Obtaining Causal Information by Merging Datasets with MAXENT

Sergio Hernan Garrido Mejia[*†]
Danmarks Tekniske Universitet
Lyngby, Denmark
sermej@amazon.com

Elke Kirschbaum[*]
Amazon Research
Tübingen, Germany
elkeki@amazon.com

Dominik Janzing
Amazon Research
Tübingen, Germany
janzind@amazon.com

Abstract

The investigation of the question "which treatment has a causal effect on a target variable?" is of particular relevance in a large number of scientific disciplines. This challenging task becomes even more difficult if not all treatment variables were or even cannot be observed jointly with the target variable. In this paper, we discuss how causal knowledge can be obtained without having observed all variables jointly, but by merging the statistical information from different datasets. We show how the maximum entropy principle can be used to identify edges among random variables when assuming causal sufficiency and an extended version of faithfulness, and when only subsets of the variables have been observed jointly.

1 INTRODUCTION

The scientific community is rich in observational and experimental studies that consider a tremendous amount of problems from an even more significant number of perspectives. All these studies have collected data containing valuable information to investigate the research question at hand. At the same time, it is often impossible to use the collected data to answer slightly different or more general questions, as the required information cannot be extracted from the already existing datasets.

*Both authors contributed equally

†This work was done while the author worked at Amazon Research

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Consider, for instance, a case in which we want to investigate the influence of the place of residence on the probability to become depressed, and we are given four different studies: (1) showing the depression rates for different regions; (2) capturing depression rate with respect to (w.r.t.) age; (3) providing information about the depression rate w.r.t. sex; and (4) showing the distribution of age and sex across different regions. We want to know whether there is a direct causal link between the place of residence and the depression rate or only an indirect link through age and/or sex. In this paper, we address the question of how we can obtain this causal information without performing a new study in which we observe all factors (age, sex, place of residence, and depression rate) at the same time, but only by merging the already collected datasets.

Since the problem of inferring the joint distribution from a set of marginals is heavily underdetermined (Kellerer, 1964), we use the maximum entropy (MAXENT) principle to infer the joint distribution that maximises the joint entropy subject to the observed marginals. This has the advantage that the MAXENT distribution contains some information about the existence of causal arrows that also hold for the true joint distribution regardless of how much the MAXENT distribution deviates from the latter.

As usual, our causal conclusions require debatable assumptions that link statistical properties of distributions from passive observations to causality. Therefore we use assumptions common in causal discovery (Spirtes et al., 1993; Pearl, 2000). Additionally, we define and intuitively justify the notion of *faithful f-expectations*, which is analogous to faithfulness in the sense of postulating genericity of parameters. This allows us to draw the following conclusions, which are the main contributions of this paper:

- The presence or absence of direct causal links can be identified only from the Lagrange multipliers of the MAXENT solution if the causal order is known (see sec. 3, corollary 1).

- For a causal graph G with N nodes for which the given constraints define all bivariate distributions uniquely, the graph constructed from the MAXENT distribution by connecting two nodes if and only if there is a non-zero Lagrange multiplier corresponding to some bivariate function of the two variables, is a supergraph of the moral graph of G (see sec. 3, theorem 2).
- Merging datasets with MAXENT improves the predictive power compared to using the observed marginal distributions (see sec. 3, theorem 3).

The remainder of this paper is structured as follows: We start by presenting the notation and assumptions used throughout this paper. Then, in sec. 2 we introduce the MAXENT principle. In sec. 3 we discuss how we can obtain causal information by merging datasets. In sec. 4 we put our work into the context of the related literature. Finally, in sec. 5 we evaluate the identification of causal edges from MAXENT on simulated and real-world datasets.

Notation Let $\mathbf{X} = \{X_1, \dots, X_N\}$ be a set of discrete random variables. Although the results in this article hold also for continuous variables with strictly positive densities ($p(\mathbf{x}) > 0$) and finite differentiable entropy, for notational convenience we consider discrete random variables with values $\mathbf{x} \in \mathcal{X}$. Further let $X_i, X_j \in \mathbf{X}$ be two variables whose causal relationship we want to investigate. We denote with $\mathbf{Z} = \mathbf{X} \setminus \{X_i, X_j\}$ the complement of $\{X_i, X_j\}$ in $\mathbf{X}$, where (by slightly overloading notation) bold variables represent sets and vectors of variables at the same time. We consider the set of functions $f = \{f_k\}$ with $f_k : \mathcal{X}_{S_k} \to \mathbb{R}$ for some $k \in \mathbb{N}$ and $\mathbf{X}_{S_k} \subseteq \mathbf{X}$. The empirical means of f for a finite sample from the joint distribution $P(\mathbf{X})$ are collected in the set $\tilde{f} = \{\tilde{f}_k\}$, and the set of true expectations we denote with $\mathbb{E}_p[f] = \{\sum_{\mathbf{x}} p(\mathbf{x}) f_k(\mathbf{x}_{S_k})\}$. Further, we denote with P the 'true' joint distribution of the variables under consideration and with $\hat{P}$ the approximate MAXENT solution satisfying the constraints imposed by the expectations of f, as described in sec. 2.

Assumptions If not stated differently, we make the following assumptions throughout this paper: The set of variables $\mathbf{X}$ is causally sufficient, that is, there is no hidden common cause $U \notin \mathbf{X}$ that is causing more than one variable in $\mathbf{X}$ (and the causing paths go only through nodes that are not in $\mathbf{X}$)(Peters et al., 2017). Furthermore, their joint distribution $P(\mathbf{X})$ satisfies the causal Markov condition and faithfulness w.r.t. a directed acyclic graph (DAG) G (see appendix A). We have L datasets, where each contains observations for only a subset of the variables, and at least one dataset contains observations for the set $\{X_i, X_j\}$. The observations are drawn from the same joint distribution $P(\mathbf{X})$.[1] Further, the set of functions f is linearly independent.

2 MAXIMUM ENTROPY

The maximum entropy (MAXENT) principle (Jaynes, 1957) is a framework to find a 'good guess' for the distribution of a system if only a set of expectations for some feature functions f is given. The MAXENT distribution is the solution to the optimisation problem

$$\max_p H_p(\mathbf{X})$$
$$\text{s.t.:} \quad \mathbb{E}_p[f] = \tilde{f} \,, \quad \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \,, \tag{1}$$

for the Shannon entropy $H_p(\mathbf{X}) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$. Often the statistical moments $f_k(x) = x^k$ for $k \in \mathbb{N}$ are used. Note that many quantities of interest are simple expressions from expectations of appropriate functions, e.g. the covariance of two random variables is $\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]$.

Approximate MAXENT The empirical means of the functions f gained from a finite sample will never be *exactly* identical to the true expectations. This implies that not even the true distribution necessarily satisfies the constraints imposed on the MAXENT distribution. This can lead to large (or even diverging) values of the parameters which overfit statistical fluctuations. To account for this, the expectations only need to be *close* to the given empirical means. This leads to the formulation of *approximate* MAXENT (Dudik et al., 2004; Altun and Smola, 2006), where the constraints in the optimisation problem in eq. (1) are replaced by approximate constraints, resulting in

$$\min_p -H_p(\mathbf{X})$$
$$\text{s.t.:} \, \|\mathbb{E}_p[f] - \tilde{f}\|_{\mathcal{B}} \le \varepsilon \,, \quad \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \,, \tag{2}$$

with $\varepsilon \ge 0$. This type of convex optimisation problems have been studied for infinite dimensional Banach spaces $\mathcal{X}$ and $\mathcal{B}$ by Altun and Smola (2006), where it was shown that eq. (2) is equivalent to

$$\max_\phi \left\langle \phi, \tilde{f} \right\rangle - \log \sum_{\mathbf{x}} \exp\left[\langle \phi, f \rangle\right] - \varepsilon \|\phi\|_{\mathcal{B}^*} \,, \tag{3}$$

with $\mathcal{B}^*$ being the dual to $\mathcal{B}$. In contrast to standard MAXENT, whose well-known dual is maximum likelihood estimation, in approximate MAXENT, the parameters ϕ are regularised depending on the choice

[1] In appendix B we sketch the case where each dataset is from a different joint distribution by introducing an additional variable for the background conditions.

of the norm in eq. (2). For instance, $\mathcal{B} = \ell_\infty$ results in a Laplace regularisation $\varepsilon\|\phi\|_1$. Appropriate choices for ε are proportional to $\mathcal{O}(1/\sqrt{M})$, where M is the sample size, although in practice ε is usually chosen using cross-validation techniques (Dudik et al., 2004; Altun and Smola, 2006).

Throughout this paper, we will use approximate MAXENT and assume that $\mathcal{X}$ and $\mathcal{B}$ are finite-dimensional. We consider the ℓ_∞ norm, which results in the element-wise constraints

$$|\mathbb{E}_p\left[f_k\right] - \tilde{f}_k| \leq \varepsilon_k \quad \forall k \tag{4}$$

for $\varepsilon_k \geq 0$. In this case, the MAXENT optimisation problem can be solved analytically using the Lagrangian formalism of constrained optimisation and the solution reads

$$\hat{p}(\mathbf{x}) = \exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k}) - \alpha\right] , \tag{5}$$

where $\lambda = \{\lambda_k\}$ are the Lagrange multipliers and $\alpha = \log \sum_\mathbf{x} \exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k})\right]$ is the partition function ensuring that $\hat{p}$ is correctly normalised. The optimal Lagrange multipliers can be found via

$$\min_\lambda - \sum_k \lambda_k \tilde{f}_k + \log \sum_\mathbf{x} \exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k})\right] + \sum_k \varepsilon_k |\lambda_k| \, . \tag{6}$$

Conditional MAXENT In cases where the marginal distribution of a subset of the variables is already known, the MAXENT approach can be natively extended to a *conditional* MAXENT. For instance, consider the variable $X_j \in \mathbf{X}$, and assume we are given the distribution $P(\bar{\mathbf{X}})$ for $\bar{\mathbf{X}} = \mathbf{X} \backslash \{X_j\}$. Additionally, we are given some expectations involving the variables $\bar{\mathbf{X}}$ and X_j. In this case, we obtain the MAXENT solution for the joint distribution of $\mathbf{X}$ by maximising the conditional entropy

$$H_p(X_j \mid \bar{\mathbf{X}}) = -\sum_\mathbf{x} p(x_j \mid \bar{\mathbf{x}}) p(\bar{\mathbf{x}}) \log p(x_j \mid \bar{\mathbf{x}}) \ , \tag{7}$$

subject to the constraints in eq. (4) imposed by the expectations of the functions f as before, but now the set of variables $\mathbf{X}_{S_k}$ the function f_k acts upon always contains the variable X_j, so $\mathbf{X}_{S_k} = \bar{\mathbf{X}}_{S_k} \cup \{X_j\}$ for $\bar{\mathbf{X}}_{S_k} \subseteq \bar{\mathbf{X}}$. In this case, the solution in the Lagrangian formalism reads

$$\hat{p}(x_j \mid \bar{\mathbf{x}}) = \exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k}) - \beta(\bar{\mathbf{x}})\right] \ , \tag{8}$$

where λ are the respective Lagrange multipliers for which optimal values can be found analogously to eq. (6), and $\beta(\bar{\mathbf{x}}) = \log \sum_{x_j} \exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k})\right]$ ensures that the marginal constraint $\hat{p}(\bar{\mathbf{x}}) = p(\bar{\mathbf{x}})$ is satisfied. The joint MAXENT distribution is then given by $\hat{p}(\mathbf{x}) = \hat{p}(x_j \mid \bar{\mathbf{x}}) p(\bar{\mathbf{x}})$.

