Distributional equivalence and structure learning for bow-free acyclic path diagrams

Christopher Nowzohour

Seminar für Statistik, ETH Zürich
e-mail: nowzohour@stat.math.ethz.ch

Marloes H. Maathuis

Seminar für Statistik, ETH Zürich
e-mail: maathuis@stat.math.ethz.ch

Robin J. Evans[*]

Department of Statistics, University of Oxford
e-mail: evans@stats.ox.ac.uk

and

Peter Bühlmann

Seminar für Statistik, ETH Zürich
e-mail: buhlmann@stat.math.ethz.ch

Abstract: We consider the problem of structure learning for bow-free acyclic path diagrams (BAPs). BAPs can be viewed as a generalization of linear Gaussian DAG models that allow for certain hidden variables. We present a first method for this problem using a greedy score-based search algorithm. We also prove some necessary and some sufficient conditions for distributional equivalence of BAPs which are used in an algorithmic approach to compute (nearly) equivalent model structures. This allows us to infer lower bounds of causal effects. We also present applications to real and simulated datasets using our publicly available R-package.

Contents

1. Introduction

We consider learning the causal structure among a set of variables from observational data. In general, the data can be modelled with a structural equation model (SEM) over the observed and unobserved variables, which expresses each variable as a function of its direct causes and a noise term, where the noise terms are assumed to be mutually independent. The structure of the SEM can be visualized as a directed graph, with vertices representing variables and edges representing direct causal relationships. We assume the structure to be recursive (acyclic), which results in a directed acyclic graph (DAG). DAGs can be understood as models of conditional independence, and many structure learning algorithms use this to find all DAGs which are compatible with the observed conditional independencies (Spirtes et al., 1993). Often, however, not all relevant variables are observed. The resulting marginal distribution over the observed variables might still satisfy some conditional independencies, but in general these will not have a DAG representation (Richardson and Spirtes, 2002). Also, there generally are additional constraints resulting from the marginalization of some of the variables (Evans, 2016; Shpitser et al., 2014).

In this paper we consider a model class which can accommodate certain hidden variables. Specifically, we assume that the graph over the observed variables is a bow-free acyclic path diagram (BAP). This means it can have directed as well as bidirected edges (with the directed part being acyclic), where the directed edges represent direct causal effects, and the bidirected edges represent

hidden confounders. The bow-freeness condition means there cannot be both a directed and a bidirected edge between the same pair of variables. The BAP can be obtained from the underlying DAG over all (hidden and observed) variables via a latent projection operation (Pearl, 2000) (if the bow-freeness condition admits this). We furthermore assume a parametrization with linear structural equations and Gaussian noise, where two noise terms are correlated only if there is a bidirected edge between the two respective nodes. In certain situations, it is beneficial to consider this restricted class of hidden variable models, as it forms a middle ground between DAG models that don't allow any hidden variables and maximal ancestral graph (MAG) models (Richardson and Spirtes, 2002) that allow arbitrarily many and general hidden variables. Such a restricted model class, if not heavily misspecified, results in a smaller distributional equivalence class, and estimation is expected to be more accurate than for more general hidden variable methods like FCI (Spirtes et al., 1993), RFCI (Colombo et al., 2012), or FCI+ (Claassen et al., 2013).

The goal of this paper is structure learning with BAPs, that is, finding the set of BAPs that best explains some observational data. Just like in other models, there is typically an equivalence class of BAPs that are statistically indistinguishable, so a meaningful structure search result should represent this equivalence class. We propose a penalized likelihood score that is greedily optimized and a heuristic algorithm (supported by some theoretical results) for finding equivalent models once an optimum is found. This method is the first of its kind for BAP models.

