

i-CMAXENT for parental discovery In Garrido Mejia et al. (2022), we showed that causal edges can be inferred from the estimated Lagrange multipliers of the solution of the CMAXENT problem. Since we have Assumption 7, which is required for those results to hold, and the solution of i-CMAXENT is an exponential family distribution, we can also use i-CMAXENT to infer causal edges. Note that we need to know which interventional distributions are identifiable without requiring knowledge of which of the *potential* causes *actually* has a direct causal link to the effect variable if we want to use i-CMAXENT for parental discovery. The following proposition shows that this is possible.

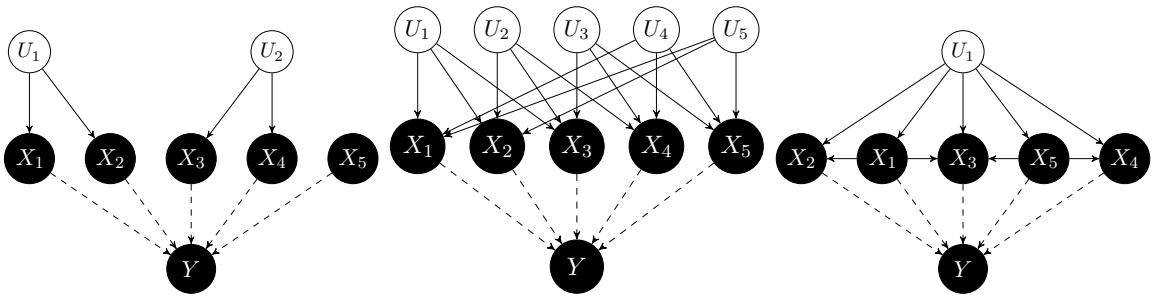
Proposition 9 (Identifiability and adjustment set of variables with only incoming arrows) *Let $\mathbf{X}$ be a set of candidate causal parents of Y, which can be confounded. Assume we know the density $\mathrm{p}(\mathbf{X})$. If the only child of $X_j \in \mathbf{X}$ is potentially, but not necessarily, Y, then $\mathrm{p}(Y \mid \mathrm{do}(X_j))$ is identifiable and a valid adjustment set for the atomic intervention is $\mathbf{X} \setminus X_j$. That is, the rest of the potential causes.*

A different way of thinking about Proposition 9 is that we can decide whether we can use interventional data in an i-CMAXENT estimation only by looking at the arrows between $\mathbf{X}$ and without any knowledge of the arrows between $\mathbf{X}$ and Y.

6. Experiments

We test i-CMAXENT both for parental discovery and for joint interventional distribution estimation. In all experiments, we find the Lagrange multipliers of the exponential family distribution by minimising the norm of the residuals between empirical averages and the corresponding expectations (see Appendix B for details on the norm minimisation and Appendix E for details about the optimisation convergence).

Synthetic data generation All used causal graphs comply with Assumption 5 and consist of three levels: unobserved confounders $\mathbf{U}$, potential causes $\mathbf{X}$, and an effect Y. While we assume causal sufficiency for the lower part of the graph, that is, no hidden confounders between $\mathbf{X}$ and Y, we do allow for hidden confounders $\mathbf{U}$ to exist among the potential causes $\mathbf{X}$. The considered graph structures are shown in Figures 2(*a*) to 2(*c*). Further details on synthetic data generation are in Appendix D.



(*a*) Structure graph (a) (*b*) Structure graph (b) (*c*) Structure graph (c)

Figure 2: (Figure 2(*a*)), (Figure 2(*b*)), and (Figure 2(*c*)) show the graph structures used for our synthetic experiments. We randomise the presence of the edges in the lower part of the graphs (dashed arrows). The solid arrows are always present in the shown way.

Causal discovery The main question of this task is: from a set of *potential* causes $\mathbf{X}$ of our variable of interest Y, which of those variables are *actual* causal parents? To keep our model as comparable as possible with previous work on CMAXENT, we use the same graph structures as in Garrido Mejia et al. (2022). For each structure in Figures 2(*a*) to 2(*c*), we sample 200 random graphs as explained in Appendix D. When randomising the true causes of Y, we ensure that there is at least one causal parent and that at least one potential parent is not a true cause, so that the ROC curve is always defined.

Parental discovery benchmarking i-CMAXENT , CMAXENT, and KCI In this setting, we compare i-CMAXENT against CMAXENT and the Kernel Conditional Independence (KCI) test (Zhang et al., 2011). For i-CMAXENT , we use constraints on the single-variable interventional densities $\mathrm{p}(Y \mid \mathrm{do}(X_i))$ (in the case of binary variables, this coincides with $\mathbb{E}\left[Y \mid \mathrm{do}(X_i)\right]$) for all potential causes X_i. Since we have five potential causes in each graph, this means we have data on five interventional distributions. For CMAXENT, we use constraints on the marginal conditional densities $\mathrm{p}(Y \mid X_i)$ (in the case of binary variables, this coincides with $\mathbb{E}\left[Y \mid X_i\right]$) for each of the five potential causes X_i. Hence, we again have data on five conditional distributions in this case. For both i-CMAXENT and CMAXENT, we consider two cases: in the first case, we are also given data on the full joint observational density $\mathrm{p}(\mathbf{X})$. In the second case, only the marginal observational densities $\mathrm{p}(X_1), \ldots, \mathrm{p}(X_5)$ are given.

In the second case, we estimate $\mathrm{p}(\mathbf{X})$ by merging the constraints on the marginals, also using MAXENT. We generate 100 observations per case and per sampled graph. From these samples, we compute the empirical averages used as constraints in the optimisation procedure. For the KCI test, we assume that KCI has access to 1,000 samples from the joint density $\mathrm{p}(\mathbf{X}, Y)$. This does not reflect the causal marginal problem we are considering in this paper. In fact, KCI could not be run in such a scenario. [1] Nevertheless, we benchmark our method against KCI to show what performance can be achieved on this task with access to the full joint observational distribution.