Using conditional means When we consider a scenario as described in the introduction, in which we want to merge the information from different studies or research papers, we might only be provided with *conditional means*, like the average depression rate given that the age is in a specific range. In this case, the given constraints would be

$$|\mathbb{E}_p\left[f_k \mid \bar{\mathbf{x}}_{S_k} = \bar{\mathbf{x}}^\nu_{S_k}\right] - \tilde{f}^\nu_k| \leq \hat{\varepsilon}^\nu_k \quad \forall k, \nu \ , \tag{9}$$

for $\nu = 1, \ldots, \mathcal{V}_k$ and $\bar{\mathbf{x}}^1_{S_k}, \ldots, \bar{\mathbf{x}}^{\mathcal{V}_k}_{S_k}$ being the possible sets of values the set of discrete random variables $\bar{\mathbf{X}}_{S_k}$ can attain. Then eq. (9) replaces the constraints in eq. (4) and the conditional MAXENT solution reads

$$\hat{p}(x_j \mid \bar{\mathbf{x}}) = \exp\left[\sum_{k,\nu} \hat{\lambda}^\nu_k f_k(\mathbf{x}_{S_k}) \delta_{\bar{\mathbf{x}}_{S_k}, \bar{\mathbf{x}}^\nu_{S_k}} - \hat{\beta}(\bar{\mathbf{x}})\right] \tag{10}$$

with the Lagrange multipliers $\hat{\lambda} = \left\{\hat{\lambda}^\nu_k\right\}$ and $\hat{\beta}(\bar{\mathbf{x}}) = \log \sum_{x_j} \exp\left[\sum_{k,\nu} \hat{\lambda}^\nu_k f_k(\mathbf{x}_{S_k}) \delta_{\bar{\mathbf{x}}_{S_k}, \bar{\mathbf{x}}^\nu_{S_k}}\right]$ and

$$\delta_{\bar{\mathbf{x}}_{S_k}, \bar{\mathbf{x}}^\nu_{S_k}} = \begin{cases} 1 & \text{if } \bar{\mathbf{x}}_{S_k} = \bar{\mathbf{x}}^\nu_{S_k} \\ 0 & \text{otherwise} \, . \end{cases} \tag{11}$$

3 OBTAINING CAUSAL INFORMATION BY MERGING DATASETS WITH MAXENT

In this section, we consider the analysis of the causal relationship between variables if not all variables have been observed jointly. All proofs of the following propositions can be found in appendix C.

First, we show how to detect the presence or absence of direct causal links in a DAG G from the Lagrange multipliers of the MAXENT distribution. We start by showing that if X_i and X_j are CI given all other variables w.r.t. the true distribution, then this is also the case w.r.t. the MAXENT distribution and reflects in the respective Lagrange multipliers being zero.

Lemma 1 (CI results in Lagrange multipliers being zero)**.** *Let P be a distribution and let $\hat{P}$ be the MAXENT distribution satisfying the constraints imposed by the expectations of the functions f which are sufficient to uniquely describe the marginal distributions*

$P(X_i, \mathbf{Z}), P(X_j, \mathbf{Z})$, *and* $P(X_i, X_j)$. *Then it holds:*

$$\begin{aligned} & X_i \perp\!\!\!\perp X_j \mid \mathbf{Z} \;\; [P] \\ \Rightarrow \;\; & X_i \perp\!\!\!\perp X_j \mid \mathbf{Z} \;\; [\hat{P}] \\ \Rightarrow \;\; & \lambda_k = 0 \quad \forall k \;\; \textit{with} \;\; \mathbf{X}_{S_k} = \{X_i, X_j\} \; . \end{aligned} \quad (12)$$

Under the stated assumptions, it directly follows from lemma 1 that if two variables are CI given all other variables, and hence not directly linked in the causal DAG G, then the respective Lagrange multipliers are zero. This, however, is not enough to draw conclusions from the Lagrange multipliers about the absence or presence of causal links. For this, we first need to show that the presence of a direct link results in a non-zero Lagrange multiplier. But to do this, we first need to postulate a property that we call *faithful f-expectations*. This property is analogous to faithfulness in postulating the genericity of parameters. For the following definition, we denote with λ_f^P and λ_f^Q the set of Lagrange multipliers of the MAXENT distribution satisfying the expectation constraints in eq. (4) entailed by the functions f w.r.t. the distributions P and Q, respectively.

Definition 1 (Faithful f-Expectations)**.** *A distribution P is said to have faithful f-expectations relative to a DAG G, if $\lambda_{f_k}^P \neq 0$ for all $f_k \in f$ where it exists a distribution Q that is Markov relative to G and for which it is $\lambda_{f_k}^Q \neq 0$.*

We rephrase this definition in the language of information geometry to show that this is just a genericity assumption like usual faithfulness, and fig. 1 illustrates the intuition behind it. By elementary results of information geometry (Amari and Nagaoka, 1993), the MAXENT distribution $\hat{P}$ can also be considered a projection of the distribution P onto the exponential manifold E_f, which is defined by the span of all functions f, containing distributions of the form $\exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k}) - \alpha\right]$ (visualised by the blue plane in fig. 1). If a Lagrange multiplier λ_k is zero, then $\hat{P}$ lies within the submanifold $E_{f\setminus\{f_k\}} \subset E_f$ which is defined through the span of all functions f without f_k (illustrated by the red, dashed line in fig. 1). Then faithful f-expectations state that the projection of P onto E_f will generically not lie in $E_{f\setminus\{f_k\}}$ unless the DAG G only allows for distributions whose projections onto E_f also lie in $E_{f\setminus\{f_k\}}$.

Further justification of faithful f-expectations via some probabilistic arguments would be a research project in its own right. After all, even the discussion on usual faithfulness is ongoing: The 'measure zero argument' by Meek (1995) is criticised in Lemeire and Janzing (2012), and it is argued that natural conditional distributions tend to be more structured. In Uhler et al. (2013) it is shown that distributions are

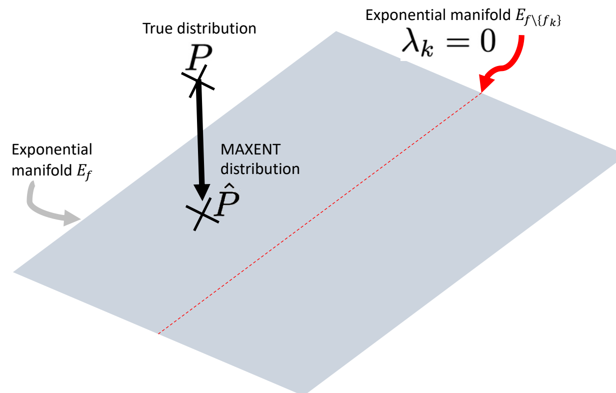


Figure 1: Intuitive explanation of the idea behind faithful f-expectations: the MAXENT distribution is a projection of the distribution P onto the exponential manifold E_f, defined by the span of the functions f. It is very unlikely that this projection falls into the submanifold $E_{f\setminus\{f_k\}}$ where $\lambda_k = 0$ just by chance.

not unlikely to be *close to* being unfaithful. Despite these concerns, faithfulness still proved to be helpful.

Postulating faithful f-expectations allows us to link the causal structure to the Lagrange multipliers:

Lemma 2 (Causally linked variables have non-zero Lagrange multipliers)**.** *Let P have faithful f-expectations relative to a causal DAG G. Then it is $\lambda_k^P \neq 0$ for any bivariate function f_k whose variables are connected in G.*

Now we have all we need to connect the structure of the causal DAG and the Lagrange multipliers.

Theorem 1 (Causal structure from Lagrange multipliers)**.** *Let P be a distribution with faithful f-expectations w.r.t. a causal DAG G, and let $\hat{P}$ be the MAXENT solution satisfying the constraints imposed by the expectations of the functions f which are sufficient to uniquely describe the marginal distributions $P(X_i, \mathbf{Z}), P(X_j, \mathbf{Z})$, and $P(X_i, X_j)$. Then the following two statements hold:*

1. *If $\mathbf{Z}$ is d-separating X_i and X_j in G, then all Lagrange multipliers λ_k are zero for all k with $\mathbf{X}_{S_k} = \{X_i, X_j\}$.*
2. *If $\lambda_k = 0$ for all k with $\mathbf{X}_{S_k} = \{X_i, X_j\}$, then there is no direct link between X_i and X_j in the DAG G.*

For the special case where we have some prior knowledge about the causal order, e.g. if we know that X_j can be causally influenced by X_i or $\mathbf{Z}$, but not the other way around, we can directly identify the absence or presence of a direct causal link between X_i and X_j:

Corollary 1 (Identification of causal links when causal order is known)**.** *Let P be a distribution with faithful f-expectations w.r.t. a causal DAG G, and let $\hat{P}$ be the MAXENT solution satisfying the constraints imposed by the expectations of the functions f which are sufficient to uniquely describe the marginal distributions $P(X_i, \mathbf{Z}), P(X_j, \mathbf{Z})$, and $P(X_i, X_j)$. If it is excluded that X_j can causally influence X_i and $\mathbf{Z}$, i.e. the DAG G cannot contain edges $X_j \to X_i$ or $X_j \to \mathbf{Z}$, then it holds*

$$X_i \text{ is not directly linked to } X_j \Leftrightarrow \lambda_k = 0 \quad \forall k \text{ with } \mathbf{X}_{S_k} = \{X_i, X_j\} \ . \tag{13}$$

This also holds for conditional MAXENT, and if conditional means are used (see eq. (9)) it holds

$$X_i \text{ is not directly linked to } X_j \Leftrightarrow \hat{\lambda}_k^{\nu} = \hat{\lambda}_k^{\nu'} \quad \forall \nu, \nu', k \text{ with } \mathbf{X}_{S_k} = \{X_i, X_j\} \ . \tag{14}$$

Note that when conditional means are used, the Lagrange multipliers need not be zero to indicate missing links, but need to be constant for all conditions. We use this result in our experiments in sec. 5, where we estimate conditional MAXENT in the causal order, which is called *causal MAXENT*, as proposed and justified by Janzing (2021).

The reader may wonder about more general statements like the question 'What information can be obtained about a DAG with N nodes if only bivariate distributions are available?'. For this scenario, we have at least a necessary condition for causal links. For this recall that for any DAG G, the corresponding *moral graph* G^m is defined as the undirected graph having edges if and only if the nodes are directly connected in G or have a common child (Lauritzen, 1996).

Theorem 2 (Graph constructed from MAXENT with only bivariate constraints is a supergraph of the moral graph)**.** *Let f be a basis for the space of univariate and bivariate functions, i.e. the set of f-expectations determine all bivariate distributions uniquely. Let P be a joint distribution that has faithful f-expectations w.r.t. the DAG G. Let G^b be the undirected graph constructed from the MAXENT distribution by connecting X_i and X_j if and only if there is a non-zero Lagrange multiplier corresponding to some bivariate function of X_i and X_j. Then G^b is a supergraph of G^m, the moral graph of G.*

Theorem 2 provides at least a candidate list for potential edges from bivariate information alone, which tells us where additional observations are needed to identify edges. The edges are candidates for being in the Markov blanket, which thus limits the number of variables that need to be considered for a prediction model.

Note that inferring causal relations via bivariate information is not uncommon: after all, many implementations of the PC algorithm (Spirtes et al., 2000; Kalisch and Bühlman, 2007; Kalisch and Bühlmann, 2008; Harris and Drton, 2013; Cui et al., 2016; Tsagris, 2019) use partial correlations instead of CIs. For real-valued variables, one can interpret this in the spirit of this paper since it infers CIs to hold whenever they are true for the multivariate Gaussian matching the observed first and second moments (i.e. the unique MAXENT distribution satisfying these constraints). These heuristics avoid the complex problem (Shah and Peters, 2020) of non-parametric CI testing. In addition to the results above, our approach also generalises the partial correlation heuristics to more general functions f_k, including multivariate and higher-order statistics.

Note also that we do not propose a *general purpose* conditional independence test because we do not fully understand what sort of conditional dependence it detects. We have concluded that it 'generically' (i.e. subject to faithful f-expectations) has power against conditional dependencies generated by a DAG. Without a DAG (whose distributions can be easily parameterised), we do not see a clear notion of genericity on which a similar statement could be based.