Example of a BAP

Consider the DAG in Figure 1a, where we observe variables $X_1, \ldots, X_4$, but do not observe H_1, H_2, H_3. The only (conditional) independency over the observed variables is $X_1 \perp\!\!\!\perp X_3 \mid X_2$, which is also represented in the corresponding BAP in Figure 1b. The parametrization of this BAP would be

$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= B_{21}X_1 + \epsilon_2 \\
X_3 &= B_{32}X_2 + \epsilon_3 \\
X_4 &= B_{43}X_3 + \epsilon_4
\end{aligned}
$$

with $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)^T \sim \mathcal{N}(\mathbf{0}, \Omega)$ and

$$
\Omega = \begin{pmatrix}
\Omega_{11} & 0 & 0 & 0 \\
0 & \Omega_{22} & 0 & \Omega_{24} \\
0 & 0 & \Omega_{33} & 0 \\
0 & \Omega_{24} & 0 & \Omega_{44}
\end{pmatrix}.
$$

Hence the model parameters in this case are B_{21}, B_{32}, B_{43}, Ω_{11}, Ω_{22}, Ω_{33}, Ω_{44}, and Ω_{24}. An example of an acyclic path diagram that is not bow-free is shown in Figure 1c.
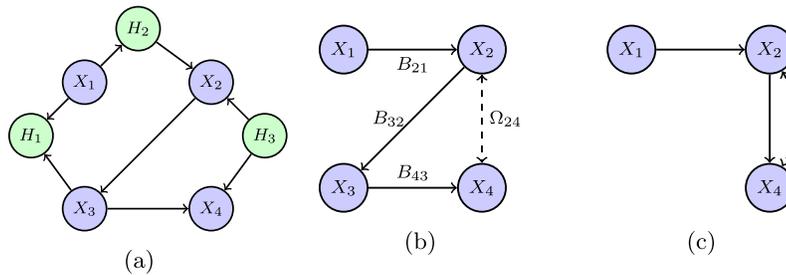
FIG 1. *(a) DAG with hidden variables H_1, H_2, H_3, (b) resulting BAP over the observed variables $X_1, \ldots, X_4$ with annotated edge weights, and (c) resulting graph if X_3 is also not observed, which is not a BAP.*

Challenges

The main challenge, like with all structure search problems in graphical modelling, is the vastness of the model space. The number of BAPs grows superexponentially. Hence, as is the case for DAGs, exhaustively scoring all BAPs and finding the global score optimum is very challenging. For DAGs, Silander and Myllymäki (2006) proposed a surprisingly simple algorithm whose runtime is exponential in the number of nodes and which is feasible for problems with up to about 30 nodes. However, extending their idea to BAPs is not straightforward, and we aim to deal with settings where the number of nodes can be significantly larger.

Another major challenge, specifically for our setting, is the fact that a graphical characterization of the (distributional) equivalence classes for BAP models is not yet known. In the DAG case, for example, it is known that models are equivalent if and only if they share the same skeleton and v-structures (Verma and Pearl, 1991). A similar result is not known for BAPs (or the more general acyclic directed mixed graphs). This makes it hard to traverse the search space efficiently, since one cannot search over the equivalence classes (like the greedy equivalence search for DAGs, see Chickering (2002)). It also makes it difficult to evaluate simulation results, since the graphs corresponding to the ground truth and the optimal solution may be distinct and yet still represent the same model.

Contributions

We provide the first structure learning algorithm for BAPs. It is a score-based algorithm and uses greedy hill climbing to optimize a penalized likelihood score. We are able to achieve a significant computational speedup by decomposing the score over the bidirected connected components of the graph and caching the score of each component. To mitigate the problem of local optima, we perform many random restarts of the greedy search.

We propose to approximate the distributional equivalence class of a BAP by using a greedy strategy for likelihood scoring. If two BAPs are similar with

respect to their penalized likelihoods within a tolerance, they should be treated as statistically indistinguishable and hence as belonging to the same class of (nearly) equivalent BAPs. Based on such greedily computed (near) equivalence classes, we can then infer bounds of total causal effects, in the spirit of Maathuis et al. (2009, 2010).