Parental discovery with combinations of interventional and conditional information using i-CMAXENT In this setting, we evaluate the performance of i-CMAXENT for parental discovery

1. Similarly, ICP (Peters et al., 2016) is not designed for the causal marginal problem. Hence, we do not compare against it.

when interventional information is available for a fraction of the variables, while for the other variables we only have conditional information. With this experiment, we want to emulate a scenario that often occurs in real datasets: namely, that not all variables are intervenable. In this scenario, we assume that we are given data on the full joint observational distribution $\mathrm{p}(\mathbf{X})$ of the potential causes.

We use Proposition 9 to decide which potential parent can provide interventional instead of conditional data. For example, in Figure 2(*c*), it follows from Proposition 9 that we can use $\mathrm{p}(Y \mid \mathrm{do}(X_2))$ as a constraint regardless of the existence of the dashed edges because it is identifiable. However, $\mathrm{p}(Y \mid \mathrm{do}(X_1))$ is not identifiable if the edge between X_1 and Y exists. As a result, we can only use conditional expectations for X_1.

In both settings, we use the relative difference estimator defined in Garrido Mejia et al. (2022) as the parameter for the ROC curves of CMAXENT and i-CMAXENT (for details see Appendix C). For the ROC curve of the KCI test, we vary the threshold on the p-value of the test.

Joint interventional distributions from single interventions We are interested in finding multivariate interventional distributions when our available data comes from single-variable interventions. For this task, we use the DAG shown in Figure 2(*a*) and sample 200 random graph instantiations.

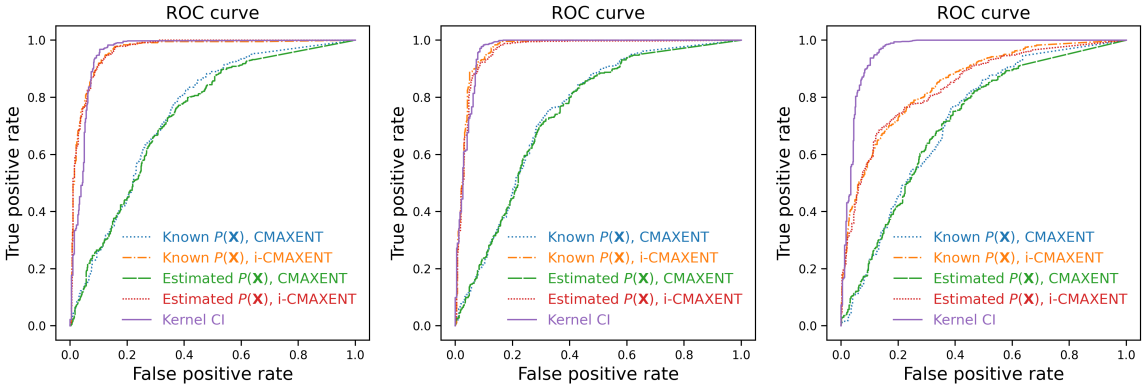
We perform five experiments to assess how varying the type and amount of constraints influences the estimation of the joint interventional density $\mathrm{p}(Y \mid \mathrm{do}(X_1, X_2))$. In all of these cases, we use 1,000 observations to compute the empirical averages: (i) Two potential causes X_i and X_j are chosen at random, and we provide i-CMAXENT with constraints on $\mathrm{p}(Y \mid X_i, X_j)$. Additionally, we provide constraints on single-variable interventional distributions for the rest of the potential causes; that is, $\mathrm{p}(Y \mid \mathrm{do}(X_k))$ for all $k \neq i, j$. (ii) We only provide constraints on $\mathrm{p}(Y \mid X_i, X_j)$. (iii) We only provide constraints on the single-variable interventional densities for all potential causes $\mathrm{p}(Y \mid \mathrm{do}(X_i))$ for all i. (iv) We only provide constraints on the single-variable conditional densities $\mathrm{p}(Y \mid X_i)$ for all potential causes. This scenario coincides with CMAXENT. (v) As an additional baseline, we finally estimate MAXENT without constraints. We then estimate the joint interventional density $\mathrm{p}(Y \mid \mathrm{do}(X_1, X_2))$ for each graph and plot the residuals against the true distribution, as shown in Figure 4.

7. Results

Parental discovery Figures 3(*a*) to 3(*c*) show the ROC curves corresponding to the graphs in Figures 2(*a*) to 2(*c*), respectively. Across all cases, i-CMAXENT consistently outperforms CMAXENT, and in graphs Figure 2(*a*) and Figure 2(*b*) achieves accuracy close to KCI despite using far less information. Even for graph Figure 2(*c*) where the potential causes interact with each other, we observe that the difference between i-CMAXENT and KCI remains modest.

Figures 3(*d*) to 3(*f*) show the performance of i-CMAXENT in scenarios where information about the interventional distributions is provided as constraints only for a fraction of the potential causes, while for the remaining only conditional distributions are used. We observe that the causes are recovered better as we increase the share of variables for which interventional data is provided.

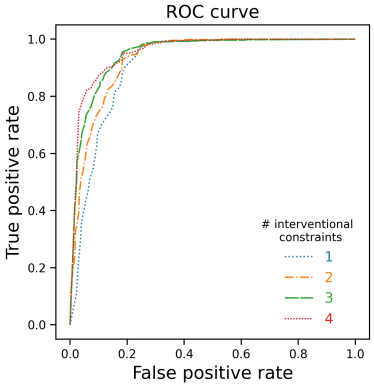
Joint interventional distributions from single variable interventions Figure 4 depicts the residuals between the estimated joint interventional distributions and the true joint interventional distributions, for interventions on X_1 and X_2. Any added constraint improves estimation, with i-CMAXENT reducing residuals to around 5% even when relying only on either single-variable conditional or interventional information.