For most applications, estimating the joint distribution of many variables is not an end in itself. Instead, one will often be interested in particular properties of the joint distributions for specific reasons. In these cases, MAXENT is already helpful if it resembles the statistical properties of interest. So far, we have shown this for some conditional independence. We will now sketch how entropy maximisation can be used for pooling predictions made from different datasets.

Theorem 3 (Predictive power of MAXENT)**.** *Let $X_j, X_i, \mathbf{Z}$ be binary variables, with $\mathbf{Z}$ possibly high dimensional. Furthermore, let $\hat{P}(X_j \mid X_i, \mathbf{Z})$ be the MAXENT solution that maximises the conditional entropy of X_j given X_i and $\mathbf{Z}$, subject to the moment constraints given by the observed pairwise distributions $P(X_j, X_i)$, $P(X_j, \mathbf{Z})$, and $P(X_i, \mathbf{Z})$. Then $\hat{P}$ is a better predictor of X_j than any of the individual bivariate probabilities, as measured by the likelihood of any point where all variables are observed, i.e. a point from $P(X_j, X_i, \mathbf{Z})$.*

This show that merging datasets with MAXENT also improves the predictive power compared to using the observed marginal distributions.

4 RELATED WORK

In the context of missing data (Rubin, 1976; Bareinboim and Pearl, 2011; Mohan and Pearl, 2021), many methods have been developed to investigate CI, how the joint distribution of the variables factorises, and how to predict the result of interventions (Pearl, 2000; Spirtes et al., 2000; Chickering, 2003; Tsamardinos et al., 2006). However, most of these approaches assume they are given one dataset in which some values are missing at random. More recently, the even more challenging task of inferring causal relationships from multiple datasets (Tillman, 2009; Ramsey et al., 2010; Triantafillou et al., 2010; Eberhardt et al., 2010; Claassen and Heskes, 2010; Tillman and Spirtes, 2011; Hyttinen et al., 2013; Tillman and Eberhardt, 2014). These approaches and our method have in common that they assume that the underlying causal structures are similar across the different datasets. Triantafillou et al. (2010); Tillman and Spirtes (2011), for instance, assume that a single causal mechanism generates the data and that the dependencies and independencies are captured by a maximal ancestral graph (MAG) and the m-separation criterion (Richardson et al., 2002). Various methods have also been proposed to discover the causal graph from multiple datasets containing measurements in different environments. Some combine statistics or constraints from the different datasets to construct a single causal graph (Claassen and Heskes, 2010; Tillman and Spirtes, 2011; Hyttinen et al., 2013, 2014; Triantafillou and Tsamardinos, 2015; Rothenhäusler et al., 2015; Forré and Mooij, 2018), while others directly combine the data from the different datasets and construct a causal graph from the pooled data (Cooper, 1997; Hauser and Bühlmann, 2012; Cooper and Yoo, 1999; Mooij and Heskes, 2013; Peters et al., 2016; Oates et al., 2016; Zhang et al., 2017; Mooij et al., 2020). In Mooij et al. (2020), for instance, the union of causal graphs in each dataset (or context) is found by jointly modelling the context variables and the observed variables. The main difference between these approaches and ours is that they all rely on statistical information that reveals CIs in each dataset individually. Hence they can only be applied if at least three variables have been observed jointly, while our approach can also be used if only pairwise observations are available.

In Gresele et al. (2022), the structural marginal question was asked: Can marginal causal models over subsets of variables with known causal graph be consistently merged? They proved that certain SCM can be falsified using only interventional and observetional data and a known graph structure. Their work differs from ours in that we are interested in interventional quantities, while they focus on counterfactual ones. As a result, the questions that can be answered with our framework are different.

In appendix E we comment on less related – but still interesting – literature on gaining statistical information from causal knowledge and other entropy-based approaches to extract and exploit causal information.

5 EXPERIMENTS

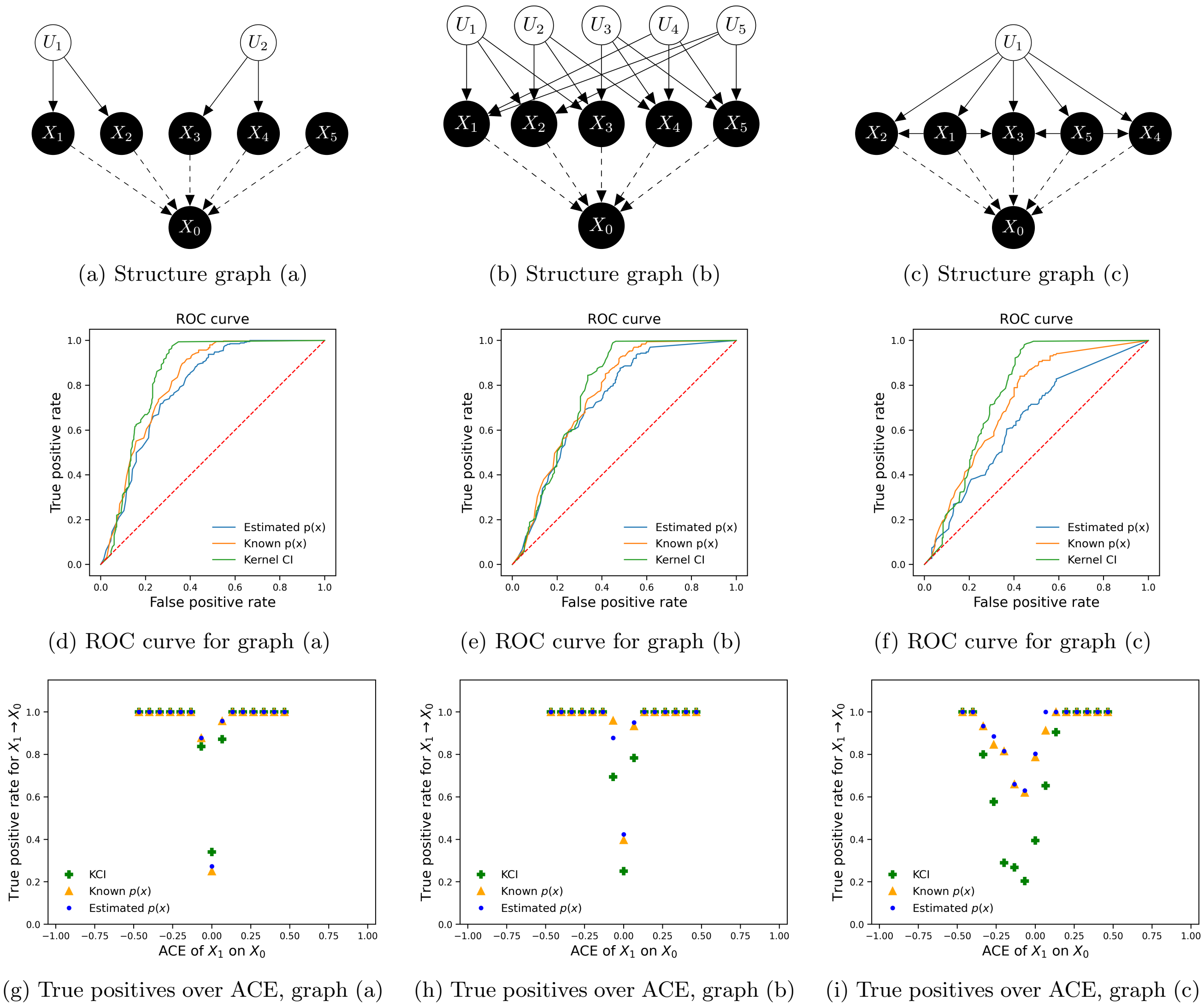
In this section, we apply the theoretical results from sec. 3 on different synthetically generated and real-world datasets. For this, we implemented the MAXENT estimation in Python (see appendix G).

Synthetic data We consider five binary variables $X_1, \dots, X_5$, which are potential causes of a sixth binary variable X_0, and we want to infer which variables X_i have a direct causal link to X_0. The ground truth DAGs for our experiments are shown in figs. 2a to 2c, and the SCMs we used for the data generation can be found in appendix H.

For the first set of experiments, we kept the structure of the confounders U_j with the potential causes fixed (solid lines) and randomised the existence of mechanisms between the potential parents and the effect variable X_0 (dashed lines). We generated 100 datasets for each graph structure by randomly picking the existing mechanisms and the parameters used in the SCM. We sample 1000 data points according to the respective SCM for each dataset. Then we artificially split these observations into five datasets that we want to merge and that always only contain bivariate information about X_0 and one of the potential causes X_i. We do this by empirically estimating the conditional means $\mathbb{E}_p\left[x_0 \mid x_i = 0\right]$ and $\mathbb{E}_p\left[x_0 \mid x_i = 1\right]$ from the samples for all $i = 1, \dots, 5$. We use these conditional means as constraints for the MAXENT optimisation problem as shown in eq. (9). We assume that X_0 cannot have a causal influence on any X_i. Therefore, we can use the results in corollary 1 to identify whether X_i is directly causally linked to X_0 or not. To decide whether the Lagrange multipliers associated with a potential cause X_i are constant – and hence X_i is not directly linked to X_0 – we use a relative difference estimator

$$\theta_i = \frac{\left|\lambda_i^1 - \lambda_i^2\right|}{\max\{|\lambda_i^1|, |\lambda_i^2|, |\lambda_i^1 - \lambda_i^2|, 1\}} \in [0, 1] \ , \qquad (15)$$

where λ_i^1, λ_i^2 are the two Lagrange multipliers for the constraints associated with X_i. We consider the Lagrange multipliers constant if θ_i is smaller than a threshold $t \in [0, 1]$. We vary the threshold t linearly between zero and one. We count the number of correctly and falsely identified edges in the 100 datasets for each threshold value. The results are summarised in the receiver operating characteristic (ROC) curves in figs. 2d to 2f. We consider two scenarios: one in which we assume that we know the marginal distri-



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

(d) ROC curve for graph (a) (e) ROC curve for graph (b) (f) ROC curve for graph (c)

(g) True positives over ACE, graph (a) (h) True positives over ACE, graph (b) (i) True positives over ACE, graph (c)

Figure 2: We show the structure of the graphs we consider in the synthetic experiments in (a), (b), and (c). In (d), (e), and (f) we show the ROC curves for the identification of missing edges. We generated 100 datasets for each graph where we varied the used SCMs and the absence and presence of the edges shown as dashed lines, whereas links represented by solid lines are always present. In (g), (h), and (i) we show how the ability to detect an edge depends on the strength of the causal effect. Here we generated another 500 datasets for each graph in which the link between X_1 and X_0 was always present, but the ACE of X_1 on X_0 varied. Although our MAXENT-based approach only uses conditional means as input, it achieves similar performance as the KCI-test that uses the full generated dataset.

bution $P(X_1, \ldots, X_5)$ for the potential causes (called 'known $p(x)$', orange line), and the second where we first infer this distribution also using MAXENT (called 'estimated $p(x)$', blue line). Further, we compare our results with a kernel-based conditional independence test (KCI-test) (Zhang et al., 2011; Strobl et al., 2019) (green line). For the KCI-test, we directly use the 1000 data points generated from the joint distribution. To generate the ROC curve, we vary the α-level of the test for the null hypothesis that X_0 is CI of X_i given all other potential causes and count the number of correct/false rejections/acceptances.

In the second set of experiments, we investigate how much our approach's ability to identify edges depends on the strength of the causal effect. We generated 500 additional datasets for each graph as described before, but this time always included a causal link from X_1 to X_0 and only varied the strength of this connection. We fixed the threshold for the identification of an edge to a randomly picked value ($t = \alpha = 0.15$). Figures 2g to 2i show how in this case the true positive rate for the identification of the link depends on the ACE of X_1 on X_0.