We present some theoretical results towards equivalence properties in BAP models, some of which generalize to acyclic path diagrams. In particular, we prove some necessary and some sufficient conditions for BAP equivalence. Furthermore, we present a Markov Chain Monte Carlo method for uniformly sampling BAPs based on ideas from Kuipers and Moffa (2015).

We obtain promising results on simulated data sets despite the challenges listed above. Comparing the highest-scoring BAPs and DAGs on real datasets exemplifies the more conservative nature of BAP models.

Related work

There are two main research communities that intersect at this topic. On the one side there are the path diagram models, going back to Wright (1934) and then being mainly developed in the behavioral sciences (Jöreskog, 1970; Duncan, 1975; Glymour and Scheines, 1986; Jöreskog, 2001). In this setting a model for the edge functions is assumed, usually a parametric model with linear edge functions and Gaussian noise. In a very general formulation, the graph over the observed variables is assumed to be an acyclic directed mixed graph (ADMG), which can have bows. While in general the parameters for these models are not identified, Drton et al. (2011) give necessary and sufficient conditions for global identifiability. Complete necessary and sufficient conditions for the more useful almost everwhere identifiability remain unknown (however, see Foygel et al. (2012) for some necessary and some sufficient conditions). BAP models are a useful subclass, since they are almost everywhere identified (Brito and Pearl, 2002). Drton et al. (2009) provided an algorithm, called residual iterative conditional fitting (RICF), for maximum likelihood estimation of the parameters for a given BAP.

On the other side there are the non-parametric hidden variable models, which are defined as marginalized DAG models (Pearl, 2000)[1]. The marginalized distributions are constrained by conditional independencies, as well as additional equality and inequality constraints (Evans, 2016). When just modelling the conditional independence constraints, the class of maximal ancestral graphs (MAGs) is sufficient (Richardson and Spirtes, 2002). Shpitser et al. (2014) have proposed the nested Markov model using ADMGs to also include the additional equality constraints. Finally, mDAGs were introduced to model all resulting constraints (Evans, 2016). In general BAPs induce independence constraints and also Verma constraints (Richardson and Spirtes, 2002, Sections 7.3 and 8),

---

[1]Strictly speaking, not all SEMs with correlated Gaussian errors can be interpreted as latent variable models, since the latent variable models have additional inequality constraints. We do not discuss this further here, but see Fox et al. (2015) for more details.

as well as other restrictions that do not apply in the non-parametric case. The BAP in Figure 1b, for example, implies a Verma constraint. Gaussian BAPs are also 'maximal', in the sense that every missing edge induces a constraint. In the non-parametric case, with each additional layer of constraints learning the graphical structure from data becomes more complicated, but at the same time more available information is utilized and a possibly more detailed structure can be learned. In the Gaussian case, however, all models are parameteric, and fitting BAPs that do not correspond to conditional independence models is essentially no different to fitting those that do. At the graphical level the search is perhaps easier, since we do not have to place the restriction of ancestrality on the structure of the graph. However, unlike for MAGs, the equivalence class of BAPs is not known, which means that one may end up fitting the same model multiple times in the form of different graphs. Furthermore, BAPs are easier to interpret as hidden variable models. This can be seen when comparing the BAP in Figure 1b with the corresponding MAG. The latter would have an additional edge between X_1 and X_4 since there is no (conditional) independency of these two variables. As can be verified, the BAP and the MAG in this example are not distributionally equivalent, since the former encodes additional non-independence constraints.

Structure search for MAGs can be done with the FCI (Spirtes et al., 1993), RFCI (Maathuis et al., 2009), or FCI+ (Claassen et al., 2013) algorithms. Silva and Ghahramani (2006) propose a fully Bayesian method for structure search in linear Gaussian ADMGs, sampling from the posterior distribution using an MCMC approach. Shpitser et al. (2012) employ a greedy approach to optimize a penalized likelihood over ADMGs for discrete parametrizations.