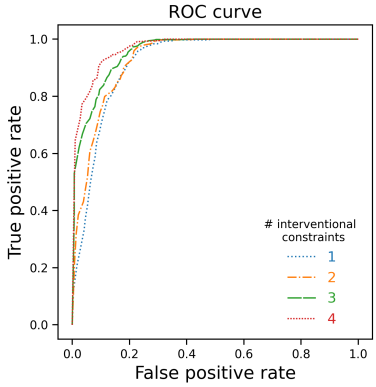
(*a*) ROC curve for graph (a)

(*b*) ROC curve for graph (b)

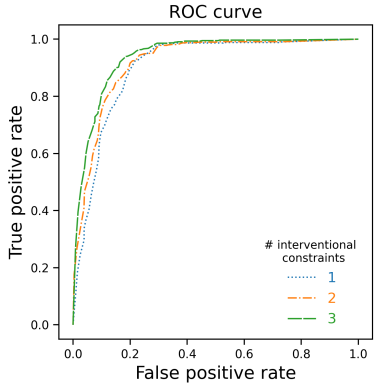
(*c*) ROC curve for graph (c)



(*d*) ROC curve for varying number of interventional means for graph (a)



(*e*) ROC curve for varying number of interventional means for graph (b)



(*f*) ROC curve for varying number of interventional means for graph (c)

Figure 3: Results for parental discovery. (Figure 3(*a*)), (Figure 3(*b*)), and (Figure 3(*c*)) show ROC curves for the identification of causal edges between the X_is and Y in setting 1. For i-CMAXENT , we use constraints on all five single-interventional densities $\mathrm{p}(Y \mid \mathrm{do}(X_i))$. For CMAXENT, we use constraints on the five single-conditional densities $\mathrm{p}(Y \mid X_i)$. The KCI test has access to observations from the joint $\mathrm{p}(\mathbf{X}, Y)$. For i-CMAXENT and CMAXENT, we consider two cases: (i) The joint observational density of the causes $\mathrm{p}(\mathbf{X})$ is known (blue and orange line). (ii) $\mathrm{p}(\mathbf{X})$ is estimated (green and red lines) from constraints on the five marginal densities $\mathrm{p}(X_i)$. Although our approach only uses single-variable interventional constraints as input, it achieves similar performance to the KCI test that uses the full generated dataset.

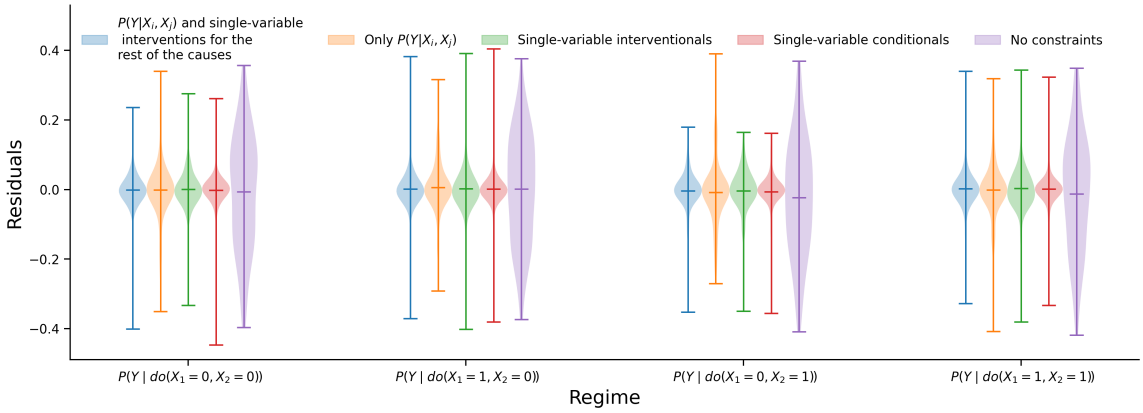


Figure 4: Residuals between true and estimated joint interventional distributions. The violin plots show the residuals between the true and the estimated joint interventional densities $\mathrm{p}(Y \mid \mathrm{do}(X_1, X_2))$ for five cases that differ in the constraints we use in the estimation. The constraints are: (i) the joint conditional $\mathrm{p}(Y \mid X_i, X_j)$ for a randomly chosen pair X_i, X_j, and single-variable interventionals $\mathrm{p}(Y \mid \mathrm{do}(X_k))$ for the rest of the variables (blue); (ii) only $\mathrm{p}(Y \mid X_i, X_j)$ (orange); (iii) single-variable interventionals $\mathrm{p}(Y \mid \mathrm{do}(X_k))$ for all causes (green); (iv) single-variable conditionals $\mathrm{p}(Y \mid X_k)$ for all causes (red); and (v) no constraints at all (purple).

Comparing residuals across settings, using only $\mathrm{p}(Y \mid X_i, X_j)$ yields the highest variance, while single-variable conditionals give slightly lower variance than the other cases. Moreover, the extrema (the horizontal marks at the end of each distribution) of the residuals have similar spreads, depending on the regime.

8. Discussion

Causal marginal problem with interventional distributions We proved in 8 that the causal marginal problem can be solved using interventional distributions as constraints within the Maximum Entropy framework, yielding solutions in the exponential family. This extends CMAXENT to i-CMAXENT and shows that interventional information from experimental data on subsets of variables can be combined without requiring joint observations. i-CMAXENT therefore provides a principled tool for applications where the joint effect of disjoint treatments must be estimated from limited interventional data.

This extension enables two key tasks: parental discovery and estimation of joint interventional distributions. In feature selection, i-CMAXENT achieves performance close to methods with full joint data (KCI, (Zhang et al., 2011)), while operating under the harder marginal setting. For joint interventional distribution estimation, i-CMAXENT with single-variable interventions performs comparably to using only single-variable conditionals. Its advantage is that it can directly incorporate experimental data when available, without distributional constraints (Saengkyongam and Silva, 2020) or functional form constraints (Kekić et al., 2025).