The results in figs. 2d to 2i show that for all graph structures our method achieves similar performance as the KCI-test. This is impressive, as our method only uses the conditional means of X_0 on only one of the potential causes. In contrast, the KCI-test uses all samples generated from the joint data distribution. That means that, although our method uses much less information than the KCI-test and even merges these little pieces of information from different datasets, our method still achieves similar performance as the KCI-test.

In addition, we want to show that the MAXENT solution can not only provide information about the causal structure but even about the strength of a causal effect. For this, we derive bounds for the ACE based only on the marginal distributions in appendix D. In fig. 3 we see that the ACE estimated based on the MAXENT distribution is always very close to the true ACE, and even in the cases where they do not precisely coincide, they are both clearly within the bounds derived based on the marginal distributions.

Real data We performed an experiment using real-world data from Gapminder (2021), a website that compiles country-level data of social, economic, and environmental nature. We chose three variables for our experiment: CO2 tonnes emission per capita (Boden et al., 2017); inflation-adjusted Gross Domestic Product (GDP) per capita (World Bank, 2019); and Human Development Index (HDI) (United Nations Development Programme, Human Development Reports, 2019). We use data from 2017 for all variables and standardise it before estimation. We consider CO2 emissions the target variable for which the other variables are potential causes. We use the unconditional mean and variance of CO2 emissions and the pairwise covariance between CO2 emissions and each of the two other variables as constraints.

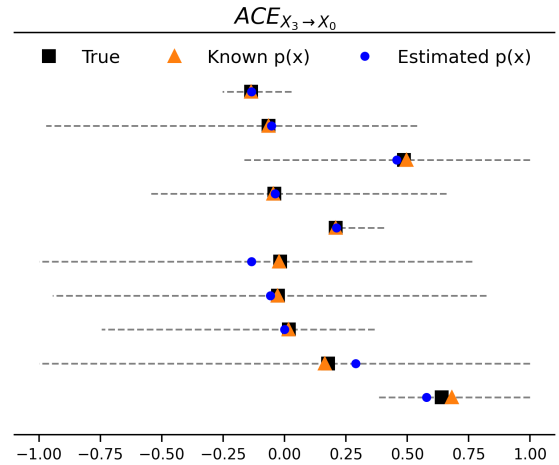


Figure 3: ACE of X_3 on X_0 for ten randomly picked examples of a variation of the graph in fig. 2c. The ACE estimated from the MAXENT solution with a known marginal distribution of the causes (orange triangle) is always close to the true ACE (black square). But even when the distribution of the causes is also inferred using MAXENT (blue dot), the estimated ACE is close to the true one and always within the bounds (grey lines) estimated from the marginal distributions.

According to the Lagrange multipliers shown in table 1, and corollary 1, we conclude that CO2 is directly linked to HDI but not to GDP. We run the same KCI-test as in the synthetic experiments above to investigate this conclusion. Table 1 also shows the obtained p-values for the null hypothesis that CO2 emission is CI of each variable conditioned on the other variable. The results of the KCI-test agree with our conclusion. The result of the KCI-test, nonetheless, does not necessarily reflect the ground truth. However, GDP only has an indirect causal influence on CO2 emissions through HDI matches our intuition. We would expect that a change in GDP not directly affects the CO2 emissions but influences the HDI – and potentially multiple other factors that we do not consider in this experiment – that then affects the CO2 emissions. In appendix F we discuss more such experiments in which we also include fertility and life expectancy as potential causes. In all of the considered cases where we used two potential causes, the conclu-

Table 1: Found Lagrange multipliers λ_i for the MAXENT solution and the p-values for the KCI-test. We indicate where the multipliers / p-values indicate the presence of a direct edge connecting X_i and CO2 emissions / that the two are not CI given the other variable.

variable X_i	λ_i	edge	p-value	no CI
GDP	-0.29	✗	0.19	✗
HDI	3.26	✓	0.02	✓

sion drawn from the Lagrange multipliers agreed with the conclusion drawn from the KCI-test. Only when we include more variables, the KCI-test indicates CI of HDI and CO2 emissions given the other variables, while our method still finds a direct link between them. However, this finding of the KCI-test is also inconsistent with the other CI statements of the KCI-test for smaller conditioning sets. Moreover, consider what is available to both methods: The KCI-test requires a sample of the joint distribution, while our method relies solely on bivariate covariances, which might not even be enough to describe the joint distribution fully.

Finally, we consider the example from the introduction, in which we want to investigate the depression rate conditioned on age, sex and place of residence. We are given the conditional means for the depression rate given age, sex, and the federal state of Germany, in addition to the joint distribution of age, sex, and state (Gesundheit Statistik, 2021a,b). Using this information, we can find the MAXENT solution for the joint distribution of all four variables (depression rate (D), age (A), sex (S), and place of residence (P)). The found Lagrange multipliers are shown in appendix F. For none of the three potential causes, the multipliers are constant. Hence, we assume that all three factors (age, sex, and place of residence) have a direct causal link to the depression rate.[2] Nevertheless, we can use the result from theorem 3, stating that the joint MAXENT solution is a better predictor than any of the given marginal distributions, to investigate questions like ‘What is the probability for a 30-year-old woman living in a certain federal state to become depressed?’. When we, for instance, consider the federal states Baden-Wuerttemberg (BW) and Berlin (BE), then the result of the MAXENT solution is

$$\begin{aligned} p(D \mid S = \text{female}, A = 30, P = \text{BW}) &= \;\; 9.5\%, \\ p(D \mid S = \text{female}, A = 30, P = \text{BE}) &= 11.2\%, \end{aligned}$$

while from the marginal distributions, we get

$$\begin{aligned} p(D \mid S = \text{female}) \;\; &= 9.7\%, p(D \mid P = \text{BW}) = 7.7\%, \\ p(D \mid A = 30 - 44) &= 7.5\%, p(D \mid P = \text{BE}) \;\; = 9.3\%. \end{aligned}$$

It seems surprising that the probability increases when conditioning on all three factors. However, since the depression rate for ‘female’ is higher than for ‘male’ (which is only 8.6%), it makes sense that the depression probability slightly increases when additionally conditioning on the sex being female. Further note that none of the above necessarily reflects the true probability. The MAXENT solution only provides a ‘better guess’ for the depression rate given all three factors than each of the marginal distributions.

6 CONCLUSION

We have derived how the MAXENT principle can identify links in causal graphs and thus obtain information about the causal structure by merging the statistical information in different datasets.

There are several directions of extension of this work. On the practical side, we believe that developing efficient ways to compute the expectations of the inferred distribution is vital. In our experiments, we merged between two and five datasets and used up to 22 constraints. In order to scale the problem to more variables and constraints, the main bottleneck is the estimation of the partition function (α and $\beta(\bar{\mathbf{x}})$ in sec. 2). Some efficient ways to compute this are developed in Wainwright and Jordan (2008), however, the properties with respect to causality remain unknown.

Another direction for future work would be to study the statistical properties of the estimated parameters. In other words, to develop a statistical of the null hypothesis of a multiplier being zero (in the case of unconditional moments, or equal to others in the case of conditional moments).

In addition to the causal insights we get from this work, we would like to highlight two ways in which this work can positively impact society. First, by using only information from expectations, a characteristic that makes MAXENT a flexible approach, we move one step forward to avoid identifying individuals in adversarial attacks. Second, by using information from different sources, we can avoid being unable to answer causal questions, or worse, giving wrong causal answers because of a lack of jointly observed data.

Acknowledgements

We thank Steffen Lauritzen for helpful remarks on undirected graphical models.

[2]Note, it is still possible that these factors only have an *indirect* influence on the depression rate via other factors that we do not consider here. Investigating *all* potential causes for the depression rate would be a research project of its right and is out of the scope of this work.

References

Altun, Y. and Smola, A. (2006). Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 139–153, Pittsburgh, PA.

Amari, S. and Nagaoka, H. (1993). *Methods of Information Geometry*. Oxford University Press.

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

Bareinboim, E. and Pearl, J. (2011). Controlling selection bias in causal inference. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence, August 7 –11*, pages 247–254.

Boden, T. A., Andres, R. J., and Marland, G. (2017). Global, regional, and national fossil-fuel co2 emissions (1751 - 2014) (v. 2017).

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1).

Chickering, D. M. (2003). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Claassen, T. and Heskes, T. (2010). Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems 22*.

Compton, S., Kocaoglu, M., Greenewald, K., and Katz, D. (2021). Entropic causal inference: Identifiability and finite sample results. *arXiv preprint arXiv:2101.03501*.

Cooper, G. F. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224.

Cooper, G. F. and Yoo, C. (1999). Causal discovery from a mixture of experimental and observational data. *arXiv preprint arXiv:1301.6686*.

Cui, R., Groot, P., and Heskes, T. (2016). Copula pc algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 377–392. Springer.

Dudik, M., Phillips, S. J., and Schapire, R. E. (2004). Performance guarantees for regularized maximum entropy density estimation. In *International Conference on Computational Learning Theory*, pages 472–486. Springer.

Eberhardt, F., Hoyer, P., and Scheines, R. (2010). Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 185–192. JMLR Workshop and Conference Proceedings.

Forré, P. and Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*.

Galles, D. and Pearl, J. (1995). Testing identifiability of causal effects. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 185–195, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gapminder (2021). Gapminder. `www.gapminder.org/data/`.

Gesundheit Statistik, G. d. B. (2021a). 12-month-prevalence of self-reported diagnosed depression (percentage of the respondents). classification: years, region, age, sex, level of education. [Online; accessed 6-October-2021].

Gesundheit Statistik, G. d. B. (2021b). European core health indicators (echi): Population by sex and age. [Online; accessed 6-October-2021].

Gresele, L., von Kügelgen, J., Kübler, J. M., Kirschbaum, E., Schölkopf, B., and Janzing, D. (2022). Causal inference through the structural causal marginal problem. *arXiv preprint arXiv:2202.01300*.

Grosse-Wentrup, M., Janzing, D., Siegel, M., and Schölkopf, B. (2016). Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage*, 125:825–833.

Harris, N. and Drton, M. (2013). Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11).

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., and Jarvisalo, M. (2013). Discovering cyclic causal models with latent variables: A general sat-based procedure. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Janzing, D. (2018). Merging joint distributions via causal model classes with low VC dimension. preprint `arXiv:1804.03206v2`.

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

Janzing, D., Sun, X., and Schölkopf, B. (2009). Distinguishing cause and effect via second order exponential models. `http://arxiv.org/abs/0910.5561`.

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

Jung, Y., Tian, J., and Bareinboim, E. (2021). Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Kalisch, M. and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3).

Kalisch, M. and Bühlmann, P. (2008). Robustification of the pc-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, 17(4):773–789.

Kallus, N., Puli, A. M., and Shalit, U. (2018). Removing hidden confounding by experimental grounding. *arXiv preprint arXiv:1810.11646*.

Kellerer, H. (1964). Maßtheoretische Marginalprobleme. *Math. Ann.*, 153:168–198. in German.

Kocaoglu, M., Dimakis, A., Vishwanath, S., and Hassibi, B. (2017). Entropic causal inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Lauritzen, S. (1996). *Graphical Models.* Clarendon Press, Oxford, New York, Oxford Statistical Science Series edition.

Lemeire, J. and Janzing, D. (2012). Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249.

Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. Proceedings of *11th Uncertainty in Artificial Intelligence (UAI), Montreal, Canada,* Morgan Kaufmann, pages 411–418.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 0(ja):1–42.

Mooij, J. and Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. *arXiv preprint arXiv:1309.6849*.

Mooij, J. M., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21:1–108.

Oates, C. J., Smith, J. Q., and Mukherjee, S. (2016). Estimating causal structure using conditional dag models. *The Journal of Machine Learning Research*, 17(1):1880–1903.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012.

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

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., and Glymour, C. (2010). Six problems for causal inference from fmri. *neuroimage*, 49(2):1545–1558.

Richardson, T., Spirtes, P., et al. (2002). Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030.