Outline of this paper

In Section 2 we give an in-depth overview of the model and its estimation from data, as well as some distributional equivalence properties. In Section 3 we present the details of our greedy algorithm with various computational speedups. In Section 4 we present empirical results on simulated and real datasets. All proofs as well as further theoretical results and justifications can be found in the Appendix.

2. Model and estimation

2.1. Graph terminology

Let $X_1, \ldots, X_d$ be a set of random variables and $V = \{1, \ldots, d\}$ be their index set. The elements of V are also called *nodes* or *vertices*. A *mixed graph* or *path diagram* G on V is an ordered tuple $G = (V, E_D, E_B)$ for some $E_D, E_B \subseteq V \times V \setminus \{(i, i) \mid i \in V\}$. If $(i, j) \in E_D$, we say there is a *directed edge* from i to j and write $i \rightarrow j \in G$. If $(i, j) \in E_B$, we must also have $(j, i) \in E_B$, and we say there is a *bidirected edge* between i and j and write $i \leftrightarrow j \in G$.

The set $\mathrm{pa}_G(i) \coloneqq \{j \mid j \to i \in G\}$ is called the *parents* of i. This definition extends to sets of nodes S in the obvious way: $\mathrm{pa}_G(S) \coloneqq \bigcup_{i \in S} \mathrm{pa}_G(i)$. The *in-degree* of i is the number of arrowheads at i. If $V' \subseteq V$, $E'_D \subseteq E_D|_{V' \times V'}$, and $E'_B \subseteq E_B|_{V' \times V'}$, then $G' = (V', E'_D, E'_B)$ is called a *subgraph* of G, and we write $G' \subseteq G$. The *induced subgraph* G_W for some vertex set $W \subset V$ is the restriction of V to vertices W. If $G' \subseteq G$ but $G' \neq G$, we call G' a *strict subgraph* of G and write $G' \subset G$. The *skeleton* of G is the undirected graph over the same node set V and with edges $i - j$ if and only if $i \to j \in G$ or $i \leftrightarrow j \in G$ (or both).

A *path* π between i and j is an ordered tuple of (not necessarily distinct) nodes $\pi = (v_0 = i, \ldots, v_l = j)$ such that there is an edge between v_k and v_{k+1} for all $k = 0, \ldots, l-1$. If the nodes are distinct, the path is called *non-overlapping* (note that in the literature a path is mostly defined as non-overlapping). The *length* of π is the number of edges $\lambda(\pi) = l$. If π consists only of directed edges pointing in the direction of j, it is called a *directed path* from i to j. A node j on a non-overlapping path π is called a *collider* if π contains a non-overlapping subpath (i, j, k) with two arrowheads into j[2]. Otherwise j is called a non-collider on the path. If j is a collider on a non-overlapping path (i, j, k), we call (i, j, k) a *collider triple*. Moreover, if (i, j, k) is a collider triple and i and k are not adjacent in the graph, then (i, j, k) is called a *v-structure*. A path without colliders is called a *trek*.

Let $A, B, C \subset V$ be three disjoint sets of nodes. The set $\mathrm{an}(C) \coloneqq C \cup \{i \in V \mid \text{there exists a directed path from } i \text{ to } c \text{ for some } c \in C\}$ is called the *ancestors* of C. A non-overlapping path π from $a \in A$ to $b \in B$ is said to be *m-connecting given* C if every non-collider on π is not in C and every collider on π is in $\mathrm{an}(C)$. If there are no such paths, A and B are *m-separated given* C, and we write $A \perp\!\!\!\perp_m B \mid C$. We use a similar notation for denoting conditional independence of the corresponding set of variables X_A and X_B given X_C: $X_A \perp\!\!\!\perp X_B \mid X_C$.