Limitations Our method requires knowing which variables were intervened upon to match expectations in the optimisation problem. This is inherent to the Maximum Entropy framework: constraints must specify the intervention. By contrast, methods like ICP only assume interventions

occurred on some subset, but they require access to full joint observational data, which we do not. Furthermore, identifiability remains an important question. It is clear that not all joint conditional distributions can be identified using the method, regardless of the empirical success we observe in Section 7. We conjecture that some potential functional form constraints, such as the Generalised Additive Model (GAM) (Hastie and Tibshirani, 1986) assumption in (Kekić et al., 2025) might be enough for identification with i-CMAXENT .

Acknowledgments

This work was partly supported by the Amazon Hub project PSMEWE A03B "Merging data sources".

References

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120, 2012.

Adam Berger, Stephen A Della Pietra, and Vincent J Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.

Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 116–125, 1999.

David Danks, Clark Glymour, and Robert Tillman. Integrating locally learned causal structures with overlapping variables. *Advances in Neural Information Processing Systems*, 21, 2008.

W Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4): 427–444, 1940.

Anish Dhir and Ciarán M Lee. Integrating overlapping datasets using bivariate causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3781–3790, 2020.

Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In Marina Meila and Xiaotong Shen, editors, *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. URL https://proceedings.mlr.press/v2/eaton07a.html.

Muhammad Qasim Elahi, Mahsa Ghasemi, and Murat Kocaoglu. Identification of average causal effects in confounded additive noise models. *arXiv preprint arXiv:2407.10014*, 2024.

Farzan Farnia and David Tse. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Sergio Hernan Garrido Mejia, Elke Kirschbaum, and Dominik Janzing. Obtaining causal information by merging datasets with maxent. In *International Conference on Artificial Intelligence and Statistics*, pages 581–603. PMLR, 2022.

Jaime Roquero Gimenez and Dominik Rothenhäusler. Causal aggregation: estimation and inference of causal effects by constraint-based data fusion. *Journal of Machine Learning Research*, 23(335): 1–60, 2022.

Luigi Gresele, Julius Von Kügelgen, Jonas Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. In *International Conference on Machine Learning*, pages 7793–7824. PMLR, 2022.

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433, 2004.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical science*, 1(3):297–310, 1986.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Reginawanti Hindersah, Agusthinus Marthin Kalay, and Abraham Talahaturuson. Rice yield grown in different fertilizer combination and planting methods: Case study in buru island, indonesia. *Open Agriculture*, 7(1):871–881, 2022.

Dominik Janzing. Causal versions of maximum entropy and principle of insufficient reason. *Journal of Causal Inference*, 9(1):285–301, 2021.

Dominik Janzing, Xiaohai Sun, and Bernhard Schölkopf. Distinguishing cause and effect via second order exponential models. *arXiv preprint arXiv:0910.5561*, 2009.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Olivier Jeunen, Ciarán Gilligan-Lee, Rishabh Mehrotra, and Mounia Lalmas. Disentangling causal effects from sets of interventions in the presence of unobserved confounders. *Advances in Neural Information Processing Systems*, 35:27850–27861, 2022.

Yonghan Jung, Iván Díaz, Jin Tian, and Elias Bareinboim. Estimating causal effects identifiable from a combination of observations and experiments. *Advances in Neural Information Processing Systems*, 36:46446–46490, 2023.

Armin Kekić, Sergio Hernan Garrido Mejia, and Bernhard Schölkopf. Learning joint interventional effects from single-variable interventions in additive models. *arXiv preprint arXiv:2506.04945*, 2025.

Hans G. Kellerer. Maßtheoretische marginalprobleme. *Mathematische Annalen*, 153(3):168–198, June 1964. doi: 10.1007/bf01360315. URL https://doi.org/10.1007/bf01360315.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sanghack Lee, Juan D Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *Uncertainty in artificial intelligence*, pages 389–398. PMLR, 2020.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *The Journal of Machine Learning Research*, 21(1):3919–4026, 2020.

Mateusz Olko, Mateusz Gajewski, Joanna Wojciechowska, Mikołaj Morzy, Piotr Sankowski, and Piotr Miłoś. Since faithfulness fails: The performance limits of neural causal discovery. In *International Conference on Machine Learning*, pages 47155–47175. PMLR, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Haonan Qiu, Shihong Yang, Zewei Jiang, Yi Xu, and Xiyun Jiao. Effect of irrigation and fertilizer management on rice yield and nitrogen loss: A meta-analysis. *Plants*, 11(13):1690, 2022.

Sorawit Saengkyongam and Ricardo Silva. Learning joint nonlinear effects from single-variable interventions in the presence of hidden confounders. In *Conference on Uncertainty in Artificial Intelligence*, pages 300–309. PMLR, 2020.

Numair Sani, Atalanti A Mastakouri, and Dominik Janzing. Bounding probabilities of causation through the causal marginal problem. *arXiv preprint arXiv:2304.02023*, 2023.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012.

Xu Shi, Ziyang Pan, and Wang Miao. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Causal inference by choosing graphs with most plausible markov kernels. In *Ninth International Symposium on Artificial Intelligence and Mathematics (AIMath 2006)*, pages 1–11, 2006.

Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521, 2001.

Jin Tian and Judea Pearl. *A general identification condition for causal effects*. eScholarship, University of California, 2002.

Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Causal effect identification from multiple incomplete data sources: A general search-based approach. *Journal of Statistical Software*, 99(5):1–40, 2021. doi: 10.18637/jss.v099.i05.

Robert Tillman and Peter Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 3–15. JMLR Workshop and Conference Proceedings, 2011.

Robert E Tillman. Structure learning with independent non-identically distributed data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1041–1048, 2009.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 16(1): 2147–2205, 2015.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Anicetus Wihardjaka, Elisabeth Srihayu Harsanti, and Asep Nugraha Ardiwinata. Effect of fertilizer management on potassium dynamics and yield of rainfed lowland rice in indonesia. *Chilean journal of agricultural research*, 82(1):33–43, 2022.

Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 632–639, 2002.

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.