Rothenhäusler, D., Heinze, C., Peters, J., and Meinshausen, N. (2015). Backshift: Learning causal cyclic graphs from unknown shift interventions. *arXiv preprint arXiv:1506.02494*.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2013). Semi-supervised learning in causal and anticausal settings. In Schölkopf, B., Luo, Z., and Vovk, V., editors, *Empirical Inference*, Festschrift in Honor of Vladimir Vapnik, pages 129–141. Springer.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538.

Sharma, A., Hofman, J. M., Watts, D. J., et al. (2018). Split-door criterion: Identification of causal effects through auxiliary outcomes. *Annals of Applied Statistics*, 12(4):2699–2733.

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, 11(5).

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York, NY.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search*. MIT, Cambridge, MA, 2nd edition.

Strobl, E. V., Visweswaran, S., Zhang, K., et al. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):1–24.

Sun, X., Janzing, D., and Schölkopf, B. (2006). Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL.

Tillman, R. and Spirtes, P. (2011). 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.

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

Tillman, R. E. and Eberhardt, F. (2014). Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, 41(1):41–64.

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

Triantafillou, S., Tsamardinos, I., and Tollis, I. (2010). Learning causal structure from overlapping variable sets. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 860–867. JMLR Workshop and Conference Proceedings.

Tsagris, M. (2019). Bayesian network learning with the pc algorithm: An improved and correct variation. *Applied Artificial Intelligence*, 33(2):101–123.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

Tsamardinos, I., Triantafillou, S., and Lagani, V. (2012). Towards integrative causal analysis of heterogeneous data sets and studies. *J. Mach. Learn. Res.*, 13(1):1097–1157.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463.

United Nations Development Programme, Human Development Reports (2019). Human development index (hdi).

Vorob'ev, N. (1962). Consistent families of measures and their extensions. *Theory Probab. Appl*, 7(2):147–163.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Wang, Y. and Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596.

World Bank (2019). Gdp per capita (constant 2010 us$).

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access.

Zhang, K., Peters, P., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. `http://uai.sis.pitt.edu/papers/11/p804-zhang.pdf`.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. (2010). Modeling interaction via the principle of maximum causal entropy. In *ICML*.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. (2013). The principle of maximum causal entropy for estimating interacting processes. *IEEE Transactions on Information Theory*, 59(4):1966–1980.

Supplementary Materials

A GRAPHICAL CAUSAL MODELS

In graphical causal models, the causal relations among random variables are described via a directed acyclic graph (DAG) G, where the expression $X_i \to X_j$ means that X_i influences X_j 'directly' in the sense that intervening on X_i changes the distribution of X_j if all other nodes are adjusted to fixed values. If no hidden variable $U \notin \mathbf{X}$ exists that causes more than one variable in $\mathbf{X}$, then the set $\mathbf{X}$ is said to be *causally sufficient* (Spirtes, 2010).

The crucial postulate that links statistical observations with causal semantics is the *causal Markov condition* (Spirtes et al., 1993; Pearl, 2000), stating that each node X_n is conditionally independent (CI) of its non-descendants given its parents $PA(X_n)$ w.r.t. the graph G. Then, the probability mass function of the joint probability distribution factorises into

$$p(x_1, \dots x_N) = \prod_{n=1}^{N} p(x_n \mid pa(x_n)) \ , \tag{16}$$

where $p(x_n \mid pa(x_n))$ are often called *Markov kernels* (Lauritzen, 1996). This entails further CIs described by the graphical criterion of *d-separation* (Pearl, 2000). The Markov condition is a necessary condition for a DAG being causal. To test the corresponding CIs is a first sanity check for a causal hypothesis.

More assumptions are required to infer causal structure from observational data. One common assumption is *faithfulness*: a distribution is faithful to a DAG G if a CI in the data implies d-separation in the graph. Inferring the entire causal DAG from passive observations, or *causal graph discovery*, is, nevertheless, an ambitious task (Spirtes et al., 1993; Peters et al., 2017). We, therefore, focus on the weaker task of inferring the presence or absence of certain causal links.

B DATASETS COMING FROM DIFFERENT JOINT DISTRIBUTIONS

Our approach implicitly assumes that all datasets are taken from the same joint distribution. This assumption deserves justification. Suppose, for instance, we are interested in statistical relations between variables $X_1, \dots, X_N$ describing different health conditions of human subjects. Assume we are given L datasets containing different subsets of variables (e.g. bivariate statistics), but the datasets are from different countries. Accordingly, we should not assume a common joint distribution $X_1, \dots, X_N$. Instead, we may introduce an additional variable C, and a dataset from country $C = c$ containing variables X_i, X_j then provides only information about $E[f(X_i, X_j)|C = c]$. We would then infer a joint distribution of $C, X_1, \dots, X_N$ via MAXENT, given the conditional expectations.

C PROOFS

Here we repeat the theorems, corollaries and lemmas from the main text and provide the complete proofs for all of them.

Lemma 1 (CI results in Lagrange multipliers being zero)**.** *Let P be a distribution and let $\hat{P}$ be the MAXENT distribution satisfying the constraints imposed by the expectations of the functions f which are sufficient to uniquely describe the marginal distributions $P(X_i, \mathbf{Z}), P(X_j, \mathbf{Z})$, and $P(X_i, X_j)$. Then it holds:*

$$\begin{aligned} & X_i \perp\!\!\!\perp X_j \mid \mathbf{Z} \;\; [P] \\ \Rightarrow \quad & X_i \perp\!\!\!\perp X_j \mid \mathbf{Z} \;\; [\hat{P}] \\ \Rightarrow \quad & \lambda_k = 0 \quad \forall k \;\; \textit{with} \;\; \mathbf{X}_{S_k} = \{X_i, X_j\} \; . \end{aligned} \tag{12}$$

Proof. We first show that CI w.r.t. P results in CI w.r.t. the MAXENT distribution. Let Q be a distribution satisfying the following two conditions:

(a) $Q(X_i, X_j) = P(X_i, X_j)$, $Q(X_i, \mathbf{Z}) = P(X_i, \mathbf{Z})$, and $Q(X_j, \mathbf{Z}) = P(X_j, \mathbf{Z})$, and

(b) $X_i \perp\!\!\!\perp X_j \mid \mathbf{Z} \;\; [Q]$.

We know that such a distribution satisfying (a) and (b) exists, as this is the case for at least P itself. Now assume that the MAXENT distribution $\hat{P}$ satisfies condition (a) but not condition (b). Then the entropy of $\hat{P}$ is

$$\begin{aligned} H_{\hat{p}}(\mathbf{X}) &= H_{\hat{p}}(X_i \mid X_j, \mathbf{Z}) + H_{\hat{p}}(X_j, \mathbf{Z}) \overset{(\not b)}{<} H_{\hat{p}}(X_i \mid \mathbf{Z}) + H_{\hat{p}}(X_j, \mathbf{Z}) \\ &\overset{(a)}{=} H_q(X_i \mid \mathbf{Z}) + H_q(X_j, \mathbf{Z}) \overset{(b)}{=} H_q(X_i \mid X_j, \mathbf{Z}) + H_q(X_j, \mathbf{Z}) = H_q(\mathbf{X}) \; . \end{aligned}$$

This violates the assumption that $\hat{P}$ maximises the entropy. Hence, the distribution satisfying the marginal constraints in (a) that maximises the entropy must satisfy the CI in (b).

Next, we show that CI w.r.t. the MAXENT distribution results in the respective Lagrange multipliers being zero. By applying Bayes' rule to the MAXENT distribution in eq. (5) it can be seen that

$$\begin{aligned} \hat{p}(x_i \mid x_j, \mathbf{z}) &= \frac{p(\mathbf{x})}{\sum_{x_i} p(\mathbf{x})} = \frac{\exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k}) + \alpha\right]}{\sum_{x_i} \exp\left[\sum_k \lambda_k f_k(\mathbf{x}_{S_k}) + \alpha\right]} \\ &= \frac{\exp\left[\sum\limits_{\substack{k \text{ with} \\ \mathbf{x}_{S_k} = \{X_i\} \cup \mathbf{Z}}} \lambda_k f_k(\mathbf{x}_{S_k}) + \sum\limits_{\substack{k \text{ with} \\ \mathbf{x}_{S_k} = \{X_i, X_j\}}} \lambda_k f_k(\mathbf{x}_{S_k}) + \alpha\right]}{\sum\limits_{x_i} \exp\left[\sum\limits_{\substack{k \text{ with} \\ \mathbf{x}_{S_k} = \{X_i\} \cup \mathbf{Z}}} \lambda_k f_k(\mathbf{x}_{S_k}) + \sum\limits_{\substack{k \text{ with} \\ \mathbf{x}_{S_k} = \{X_i, X_j\}}} \lambda_k f_k(\mathbf{x}_{S_k}) + \alpha\right]} \end{aligned}$$

Using the linear independence of the functions f, it directly follows that

$$\begin{aligned} & \hat{p}(x_i \mid x_j, \mathbf{z}) = \hat{p}(x_i \mid \mathbf{z}) \\ \Rightarrow \quad & \lambda_k = 0 \quad \forall k \;\; \text{with} \;\; \mathbf{X}_{S_k} = \{X_i, X_j\} \end{aligned}$$

and from this, it directly follows the assertion. □

An alternative way to prove lemma 1 is by considering an undirected graphical model and using insights from information geometry. To do this, we consider an undirected graph G_U with a vertex set corresponding to the random variables $\mathbf{X}$. Furthermore, let the joint distribution $P(\mathbf{X})$ satisfy the global Markov condition on G_U and have strictly positive density $p(\mathbf{x}) > 0$. Then the Hammersley-Clifford theorem (Lauritzen, 1996) tells us that the joint density $p(\mathbf{x})$ can be factorised into

$$p(\mathbf{x}) = \frac{1}{\tilde{\alpha}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C) \tag{17}$$

with

$$\tilde{\alpha} = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C) \tag{18}$$

for some clique potentials $\psi_C : \mathcal{X}_C \to [0, \infty)$, where $\mathcal{C}$ is the set of maximal cliques of the graph G_U and $\mathbf{X}_C$ are the variables corresponding to the nodes in clique C. In a log-linear model, we can formulate the clique potentials as

$$\psi_C(\mathbf{x}_C) = \exp\left[\sum_{k=1}^{K} \theta_{C,k} h_k(\mathbf{x}_C)\right] \tag{19}$$

for some measurable functions $h_k : \mathcal{X}_C \to \mathbb{R}$. Hence, the joint density can be written in the form

$$p(\mathbf{x}) = \exp\left[\sum_{C,k} \theta_{C,k} h_k(\mathbf{x}_C) - \tilde{\alpha}\right] . \tag{20}$$

This strongly resembles the MAXENT distribution (see eq. (5)). And indeed, if the subsets of variables $\mathbf{X}_{S_k}$ observed in the different datasets would be equal to the maximal cliques of the undirected graph, there would be a one-on-one correspondence between the MAXENT solution and the factorised true distribution. As a result, we would directly get equivalence in eq. (12) in lemma 1. In general, however, this is not the case. Nevertheless, the clique potential formalism provides an additional way to prove lemma 1.