A graph G is called *cyclic* if there are at least two distinct nodes i and j such that there are directed paths both from i to j and from j to i. Otherwise G is called *acyclic* or *recursive*. An acyclic path diagram is also called an acyclic directed mixed graph (ADMG). An acyclic path diagram having at most one edge between each pair of nodes is called a *bow-free*[3] *acyclic path diagram (BAP)*. An ADMG without any bidirected edges is called a *directed acyclic graph (DAG)*.

2.2. The model

A *linear structural equation model (SEM)* M is a set of linear equations involving the variables $\mathbf{X} = (X_1, \ldots, X_d)^T$ and some error terms $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_d)^T$:

$$\mathbf{X} = B\mathbf{X} + \boldsymbol{\epsilon}, \tag{1}$$

where B is a real matrix, $\mathrm{cov}(\boldsymbol{\epsilon}) = \Omega$ is a positive semi-definite matrix, and we assume that all variables $\mathbf{X}$ have been normalized to mean zero. M has an

---

[2]That is, one of the following structures: $\to\leftarrow, \leftrightarrow\leftarrow, \to\leftrightarrow, \leftrightarrow\leftrightarrow$.
[3]The structure $i \to j$ together with $i \leftrightarrow j$ is also known as *bow*.

associated graph G that reflects the structure of B and Ω. For every non-zero entry B_{ij} there is a directed edge from j to i, and for every non-zero entry Ω_{ij} there is a bidirected edge between i and j. Thus we can also write (1) as:

$$X_i = \sum_{j \in \mathrm{pa}_G(i)} B_{ij} X_j + \epsilon_i, \qquad \text{for all } i \in V, \tag{2}$$

with $\mathrm{cov}(\epsilon_i, \epsilon_j) = \Omega_{ij}$ for all $i, j \in V$.

Our model is a special type of SEM, which we refer to as *Gaussian BAP model*[4]. In particular, we make the following assumptions:

(A1) The errors $\boldsymbol{\epsilon}$ follow a multivariate Normal distribution $\mathcal{N}(\mathbf{0}, \Omega)$.
(A2) The associated graph G is a BAP.

Assumption (A1) is not strictly needed for our equivalence results, but we rely on it for fitting the models in practice using the RICF method of Drton et al. (2009).

Often M is specified via its graph G, and we are interested to find parameters θ_G compatible with G. We thus define the parameter spaces for the edge weight matrices B (directed edges) and Ω (bidirected edges) for a given BAP G as

$$\mathcal{B}_G = \{B \in \mathbb{R}^{d \times d} \mid B_{ij} = 0 \text{ if } j \to i \text{ is not an edge in } G\}$$
$$\mathcal{O}_G = \{\Omega \in \mathbb{R}^{d \times d} \mid \Omega_{ij} = 0 \text{ if } i \neq j \text{ and } i \leftrightarrow j \text{ is not an edge in } G;$$
$$\Omega \text{ is symmetric positive semi-definite}\}$$

and the combined parameter space as

$$\Theta_G = \mathcal{B}_G \times \mathcal{O}_G.$$

The covariance matrix for $\mathbf{X}$ is given by:

$$\phi(\theta) = (I - B)^{-1} \Omega (I - B)^{-T}, \tag{3}$$

where $\phi : \Theta_G \to \mathcal{S}_G$ maps parameters to covariance matrices, and $\mathcal{S}_G := \phi(\Theta_G)$ is the set of covariance matrices compatible with G. Note that $\phi(\theta)$ exists since G is acyclic by (A2) and therefore $I - B$ is invertible.

We assume that the variables are normalized to have variance 1, that is, we are interested in the subset $\bar{\mathcal{S}}_G \subset \mathcal{S}_G$, where $\bar{\mathcal{S}}_G = \{\Sigma \in \mathcal{S}_G \mid \Sigma_{ii} = 1 \text{ for all } i = 1, \dots, d\}$, and its preimage under ϕ, $\bar{\Theta}_G := \phi^{-1}(\bar{\mathcal{S}}_G) \subset \Theta_G$.