K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

Appendix A. Proofs of the main results

Proposition 10 (Nonparametric identification through observational data combination) *Let* $\mathbf{X}, Y$ *be binary random variables and suppose* $\mathrm{p}(\mathbf{x}) > 0$*, for all* $\mathbf{x}$*. Then the joint interventional distribution* $\mathrm{p}(Y \mid \mathrm{do}(X_1 = x_1, X_2 = x_2))$ *is identifiable using* $\mathrm{p}(\mathbf{X})$ *and* $\mathrm{p}(Y \mid X_i)$ *for* $i = 1, 2$*.*

Proof We have

$$\begin{aligned}
\mathrm{p}(Y \mid X_1 = 0) =& \mathrm{p}(Y \mid X_1 = 0, X_2 = 0)\mathrm{p}(X_2 = 0 \mid X_1 = 0) \\
&+ \mathrm{p}(Y \mid X_1 = 0, X_2 = 1)\mathrm{p}(X_2 = 1 \mid X_1 = 0) && (6)\\
\mathrm{p}(Y \mid X_1 = 1) =& \mathrm{p}(Y \mid X_1 = 1, X_2 = 0)\mathrm{p}(X_2 = 0 \mid X_1 = 1) \\
&+ \mathrm{p}(Y \mid X_1 = 1, X_2 = 1)\mathrm{p}(X_2 = 1 \mid X_1 = 1) && (7)\\
\mathrm{p}(Y \mid X_2 = 0) =& \mathrm{p}(Y \mid X_1 = 0, X_2 = 0)\mathrm{p}(X_1 = 0 \mid X_2 = 0) \\
&+ \mathrm{p}(Y \mid X_1 = 1, X_2 = 0)\mathrm{p}(X_1 = 1 \mid X_2 = 0) && (8)\\
\mathrm{p}(Y \mid X_2 = 1) =& \mathrm{p}(Y \mid X_1 = 0, X_2 = 1)\mathrm{p}(X_1 = 0 \mid X_2 = 1) \\
&+ \mathrm{p}(Y \mid X_1 = 1, X_2 = 1)\mathrm{p}(X_1 = 1 \mid X_2 = 1). && (9)
\end{aligned}$$

Using Equations (6) to (9) we can find the following expressions:

$$\begin{aligned}
\mathrm{p}(Y \mid X_1 = 0, X_2 = 0) =& \frac{\mathrm{p}(Y \mid X_1 = 0) - \mathrm{p}(Y \mid X_1 = 0, X_2 = 1)\mathrm{p}(X_2 = 1 \mid X_1 = 0)}{\mathrm{p}(X_2 = 0 \mid X_1 = 0)} && (10)\\
\mathrm{p}(Y \mid X_1 = 1, X_2 = 1) =& \frac{\mathrm{p}(Y \mid X_1 = 1) - \mathrm{p}(Y \mid X_1 = 1, X_2 = 0)\mathrm{p}(X_2 = 0 \mid X_1 = 1)}{\mathrm{p}(X_2 = 1 \mid X_1 = 1)} && (11)\\
\mathrm{p}(Y \mid X_1 = 1, X_2 = 0) =& \frac{\mathrm{p}(Y \mid X_2 = 0) - \mathrm{p}(Y \mid X_1 = 0, X_2 = 0)\mathrm{p}(X_1 = 0 \mid X_2 = 0)}{\mathrm{p}(X_1 = 1 \mid X_2 = 0)} && (12)\\
\mathrm{p}(Y \mid X_1 = 0, X_2 = 1) =& \frac{\mathrm{p}(Y \mid X_2 = 1) - \mathrm{p}(Y \mid X_1 = 1, X_2 = 1)\mathrm{p}(X_1 = 1 \mid X_2 = 1)}{\mathrm{p}(X_1 = 0 \mid X_2 = 1)}, && (13)
\end{aligned}$$

with which we can replace recursively to find the following expression for $\mathrm{p}(Y \mid X_1 = 0, X_2 = 0)$:

$$\begin{aligned}
\mathrm{p}(Y \mid X_1 = 0, X_2 = 0) = \mathrm{p}(Y \mid X_1 = 1, X_2 = 1)&\mathrm{p}(X_2 = 0 \mid X_1 = 0)^{-1} \\
&[\mathrm{p}(Y \mid X_1 = 0) - \mathrm{p}(X_2 = 1 \mid X_1 = 0)\mathrm{p}(X_1 = 0 \mid X_2 = 1)^{-1} \\
&\quad [\mathrm{p}(Y \mid X_2 = 1) - \mathrm{p}(X_1 = 1 \mid X_2 = 1)\mathrm{p}(X_2 = 1 \mid X_1 = 1)^{-1} \\
&\quad\quad [\mathrm{p}(Y \mid X_1 = 1) - \mathrm{p}(X_1 = 0 \mid X_2 = 0)\mathrm{p}(X_1 = 1 \mid X_2 = 0)^{-1} \\
&\quad\quad\quad [\mathrm{p}(Y \mid X_2 = 0) - \mathrm{p}(Y \mid X_1 = 0, X_2 = 0)]]]], \qquad (14)
\end{aligned}$$

and after some algebra we can find an expression of $\mathrm{p}(Y \mid X_1 = 0, X_2 = 0)$ as a function of only bivariate distributions, which we can replace again in Equations (6) to (9) to find the other distributions of interest. ■

Theorem 8 (Exponential family of i-CMAXENT) *Using the Lagrange multiplier formalism, the solution of Equation* (1) *with the additional constraint from Equation* (4) *is given by the following*

exponential family:

$$\begin{aligned} \mathrm{p}_\lambda(y \mid \mathbf{x}) = \exp\Bigg(& \sum_{k=1}^{K} \lambda_k f_k(y, \mathbf{x}) + \sum_{j=1}^{J} \sum_{\mathbf{x}_{S_j}} \lambda_j^{\mathbf{x}_{S_j}} g_j(y, \mathbf{x}) \\ & + \sum_{l=1}^{L} \sum_{\mathbf{x}_{S_l^I}, \mathbf{x}_{S_l^C}} \lambda_l^{\mathbf{x}_{S_l^I}, \mathbf{x}_{S_l^C}} h_l(y, \mathbf{x}_{S_l^I}, \mathbf{x}_{S_l^C}) + \beta(\mathbf{x}) \Bigg), \end{aligned} \tag{5}$$

where $\beta(\mathbf{x})$ is the normalising constant, as in the conditional case.