Alternative proof for lemma 1. Let us, without loss of generality, assume that $Z = \mathbf{Z}$ is one (vector-valued) variable. Then $X_i \perp\!\!\!\perp X_j \mid Z$ w.r.t. P implies that P can be represented by the undirected graphical model $X_i - Z - X_j$ (Lauritzen, 1996) or a subgraph of it (in case X_i or X_j are also independent of Z). Accordingly, P factorises according to the clique potentials of this graph and eq. (20). Thus P lies in the exponential manifold (Amari and Nagaoka, 1993) $\tilde{E}$ of distributions given by $\exp[h_1(x_i, z) + h_2(x_j, z) - \tilde{\alpha}]$ with arbitrary functions h_1, h_2. Let $E \supset \tilde{E}$ be the exponential manifold of distributions $\exp[h_1(x_i, z) + h_2(x_j, z) + h_3(x_i, x_j) - \tilde{\alpha}]$ with arbitrary functions h_1, h_2, h_3. By elementary results of information geometry (Amari and Nagaoka, 1993), $\hat{P}$ can also be defined as the projection of P onto E. Since P lies in $\tilde{E}$, and thus also in E, it follows that $P = \hat{P}$ in this case. This also implies that $h_3(x_i, x_j) = 0$ in the MAXENT distribution and thus $\sum_k \lambda_k f_k(x_i, x_j) = 0$. Due to the linear independence of the functions f this implies that $\lambda_k = 0$ for all k with $\mathbf{X}_{S_k} = \{X_i, X_j\}$. ☐

Lemma 2 (Causally linked variables have non-zero Lagrange multipliers)**.** *Let P have faithful f-expectations relative to a causal DAG G. Then it is $\lambda_k^P \neq 0$ for any bivariate function f_k whose variables are connected in G.*

Proof. If X_i and X_j are connected in G, the distribution $q(\mathbf{x}) \sim \exp[f_k(x_i, x_j)]$ is Markov relative to G. Obviously, it is $\lambda_k^Q = 1 \neq 0$, and due to P having faithful f-expectations it is also $\lambda_k^P \neq 0$. ☐

Theorem 1 (Causal structure from Lagrange multipliers)**.** *Let P be a distribution with faithful f-expectations w.r.t. a causal DAG G, and let $\hat{P}$ be the MAXENT solution satisfying the constraints imposed by the expectations of the functions f which are sufficient to uniquely describe the marginal distributions $P(X_i, \mathbf{Z}), P(X_j, \mathbf{Z})$, and $P(X_i, X_j)$. Then the following two statements hold:*

1. *If $\mathbf{Z}$ is d-separating X_i and X_j in G, then all Lagrange multipliers λ_k are zero for all k with $\mathbf{X}_{S_k} = \{X_i, X_j\}$.*
2. *If $\lambda_k = 0$ for all k with $\mathbf{X}_{S_k} = \{X_i, X_j\}$, then there is no direct link between X_i and X_j in the DAG G.*

Proof. The first statement follows from lemma 1, and the second statement directly follows from lemma 2. ☐

Corollary 1 (Identification of causal links when causal order is known)**.** *Let P be a distribution with faithful f-expectations w.r.t. a causal DAG G, and let $\hat{P}$ be the MAXENT solution satisfying the constraints imposed by the expectations of the functions f which are sufficient to uniquely describe the marginal distributions*

$P(X_i, \mathbf{Z}), P(X_j, \mathbf{Z})$, and $P(X_i, X_j)$. If it is excluded that X_j can causally influence X_i and $\mathbf{Z}$, i.e. the DAG G cannot contain edges $X_j \to X_i$ or $X_j \to \mathbf{Z}$, then it holds

$$\begin{aligned} &X_i \text{ is not directly linked to } X_j \\ \Leftrightarrow \quad &\lambda_k = 0 \quad \forall k \text{ with } \mathbf{X}_{S_k} = \{X_i, X_j\} \ . \end{aligned} \tag{13}$$

This also holds for conditional MAXENT, and if conditional means are used (see eq. (9)) it holds

$$\begin{aligned} &X_i \text{ is not directly linked to } X_j \\ \Leftrightarrow \quad &\hat{\lambda}_k^{\nu} = \hat{\lambda}_k^{\nu'} \quad \forall \nu, \nu', k \text{ with } \mathbf{X}_{S_k} = \{X_i, X_j\} \ . \end{aligned} \tag{14}$$

Proof. This directly follows from theorem 1. □

Theorem 2 (Graph constructed from MAXENT with only bivariate constraints is a supergraph of the moral graph)**.** *Let f be a basis for the space of univariate and bivariate functions, i.e. the set of f-expectations determine all bivariate distributions uniquely. Let P be a joint distribution that has faithful f-expectations w.r.t. the DAG G. Let G^b be the undirected graph constructed from the MAXENT distribution by connecting X_i and X_j if and only if there is a non-zero Lagrange multiplier corresponding to some bivariate function of X_i and X_j. Then G^b is a supergraph of G^m, the moral graph of G.*

Proof. The undirected graph G^b contains all edges of G due to lemma 2. It only remains to show that G^b also connects pairs with a common child. To show that G^b also connects pairs X_i, X_j with a common child X_c, we first consider the 3-node DAG $X_i \to X_c \leftarrow X_j$, and construct an example distribution, that is Markovian for this DAG, which uses only pair-interactions, including an interaction term X_i, X_j. By embedding this distribution into a general joint distribution, we conclude that common children can result in interaction terms after projection on pair interactions.

We define a Markovian distribution P via $P(X_i)P(X_j)P(X_c \mid X_i, X_j)$, with

$$P(X_c \mid X_i, X_j) := \exp\left[\phi_i(X_c, X_i) + \phi_j(X_c, X_j) - \log z(X_i, X_j)\right] \ ,$$

where the partition function z reads

$$z(X_i, X_j) := \sum_{x_c} \exp\left[\phi_i(x_c, X_i) + \phi_j(x_c, X_j)\right] \ .$$

By construction, P lies in the exponential manifold spanned by univariate and bivariate functions. It therefore coincides with the MAXENT distribution subject to all bivariate marginals. Thus, G^b contains the edge $X_i - X_j$ whenever $z(X_i, X_j)$ depends on both X_i and X_j. This dependence can be easily checked, for instance, for $\phi_i(x_c, x_i) := \delta_{x_c}\delta_{x_i}$ and $\phi_j(x_c, x_j) := \delta_{x_c}\delta_{x_j}$, where $\delta_{x_i}, \delta_{x_j}, \delta_{x_c}$ are indicator functions for arbitrary values, as defined in eq. (11).

For any DAG G with N variables containing the collider above as subgraph, $P(X_1, \ldots, X_N) \sim P(X_i, X_j, X_c)$ is also Markov relative to G and, at the same time, coincides with the MAXENT solution subject to the bivariate constraints. Hence the moral graph G^m still has an edge $X_i - X_j$ because there exists a distribution, Markovian to G, that has a bivariate term depending on X_i and X_j in the MAXENT distribution subject to all bivariate distributions. □

Theorem 3 (Predictive power of MAXENT)**.** *Let $X_j, X_i, \mathbf{Z}$ be binary variables, with $\mathbf{Z}$ possibly high dimensional. Furthermore, let $\hat{P}(X_j \mid X_i, \mathbf{Z})$ be the MAXENT solution that maximises the conditional entropy of X_j given X_i and $\mathbf{Z}$, subject to the moment constraints given by the observed pairwise distributions $P(X_j, X_i)$, $P(X_j, \mathbf{Z})$, and $P(X_i, \mathbf{Z})$. Then $\hat{P}$ is a better predictor of X_j than any of the individual bivariate probabilities, as measured by the likelihood of any point where all variables are observed, i.e. a point from $P(X_j, X_i, \mathbf{Z})$.*

Proof. The proof follows directly from the duality of MAXENT and maximum likelihood (Wainwright and Jordan, 2008). However, we prove it here using the Lagrange multipliers found by the optimisation procedure.

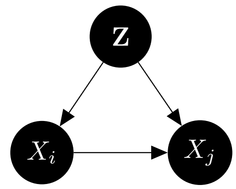


Figure 4: DAG for a treatment variable X_i influencing a target X_j in the presence of confounders $\mathbf{Z}$.

By the definition of maximum likelihood and MAXENT, we can write:

$$\mathbb{E}_{P(X_j,X_i,\mathbf{Z})}\left[\log \hat{P}(x_j \mid x_i, \mathbf{z})\right] = \mathbb{E}_{P(X_j,X_i,\mathbf{Z})}\left[\log \max_{\lambda} \exp\left(\sum_k \lambda_k f_k(x_j, \mathbf{z}) + \sum_l \lambda_l g_l(x_j, x_i) - \beta(x_i, \mathbf{z})\right)\right] \quad (21)$$

On the other hand, if we do not use the MAXENT solution, the maximum likelihood estimate we can attain consistent with $P(X_j, X_i)$ is $P(X_j \mid X_i)$. From eq. (21), we can attain that solution by setting all λ_k to zero. This means that if $P(X_j, \mathbf{Z})$ is not valuable in predicting the multipliers, then we attain the same solution as not using the information from $P(X_j, \mathbf{Z})$. However, if there is information to be exploited from the moments given by $P(X_j, \mathbf{Z})$, then the multipliers are not set to zero, attaining a higher likelihood. □

D OBTAINING INFORMATION ABOUT THE STRENGTH OF CAUSAL EFFECTS BY MERGING DATASETS

Another similarly essential and challenging task is to quantify the causal influence of a treatment on a target in the presence of confounders. In this section, we consider a scenario where we want to investigate the causal effect of a treatment variable X_i (e.g. the place of residence) on a target variable X_j (e.g. the depression rate) in the presence of confounders $\mathbf{Z}$ (e.g. the age that can influence both the depression rate and the place of residence, as displayed in fig. 4). Only pairwise observations for treatment – target and treatment – confounders are available in this scenario. To investigate the causal effect of X_i on X_j, first, we can use the results from sec. 3 and the MAXENT distribution to identify if there is a direct causal link from X_i to X_j. If this is the case, this section provides further insights into the causal relationship between X_i and X_j. Even without observing all variables jointly, we can derive bounds on the interventional distribution $P(X_j \mid do(X_i))$ and the ACE of X_i on X_j.

Background One of the core tasks in causality is computing interventional distributions. By answering the question "what would happen if variable X_i was set to value x_i?" they provide valuable information without actually having to perform an experiment in which X_i is set to the value x_i. Pearl's *do-calculus* (Pearl, 2000) provides the tools to compute the distribution $P(X_j \mid do(X_i))$ of X_j when intervening on X_i. In the infinite sample limit, the interventional distribution can be computed non-parametrically using backdoor adjustment

$$p(x_j \mid do(x_i)) = \sum_{\mathbf{z}} p(x_j \mid x_i, \mathbf{z}) p(\mathbf{z}) \;, \quad (22)$$

if $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$ is a set of nodes that contains no descendent of X_i and blocks all paths from X_i to X_j that contain an arrow into X_i (Pearl, 2000). In the case of binary variables, the interventional distribution can also be used to compute the average causal effect (ACE) of X_i on X_j:

$$ACE_{X_i \to X_j} = p(x_j = 1 \mid do(x_i = 1)) - p(x_j = 1 \mid do(x_i = 0)) \;. \quad (23)$$

Deriving Bounds on the Interventional Distribution and the ACE Using the backdoor adjustment in eq. (22) we can bound the interventional distribution based only on the observed marginal distributions.

Theorem 4. *Let X_i, X_j, and $\mathbf{Z}$ be discrete random variables in the causal DAG shown in fig. 4 with known marginal distributions $P(X_i, X_j)$ and $P(X_i, \mathbf{Z})$. Then the interventional distribution $P(X_j \mid do(X_i))$ is bounded as follows:*

$$\frac{p(x_j, x_i = x_i')}{\max_{\mathbf{z}} p(x_i = x_i' \mid \mathbf{z})} \leq p(x_j \mid do(x_i = x_i')) \leq \frac{p(x_j, x_i = x_i')}{\min_{\mathbf{z}} p(x_i = x_i' \mid \mathbf{z})} \;. \quad (24)$$

Proof. Using Pearl's backdoor adjustment in eq. (22) and Bayes' rule, we find

$$p(x_j \mid do(x_i = x_i')) = \sum_{\mathbf{z}} p(x_j \mid x_i = x_i', \mathbf{z}) p(\mathbf{z}) = \sum_{\mathbf{z}} p(x_j \mid x_i = x_i', \mathbf{z}) p(\mathbf{z}) \cdot \frac{p(x_i = x_i' \mid \mathbf{z})}{p(x_i = x_i' \mid \mathbf{z})}$$
$$\leq \frac{\sum_{\mathbf{z}} p(x_j \mid x_i = x_i', \mathbf{z}) p(\mathbf{z}) p(x_i = x_i' \mid \mathbf{z})}{\min_{\mathbf{z}} p(x_i = x_i' \mid \mathbf{z})} = \frac{p(x_j, x_i = x_i')}{\min_{\mathbf{z}} p(x_i = x_i' \mid \mathbf{z})} \,.$$

The lower bound can be derived analogously. □

In the case where X_i and X_j are binary, we can use eq. (24) also to bound the ACE of X_i on X_j.