Proof We start by setting up the Lagrangian, where we have one λ for each constraint in our optimisation problem.

$$\begin{aligned} \mathcal{L} = & -\sum_{y,\mathbf{x}} \mathrm{p}(y \mid \mathbf{x}) \mathrm{p}(\mathbf{x}) \log \mathrm{p}(y \mid \mathbf{x}) \\ & + \sum_{k=1}^{K} \lambda_k \Big(\sum_{y,\mathbf{x}} \mathrm{p}(y \mid \mathbf{x}) \mathrm{p}(\mathbf{x}) f_k(y, \mathbf{x}) - \tilde{f}_k \Big) \\ & + \sum_{j=1}^{J} \sum_{\mathbf{x}_{S_j}} \lambda_j^{\mathbf{x}_{S_j}} \Big(\sum_{y} \mathrm{p}(y \mid \mathbf{x}_{S_j}) g_j(y, \mathbf{x}_{S_j}) - \tilde{g}_j(\mathbf{x}_{S_j}) \Big) \\ & + \sum_{l=1}^{L} \sum_{\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}} \lambda_l^{\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}} \Big(\sum_{y} \mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C}) h_l(y, \mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}) - \tilde{h}_l(\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}) \Big) \\ & + \sum_{\mathbf{x}} \lambda^{\mathbf{x}} \Big(\sum_{y} \mathrm{p}(y \mid \mathbf{x}) - 1 \Big) \end{aligned} \tag{15}$$

The above Lagrangian contains an interventional distribution that depends on $\mathrm{p}(y \mid \mathbf{x})$. In order to optimise the Lagrangian we need to replace $\mathrm{p}(y \mid \mathrm{do}(x))$ as a function of $\mathrm{p}(y \mid x)$. In the following, we will use s_j and s_l^I to denote elements of the set S_j and S_l^I. In the following derivation, we will denote the complement of S_j in $\mathbf{X}$ as S_j'.

$$\mathrm{p}(y \mid \mathbf{x}_{S_j}) = \sum_{\mathbf{x}_{S_j'}} \mathrm{p}(y, \mathbf{x}_{S_j'} \mid \mathbf{x}_{S_j}) \tag{16}$$

$$= \sum_{\mathbf{x}_{S_j'}} \mathrm{p}(y \mid \mathbf{x}_{S_j'}, \mathbf{x}_{S_j}) \mathrm{p}(\mathbf{x}_{S_j'} \mid \mathbf{x}_{S_j}) \tag{17}$$

$$= \sum_{\mathbf{x}_{S_j'}} \mathrm{p}(y \mid \mathbf{x}) \mathrm{p}(\mathbf{x}_{S_j'} \mid \mathbf{x}_{S_j}). \tag{18}$$

Using this technique, we can express $\mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C})$ as a function of $\mathrm{p}(y \mid \mathbf{x})$, that is, $\mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C}) = \mathrm{p}(y \mid \mathbf{x}) \overline{\mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C})}$, where $\overline{\mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C})}$ is a combination of sums and products of known observable quantities.

Differentiating with respect to each $\mathrm{p}(Y = y \mid \mathbf{X} = \mathbf{x})$ and all the multipliers, we obtain

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathrm{p}(y \mid \mathbf{x})} =& -\mathrm{p}(\mathbf{x})[\log \mathrm{p}(y \mid \mathbf{x}) + 1] \\
&+ \sum_{k=1}^{K} \lambda_k \mathrm{p}(\mathbf{x}) f_k(y, \mathbf{x}) \\
&+ \sum_{j=1}^{J} \sum_{\mathbf{x}_{S_j}} \lambda_j^{\mathbf{x}_{S_j}} \left[\sum_{\mathbf{x}_{S_j^C}} \mathrm{p}(\mathbf{x}_{S_j^C} \mid \mathbf{x}_{S_j}) g_j(y, \mathbf{x}_{S_j}) \right] \\
&+ \sum_{l=1}^{L} \sum_{\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}} \lambda_l^{\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}} \overline{\mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C})} h_l(y, \mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}) \\
&+ \lambda^{\mathbf{x}}
\end{aligned} \tag{19}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = \sum_{y, \mathbf{x}} \mathrm{p}(y \mid \mathbf{x}) \mathrm{p}(\mathbf{x}) f_k(y, \mathbf{x}) - \tilde{f}_k, \, \forall \, k \tag{20}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_j^{\mathbf{x}_{s_j}}} = \sum_{y} \mathrm{p}(y \mid \mathbf{x}_{s_j}) g_j(y, \mathbf{x}_{s_j}) - \tilde{g}_j(\mathbf{x}_{s_j}), \, \forall \, s_j \in S_j, \, \forall \, j \tag{21}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_l^{\mathbf{x}_{s_l^C}, \mathbf{x}_{s_l^I}}} = \sum_{y} \mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C}) h_l(y, \mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}) - \tilde{h}_l(\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}), \, \forall \, s_l^C \in S_l^C, \forall \, s_l^I \in S_l^I, \, \forall \, l \tag{22}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda^{\mathbf{x}}} = \sum_{y} \mathrm{p}(y \mid \mathbf{x}) - 1, \, \forall \, \mathbf{x} \tag{23}$$

Notice that the derivative with respect to the Lagrangian in Equation (19) is 0 for those functions f_k, g_j and h_l for which $\mathbf{x} \notin S_j, S_k, S_l^C, S_l^I$.