Lemma 3. *In the setting described in theorem 4 the ACE of X_i on X_j is bounded as follows:*

$$\frac{p(x_j{=}1, x_i{=}1)}{\max_{\mathbf{z}} p(x_i{=}1 \mid \mathbf{z})} - \frac{p(x_j{=}1, x_i{=}0)}{\min_{\mathbf{z}} p(x_i{=}0 \mid \mathbf{z})} \leq ACE_{X_i \to X_j} \leq \frac{p(x_j{=}1, x_i{=}1)}{\min_{\mathbf{z}} p(x_i{=}1 \mid \mathbf{z})} - \frac{p(x_j{=}1, x_i{=}0)}{\max_{\mathbf{z}} p(x_i{=}0 \mid \mathbf{z})} \,. \tag{25}$$

Proof. This directly follows from theorem 4 and eq. (23). □

Lemma 3 provides us with at least an approximate insight of the strength of the causal effect of the variable X_i on X_j. In addition, the bounds provide a correct scale of the bounds that could be found using the MAXENT solution. However, as the MAXENT solution is an approximation to the true distribution, this will also be an approximation, and as it is a point estimate, we do not know how close or far we are from the true ACE. Therefore we report bounds to show what can be said from marginal distributions about the ACE even without MAXENT.

Related Work on Confounder Correction The classical task of confounder correction is to estimate the effect of a treatment variable on a target in the presence of unobserved confounders. In this paper, however, we consider the scenario shown in fig. 4 and assume that we have observations for the confounders $\mathbf{Z}$, but not for X_i, X_j, and $\mathbf{Z}$ jointly. If X_i, X_j, and $\mathbf{Z}$ were observed jointly, the causal effect of X_i on X_j would be identifiable and could be computed using Pearl's backdoor adjustment (see appendix A). In cases where $\mathbf{Z}$ is unobserved, the causal effect of X_i on X_j is not directly identifiable. One exception is if a set of observed variables satisfies the front-door criterion (Pearl, 2000). In Galles and Pearl (1995); Pearl (2000) and Kuroki and Pearl (2014) more general conditions were presented for which do-calculus and proxy-variables of unobserved confounders, respectively, make the causal effect identifiable. Another approach to confounder correction is by phrasing the problem in the potential outcome framework, e.g., using instrumental variables (Angrist et al., 1996; Grosse-Wentrup et al., 2016) or principal stratification (Rubin, 2004). Other, more recent approaches include, for instance: double/debiased machine learning (Chernozhukov et al., 2018; Jung et al., 2021); combinations of unsupervised learning and predictive model checking to perform causal inference in multiple-cause settings (Wang and Blei, 2019); methods that use limited experimental data to correct for hidden confounders in causal effect models (Kallus et al., 2018); and the split-door criterion which considers time series data where the target variable can be split into two parts of which one is only influenced by the confounders and the other is influenced by the confounders and the treatment, reducing the identification problem to that of testing for independence among observed variables (Sharma et al., 2018). Although confounder correction is a common and well-studied problem, we are unaware of approaches based on pairwise observations for treatment – target and treatment – confounders only.

E ADDITIONAL RELATED WORK

Gaining statistical information from causal knowledge One approach to using causal information to improve the approximation to the true joint distribution is *causal* MAXENT (Sun et al., 2006; Janzing et al., 2009), a particular case of conditional MAXENT, where the entropy of the variables is maximised along the causal order. For cause-effect relations, it just amounts to first maximising the entropy of the cause subject to all constraints that refer to it. Then, it maximises the conditional entropy for the effect given the cause subject to all constraints. Maximising the entropy in the causal order results in a distribution with lower entropy than maximising the entropy jointly. Consequently, the distribution learned in the causal order will have better predictive power. Another simple example of how causal information can help gaining statistical insights is the following: Imagine we are given the bivariate marginal distributions $P(X_1, X_2)$ and $P(X_2, X_3)$. In the general case, where we do not know the causal graph, we could not identify the joint distribution. However, when we know that the three variables form a causal chain $X_1 \to X_2 \to X_3$, this causal information is enough to identify the joint distribution uniquely (Janzing, 2018). [3] Even perfect causal knowledge does not uniquely determine the joint distribution for less simplistic scenarios. But causal information may still help to get *some* properties of the joint distribution. In Tsamardinos et al. (2012), for instance, the causal structure is used to predict CI of variables that have not been observed together. This paper approaches a complementary problem: gaining causal insights by merging statistical information from different datasets.

Entropy based approaches to extract or exploit causal information Different methods exposing the relationship between information theory and causality are present in the literature. In Kocaoglu et al. (2017); Compton et al. (2021) properties of the entropy are used to infer the causal direction between categorical variable pairs. Their main idea is that if the entropy of the exogenous noise of a functional assignment in a structural causal model (SCM) is low, then the causal direction often becomes identifiable. Their approach differs from ours in several respects: first, we investigate the absence and presence of causal edges from merged data, as opposed to trying to infer the causal direction; second, we are not constrained to variable pairs; finally, we use the entropy as the function we want to optimise directly while they compare the entropy of each of the noise variables to decide the causal direction. In Ziebart et al. (2010, 2013) the maximum causal entropy is introduced to solve inverse reinforcement learning problems. Their approach is based on having knowledge about a possible causal graph, making the MAXENT computation cheaper by exploiting the causal structure of the data. Their work differs from ours on the type of insights we get from the MAXENT estimation: while they are trying to save computation, we are trying to identify causal edges from the Lagrange multipliers.

Semi-supervised learning (SSL) The relation to semi-supervised learning (SSL) is interesting but still unexplored. The high-level connection, which we can mention, is that SSL uses $P(X)$ to infer properties of $P(X, Y)$, which has been claimed to be only possible if Y is the cause and X the effect, but not vice versa Schölkopf et al. (2013). Hence, SSL also infers joint properties from the marginal but relies on model assumptions like cluster assumption, manifold assumption, smoothness of decision boundaries. To relate this inductive bias with MAXENT probably requires defining the correct type of functions f.

F ADDITIONAL RESULTS

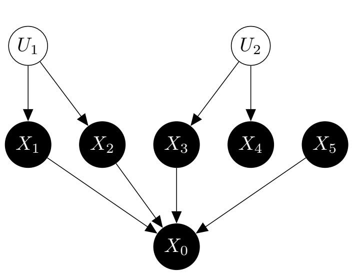
In this section, we provide exemplary plots of the Lagrange multipliers for the synthetic experiments discussed in sec. 5, as well as further results for the experiments on the two real-world datasets.

Synthetic Data In fig. 5 we show exemplary results for the Lagrange multipliers for the experiments with synthetic data discussed in sec. 5. For each of the three graphs in figs. 2a to 2c we randomly picked one dataset, for which we show the exact used graph structure in figs. 5a to 5c and the found Lagrange multipliers in figs. 5d to 5f. In all three cases, the difference between the multiplier associated with $\mathbb{E}[X_0 \mid X_i = 0]$ and the one associated with $\mathbb{E}[X_0 \mid X_i = 1]$ is very small whenever there is no edge from X_i to X_0, and relatively large whenever there is an edge connecting X_i and X_0.

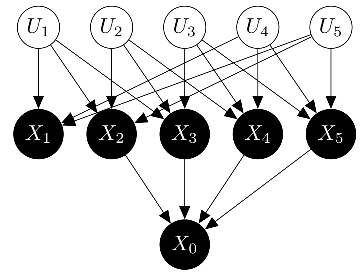
Real Data First, we further investigate the results from the experiment on the data from Gapminder (2021). For this, we consider different subsets of the variables

- children per woman / fertility (FER) (Gapminder, 2021);
- inflation-adjusted Gross Domestic Product (GDP) per capita (World Bank, 2019);
- Human Development Index (HDI) (United Nations Development Programme, Human Development Reports, 2019); and
- life expectancy (LE) in years (Gapminder, 2021).

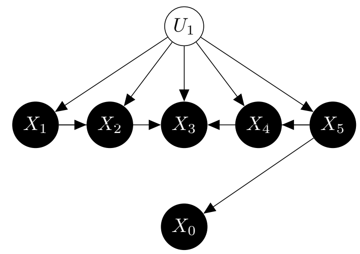
[3] In general, it is a non-trivial problem to decide whether a set of marginal distributions of different but non-disjoint sets of variables are consistent with a joint distribution (the so-called 'marginal problem' (Vorob'ev, 1962)).



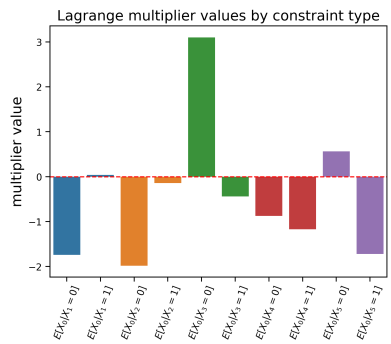
(a) Exemplary graph (a)



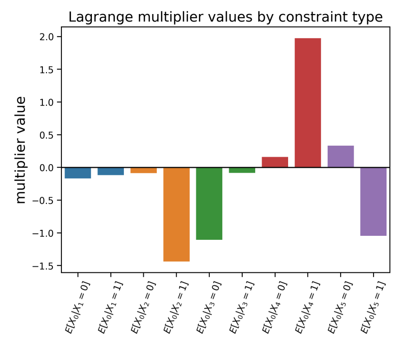
(b) Exemplary graph (b)



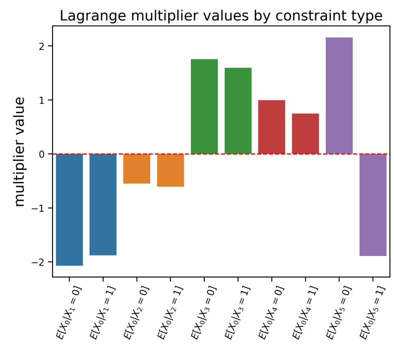
(c) Exemplary graph (c)



(d) Exemplary result for the Lagrange multipliers for graph (a)



(e) Exemplary result for the Lagrange multipliers for graph (b)



(f) Exemplary result for the Lagrange multipliers for graph (c)

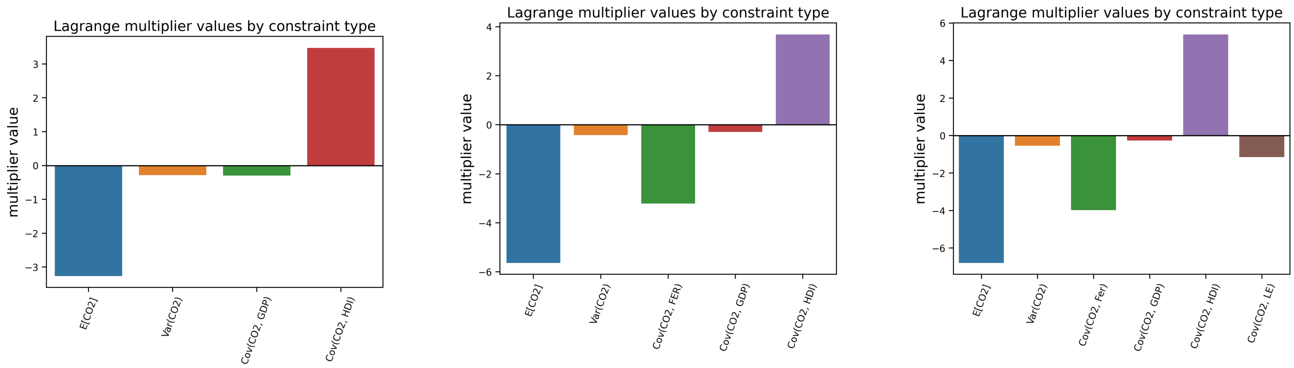
Figure 5: We show the found Lagrange multipliers for a randomly picked dataset for the displayed graphs. One can see that in all three cases, the multipliers are very close to being constant whenever an edge is missing. On the other hand, the differences between them are significant in the cases where there is an edge from X_i to X_0.

as potential causes of the target variable CO2 tonnes emission per capita (Boden et al., 2017). We always use data from 2017 for all variables and standardise it before estimation. We always use the unconditional mean and variance of CO2 emissions, and the pairwise covariance between CO2 emissions and each considered variable as constraints. We compare our results with the output of the KCI-test, where we use a significance threshold of $\alpha = 0.05$.