We then find the solution to $\mathrm{p}(y \mid \mathbf{x})$ when the above equations are equal to 0. From Equation (19) we get

$$\begin{aligned}
\mathrm{p}(y \mid \mathbf{x}) = \exp \Bigg[& \sum_{k=1}^{K} \lambda_k f_k(y, \mathbf{x}) + \frac{1}{\mathrm{p}(\mathbf{x})} \sum_{j=1}^{J} \sum_{\mathbf{x}_{S_j}} \lambda_j^{\mathbf{x}_{S_j}} g_j(y, \mathbf{x}_{S_j}) \sum_{\mathbf{x}_{S_j^c}} \mathrm{p}(\mathbf{x}_{S_j^c} \mid \mathbf{x}_{S_j}) \\
& + \frac{1}{\mathrm{p}(\mathbf{x})} \sum_{l=1}^{L} \sum_{\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}} \lambda_l^{\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}} \overline{\mathrm{p}(y \mid \mathrm{do}(\mathbf{x}_{S_l^I}), \mathbf{x}_{S_l^C})} h_l(y, \mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}) + \frac{1}{\mathrm{p}(\mathbf{x})} \lambda^{\mathbf{x}} \Bigg].
\end{aligned} \tag{24}$$

Notice that this equation is well-defined as long as $\mathrm{p}(\mathbf{x}) > 0$ for all $\mathbf{x}$. Because the elements inside the exponential depend on $\mathbf{x}$, we can rename λ with $\tilde{\lambda}$. In addition, we gather the constants into the

normalizing constant, which depends on $\mathbf{x}$, giving us

$$\mathrm{p}(y \mid \mathbf{x}) = \exp\left[\sum_{k=1}^{K} \lambda_k f_k(y, \mathbf{x}) + \sum_{j=1}^{J} \sum_{\mathbf{x}_{S_j}} \tilde{\lambda}_j^{\mathbf{x}_{S_j}} g_j(y, \mathbf{x}_{S_j}) + \sum_{l=1}^{L} \sum_{\mathbf{x}_{S_l^C}} \sum_{\mathbf{x}_{S_l^I}} \tilde{\lambda}_l^{\mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}} h_l(y, \mathbf{x}_{S_l^C}, \mathbf{x}_{S_l^I}) + \beta(\mathbf{x})\right] \tag{25}$$

as required. ■

Proposition 11 (Identifiability and adjustment set of variables with only incoming arrows) *Let $\mathbf{X}$ be a set of candidate causal parents of Y, which can be confounded. Assume we know the density $\mathrm{p}(\mathbf{X})$. If the only child of $X_j \in \mathbf{X}$ is potentially, but not necessarily, Y, then $\mathrm{p}(Y \mid \mathrm{do}(X_j))$ is identifiable and a valid adjustment set for the atomic intervention is $\mathbf{X} \setminus X_j$. That is, the rest of the potential causes.*

Proof Because of the assumed generative process, the causal sufficiency assumption, and the fact that Y is the only potential child of X_j, there are no "bidirectional" arrows connected to any child of X_j in the sense of Tian and Pearl (2002). As a result, the conditions of Theorem 2 in Tian and Pearl (2002) apply and $\mathrm{p}(Y \mid \mathrm{do}(X_j))$ is identifiable using observational quantities.

We have just proved the identifiability of the interventional distribution. Now we would like to prove that the set $\mathbf{X}' = \mathbf{X} \setminus X_j$ is a valid adjustment set. This is true for the following reasons. First, the assumption of the three levels of the generative process, which has as a consequence that there is no collider (or descendants of a collider) between Y and the elements in $\mathbf{X}'$. Second, Assumption (1) which states there is no unobserved confounder between $\mathbf{X}$ and Y, thus $\mathbf{X}'$ blocks any potential backdoor path between X_j and Y through any confounder. ■

Appendix B. Computation of Maxent through norm minimisation

As shown in Theorem 8, the interventional maximum entropy solution is equivalent to maximum likelihood where the dual variables are the parameters of the exponential family. We will now show how the maximum likelihood problem can be expressed as the minimisation between the difference between the given empirical averages and the expectations of the functions given the exponential family distribution. The way we maximise Equation (25) for all our data, with respect to the Lagrange multipliers (the λ_i), is by making equal to 0 the derivative of the log-likelihood with respect to the parameters:

$$\frac{\partial \log \mathrm{p}(Y = y \mid \mathbf{X} = \mathbf{x})}{\partial \lambda_i} = \frac{1}{N} \sum_{n=1}^{N} f_k(y^n, \mathbf{x}^n) - \frac{f_k(y, \mathbf{x}) \sum_k \lambda_k f_k(y, \mathbf{x})}{\exp \sum_y \sum_k \lambda_k f_k(y, \mathbf{x})} = 0. \tag{26}$$

Because we have empirical averages, the previous equation (for the whole data) becomes

$$\tilde{f}_k(y, \mathbf{x}) - \mathbb{E}_{\mathrm{p}_\lambda}[f_k(y, \mathbf{x})] = 0, \tag{27}$$

which we can compute by using any method that minimises the difference between the observed empirical average and the entailed expectation using the exponential family distribution. That is, we can compute the solution to the dual problem of the exponential family with

$$\lambda = \arg\min_{\lambda} \| \tilde{f}_k(y, \mathbf{x}) - \mathbb{E}_{\mathrm{p}_\lambda}[f_k(y, \mathbf{x})] \| \tag{28}$$

In the synthetic experiments Section (6), there were situations where $\mathrm{p}(\mathbf{x}) = 0$. We fixed this by adding a machine epsilon to all possible combinations and renormalizing.

Appendix C. Relative difference estimator

The relative difference estimator introduced in (Garrido Mejia et al., 2022) is an estimator of how close two parameters are to each other so that an analyst can decide whether there exist conditional independence. The relative difference estimator can take values between 0 and 1. However, there is no probabilistic analysis of the estimator, so one cannot consider the value of the relative difference estimator as a well-calibrated probability of the difference between multipliers (or difference between a multiplier and 0). The estimator is defined as

$$\theta_i = \frac{\left|\lambda_i^1 - \lambda_i^2\right|}{\max\{|\lambda_i^1|, |\lambda_i^2|, \left|\lambda_i^1 - \lambda_i^2\right|, 1\}} \in [0, 1] \ , \tag{29}$$

where λ_i^1, λ_i^2 are the two Lagrange multipliers for the constraints associated with X_i.

Appendix D. Details of data generation processes

The data generation process consists of two steps. For each of the three graph structures we first sample the parameters of a generative process. For this we sample for each variable $Z \in \{\mathbf{U}, \mathbf{X}, Y\}$ a value for $\mathrm{p}(Z{=}1 \mid PA_Z)$ by sampling from a uniform distribution between 0.1 and 0.9. We do this for each combination of values of the parents PA_Z of Z. For $\mathbf{X}$ and $\mathbf{U}$ he parents are fixed by the respective graph structure, but for Y we randomise the parents with probability 0.5 for including any particular X_i as a parent of Y. Next, we sample observations using this generative process. That is, we do ancestral sampling where we sample each variable Z using a Bernoulli distribution with the before chosen probability $\mathrm{p}(Z = 1 \mid PA_Z)$. To obtain the data from the interventional distributions, we simply set the variable to the particular intervened value and then proceed in the generative process.

Appendix E. Optimisation and convergence

To find the Lagrange multipliers, using the norm minimisation procedure explained in Appendix B, we use the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm implemented in the JAX Python library (Bradbury et al., 2018). We consider an estimation as converged if the norm (the sum of the squared of the residuals between the empirical averages given as constraints and the expectations entailed by the exponential family distribution) are less than 0.01, or if the optimiser terminates successfully. In some cases, convergence in this sense is not achieved on the first application of the optimisation algorithm. When this happens, we simply use the optimisation algorithm again using the previously found Lagrange multipliers as the initial values for the optimisation. We repeat

this process until convergence in the above sense is achieved. We found that in most cases one extra optimisation is enough to achieve convergence, there were four cases where we had to run the optimiser more than two times: twice for three, once for four, and once for five.

Appendix F. Other faithfulness assumptions

The main goal of faithfulness assumptions is to make conclusions about the causal graph from statistical conclusions. The most important use of this is to do causal discovery from finite sample data. We begin with one of the simplest notions of faithfulness (Spirtes et al., 2000):

Assumption 12 (Faithfulness) *A distribution* P *is faithful to a DAG* $\mathcal{G}$ *if no conditional independence relations other than the ones entailed by the Markov property are present.*

Since Theorem 12 does not depend parametrically on how we measure conditional independence, Zhang and Spirtes (2002) modify this notion to include a λ parameter:

Assumption 13 (λ-strong Faithfulness) *Given* $\lambda \in (0, 1)$*, a multivariate Gaussian distribution* P *is said to be* λ*-strong-faithful to a DAG* $\mathcal{G} = (V, E)$ *if for any* $i, j \in V$ *and any* $S \subseteq V \setminus \{i, j\}$ *such that* j *is d-separated from* i *given* S *if and only if* $|\mathrm{Corr}(X_i, X_j \mid X_S)| \leq \lambda$*.*

As Uhler et al. (2013) prove, and in contrast with the usual Faithfulness assumption, the set of distributions for which Theorem 13 hold does not have Lebesgue measure zero. There is an important implication: it has been proven that the usual faithfulness assumption holds "almost surely" as the set of distributions for which the assumption does not hold has measure zero. However, if this were to change, that is, if this set of distributions did not have Lebesgue measure zero, then there are uncountably many such distributions, and the possibility of being in such scenario is non-zero. Of course if faithfulness does not hold, the causal graphs built from conditional independence tests might be wrong. More recently Olko et al. (2025) studied strong faithfulness in the context of neural causal discovery and non linear models, confirming the results found by Uhler et al. (2013) in the linear case and the PC algorithm.

Other notions of faithfulness have been introduced in Zhang and Spirtes (2008):

Assumption 14 (Restricted λ-strong Faithfulness) *Given* $\lambda \in (0, 1)$*, a multivariate Gaussian distribution* P *is said to be restricted* λ*-strong-faithful to a DAG* $\mathcal{G} = (V, E)$ *if both of the following hold:*

(i) $\min\{|\mathrm{Corr}(X_i, X_j \mid X_S)| : (i, j) \in E,\ S \subseteq V \setminus \{i, j\},$ *such that* $|S| \leq deg(\mathcal{G})\} > \lambda$*, where* $deg(G)$ *denotes the sum of the indegree and the outdegree of nodes in* $\mathcal{G}$*. This condition is also known as Adjacency faithfulness (Zhang and Spirtes, 2008), according to Uhler et al. (2013).*

(ii) $\min\{|\mathrm{Corr}(X_i, X_j \mid X_S)| : (i, j, S) \in N_{\mathcal{G}}\} > \lambda$*, where* $N_{\mathcal{G}}$ *is the set of triples* (i, j, S) *such that* i, j *are not adjacent but there exists* $k \in V$ *making* (i, j, k) *an unshielded triple (a triple of the form* $i \to k \leftarrow j$*) and* i, j *are not d-separated given* S*. This condition is also known as Orientation faithfulness (Zhang and Spirtes, 2008), according to Uhler et al. (2013).*

Assumption 15 (Triangle Faithfulness) *Suppose the true causal DAG of a set of nodes* V *is* $\mathcal{G}$*. Let* X, Y, Z *be any three variables that form a triangle in* $\mathcal{G}$ *(i.e., each pair of vertices is adjacent):*

(i) *If Y is a noncollider on the path $X - Y - Z$, then X and Z are not independent conditional on any subset of $V \setminus \{X, Z\}$ that does not contain Y.*

(ii) *If Y is a collider on the path, (i.e., $X \to Y \leftarrow Z$), then X and Z are not independent conditional on any subset of $V \setminus \{X, Z\}$ that contains Y.*

Out of these definitions of faithfulness only Theorems 12 and 13 have been studied to a certain degree, given the difficulty of analysing these assumptions. We believe that the assumption of faithful f-expectations possess the same genericity as Theorem 12.