The results in fig. 6 and table 2 show that the conclusions drawn from the Lagrange multipliers are consistent over the different sets of considered potential causes. For the KCI-test, on the other hand, we get separate statements about the CIs of CO2 and HDI, and CO2 and FER, depending on the considered conditioning set. At first glance, this might be not surprising as, of course, the CI relationships can change when considering more variables. For instance, one could imagine that the causal effect of HDI on CO2 is only via FER. This would explain the behaviour of the KCI-test w.r.t. HDI. To check if this is the case – which would contradict the result from the Lagrange multipliers –, we perform another KCI-test for CO2 and HDI conditioned only on FER. Summarising the obtained CI statements, we get:

$$CO2 \not\perp\!\!\!\perp HDI \mid GDP \tag{26}$$
$$CO2 \not\perp\!\!\!\perp HDI \mid FER \tag{27}$$
$$CO2 \perp\!\!\!\perp HDI \mid GDP, FER \tag{28}$$
$$CO2 \perp\!\!\!\perp GDP \mid HDI \tag{29}$$
$$CO2 \perp\!\!\!\perp GDP \mid HDI, FER \tag{30}$$
$$CO2 \not\perp\!\!\!\perp FER \mid HDI, GDP \tag{31}$$

If we now try to draw a causal DAG for these four variables based on the CIs in eqs. (26) to (31), we see that this is not so simple. In fact, it is not possible to construct a DAG over these four variables that is consistent with eqs. (26) to (31). There are, of course, many possible reasons why the KCI-test provides these seemingly inconsistent results. For instance, one could argue that choosing $\alpha = 0.05$ as a threshold for the decision is very arbitrary and maybe a non-optimal choice. All obtained p-values were relatively small (less than 0.2), which



(a) considering GDP and HDI

(b) considering FER, GDP, and HDI

(c) considering FER, GDP, HDI, and LE

Figure 6: Lagrange multipliers for the real-world dataset from Gapminder (2021) when considering different sets of variables as potential causes for CO2 emissions. We see that in all three cases, the results for GDP and HDI are consistent, indicating that HDI is directly linked to CO2 and GDP not.

might indicate that the question ‘CI or no CI?’ might not be so easy to answer in this case. Furthermore, we do not know whether the considered example violates some of the assumptions made in the KCI-test.

This shows that the KCI-test should not be mistaken with ‘ground truth’, and the fact that the conclusions drawn from the Lagrange multipliers do not always coincide with the conclusions drawn from the KCI-test is not necessarily a problem of our proposed approach.

Finally, we show in fig. 7 the Lagrange multipliers for the experiment on depression rate w.r.t. place of residence, age, and sex. We see that for all three factors the multipliers are *not* constant across the various conditions. This indicates that all three factors might directly cause the depression rate.

G IMPLEMENTATION DETAILS

We implemented MAXENT on Python using JAX (Bradbury et al., 2018) optimisation procedures. We minimise the sum of the squares of the difference between the moments given as constraints and the moments estimated using the MAXENT distribution entailed by the Lagrange multipliers. If the absolute difference between the data expectations and the MAXENT expectations were smaller than 0.001, the procedure was considered convergent. In our current implementation, we estimate the normalising constant, although there is the possibility to use approximation methods to make the computation faster, if required (Wainwright and Jordan, 2008).

H EXPERIMENTAL SETUP

In all experiments, we observe only expectations associated with the X_i variables. To build the ROC curves for each of the samples obtained, we first generated a vector p of probabilities from a $\mathcal{U}(0.1, 0.9)$ distribution. In all the following examples, we generated 1000 observations for 100 repetitions of the SCM and estimated the empirical expectations from that sample. If the procedure did not converge, we did not take it into account for the ROC. We also randomised the logical relation between the causes X_i and the effect X_0. We denote this logical relation below as $\odot \in \{\wedge, \vee, \oplus\}$. The generative processes for the shown experiments with synthetically generated data are the following:

First, we select the used parameters as follows:

$$\begin{aligned} u_l &\sim \mathcal{N}(0,1) && \text{for} \quad l \in \{1, \dots, 5\} \\ p_k &\sim \mathcal{U}(0.1, 0.9) && \text{for} \quad k = 0, \dots, 5 \\ a_i &\sim \mathcal{N}(0,1) && \text{for} \quad i = 1, \dots, 5 \\ b_{i,j} &\sim \mathcal{N}(0,1) && \text{for} \quad j = 1, \dots, 5 \end{aligned}$$

Table 2: We show the found Lagrange multipliers λ_i for the MAXENT solution (see (a) to c)) together with the p-values for the KCI-test (see (d) to f)) . We indicate where the multipliers and p-values indicate the presence of a direct edge connecting X_i and CO2 emissions, or, respectively, that the two are not CI given the other considered variable(s). We see that the conclusions drawn from the Lagrange multipliers are consistent across the different considered sets of potential causes, while the CI statements of the KCI-test change when changing the conditioning set.

(a) considering GDP and HDI as potential causes

variable X_i	λ_i	edge
GDP	-0.29	✗
HDI	3.26	✓

(b) considering FER, GDP, and HDI as potential causes

variable X_i	λ_i	edge
FER	-3.22	✓
GDP	-0.29	✗
HDI	3.69	✓

(c) considering FER, GDP, HDI, and LE as potential causes

variable X_i	λ_i	edge
FER	-3.98	✓
GDP	-0.27	✗
HDI	5.40	✓
LE	-1.15	✓

(d) considering GDP and HDI as potential causes

variable X_i	p-value	no CI
GDP	0.19	✗
HDI	0.02	✓

(e) considering FER, GDP, and HDI as potential causes

variable X_i	p-value	no CI
FER	0.03	✓
GDP	0.13	✗
HDI	0.13	✗

(f) considering FER, GDP, HDI, and LE as potential causes

variable X_i	p-value	no CI
FER	0.08	✗
GDP	0.16	✗
HDI	0.14	✗
LE	0.00	✓

Then we use these parameters to generate the data for the variables X_0 to X_5.

For the experiment in fig. 2a the data is generated according to:

$$
\begin{aligned}
x_1 &\sim |\mathrm{Ber}(p_1) - (u_1 > 0)| \\
x_2 &\sim |\mathrm{Ber}(p_2) - (u_1 < 0.25)| \\
x_3 &\sim |\mathrm{Ber}(p_3) - (u_2 > 0)| \\
x_4 &\sim |\mathrm{Ber}(p_4) - (u_2 > 0.25)| \\
x_5 &\sim \mathrm{Ber}(p_5) \\
x_0 &\sim \mathbf{1}_{>0}\left[\left(\sum_i a_i X_i + \sum_{i,j} b_{i,j} X_i X_j\right)\right] \odot \mathrm{Ber}(p_0)
\end{aligned}
$$

And finally, for the experiment in fig. 2b the data is generated according to:

$$
\begin{aligned}
x_1 &\sim |\mathrm{Ber}(p_1) - (u_1 > 0 \vee u_2 > 0.25 \vee u_3 > 0.5)| \\
x_2 &\sim |\mathrm{Ber}(p_2) - (u_2 < 0.5 \vee u_3 < 0.25 \vee u_4 < 0)| \\
x_3 &\sim |\mathrm{Ber}(p_3) - (u_3 > 0 \vee u_4 < 0.25 \vee u_5 > 0.5)| \\
x_4 &\sim |\mathrm{Ber}(p_4) - (u_4 < 0.5 \vee u_5 > 0.25 \vee u_1 < 0)| \\
x_5 &\sim |\mathrm{Ber}(p_5) - (u_5 > 0 \vee u_1 < 0.25 \vee u_2 > 0.5)| \\
x_0 &\sim \mathbf{1}_{>0}\left[\left(\sum_i a_i X_i + \sum_{i,j} b_{i,j} X_i X_j\right)\right] \odot \mathrm{Ber}(p_0)
\end{aligned}
$$

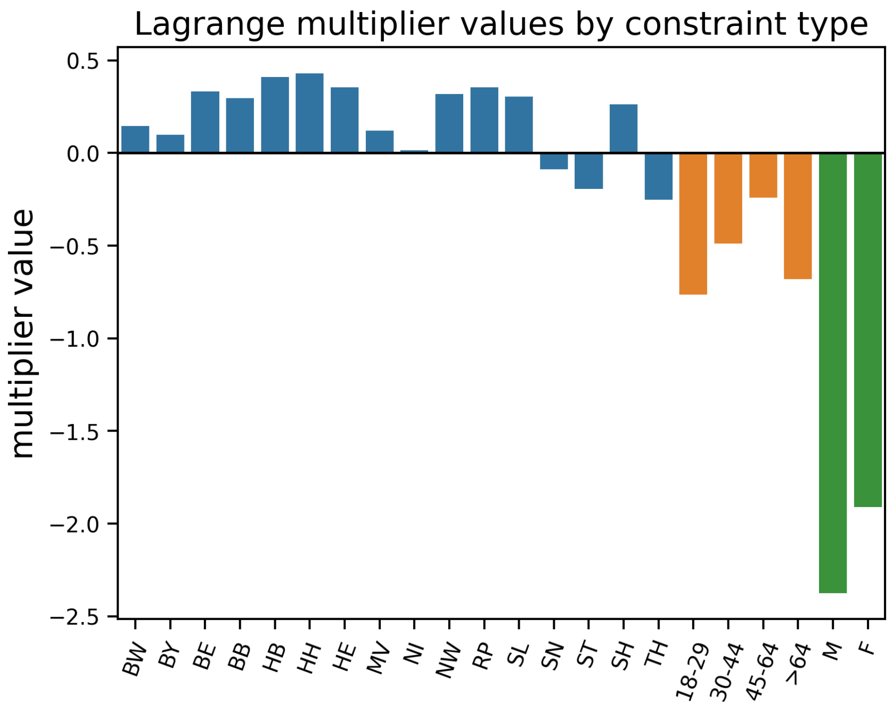


Figure 7: Lagrange multipliers for the depression dataset. We see that for none of the three potential causes place of residence, age, and sex the multipliers are close to being constant. Hence we conclude that all three factors can directly have impact on the depression rate.

For the experiment in fig. 2c we used the following generative process:

$$
\begin{aligned}
x_1 &\sim |\mathrm{Ber}(p_1) - (u_1 > 0)| \\
x_5 &\sim |\mathrm{Ber}(p_5) - (-0.25 < u_1 < 0.25)| \\
x_2 &\sim |\mathrm{Ber}(p_2) - (u_1 < 0)| \vee x_1 \\
x_4 &\sim |\mathrm{Ber}(p_4) - (u_1 < -0.25)| \vee x_5 \\
x_3 &\sim |\mathrm{Ber}(p_3) - (u_1 > 0.25)| \vee (x_1 \oplus x_5) \\
x_0 &\sim \mathbf{1}_{>0}\left[\left(\sum_i a_i X_i + \sum_{i,j} b_{i,j} X_i X_j\right)\right] \odot \mathrm{Ber}(p_0)
\end{aligned}
$$