One of the main motivations of working with BAP models is parameter identifiability. This is defined below:

Definition 1. *A normalized parameter $\theta_G \in \bar{\Theta}_G$ is identifiable if there is no $\theta'_G \in \bar{\Theta}_G$ such that $\theta_G \neq \theta'_G$ and $\phi(\theta_G) = \phi(\theta'_G)$.*

---

[4] All BAP models in this paper are assumed to have a Gaussian parametrization unless otherwise stated.

Brito and Pearl (2002) show that for any BAP G, the set of normalized non-identifiable parameters has measure zero.

The causal interpretation of BAPs is the following. A directed edge from X to Y represents a direct causal effect of X on Y. A bidirected edge between X and Y represents a hidden variable which is a cause of both X and Y, see also Figure 1. In practice, one is often interested in predicting the effect of an intervention at X_j on another variable X_i. This is called the *total causal effect* of X_j on X_i and can be defined as $E_{ij} = \frac{\partial}{\partial x}\mathbb{E}[X_i \mid do(X_j = x)]$, where the $do(X_j = x)$ means replacing the respective equation in (2) with $X_j = x$ (Pearl, 2000). For linear Gaussian path diagrams such as in (1) or (2), this is a constant quantity given by

$$E_{ij} = \left((I - B)^{-1}\right)_{ij}.$$ (4)

2.3. Penalized maximum likelihood

Consider a BAP G. A first objective is to estimate the parameters θ_G from n i.i.d. samples of model (2), denoted by $\{x_i^{(s)}\}$ ($i = 1, \ldots, d$ and $s = 1, \ldots, n$). This can be done by maximum likelihood estimation using the RICF method of Drton et al. (2009). Given the Gaussianity assumption (A1) and the covariance formula (3), one can express the log-likelihood for some given parameters θ_G and the sample covariance matrix S as:

$$l(\theta_G; S) = -\frac{n}{2}\left(\log|2\pi\Sigma_G| + \frac{n-1}{n}\text{tr}(\Sigma_G^{-1}S)\right),$$ (5)

where $\Sigma_G = \phi(\theta_G)$ is the covariance matrix implied by parameters θ_G, see for example Mardia et al. (1979, (4.1.9)). However, due to the structural constraints on B and Ω it is not straightforward to maximize this for θ_G. RICF is an iterative method to do so, yielding the maximum likelihood estimate:

$$\hat{\theta}_G = \underset{\theta_G \in \Theta_G}{\arg\max}\, l(S; \theta_G).$$ (6)

We now extend this to the scenario where the graph G is also unknown, using a regularized likelihood score with a BIC-like penalty term that increases with the number of edges. Concretely, we use the following score for a given BAP G:

$$s(G) := \frac{1}{n}\left(\max_{\theta_G \in \Theta_G} l(S; \theta_G) - (\#\{\text{nodes}\} + \#\{\text{edges}\})\log n\right).$$ (7)

We have scaled the log-likelihood and penalty with $1/n$ so that the score is expected to be $O(1)$ as n increases. Compared with the usual BIC penalty, we chose our penalty to be twice as large, since this led to better performance in simulations studies[5]. The number of nodes is typically fixed, so does not

---

[5]In practice, one could also treat the penalty coefficient as a hyperparameter and choose it via cross-validation.

matter for comparing graphs over the same vertex set. We included it to make explicit the penalization of the model parameters (which correspond to nodes and edges).

In our search for the true causal graph G, we assume that if $\Sigma \in \bar{\mathcal{S}}_H$ for any other graph H, then $\bar{\mathcal{S}}_H \supseteq \bar{\mathcal{S}}_G$, that is H represents a strict supermodel of G. This rules out the possibility that 'by chance' we land on a distribution contained in a submodel, and is a minimal requirement for causal learning. The set of matrices that violate the requirement has measure zero within the model $\bar{\mathcal{S}}_G$ (assuming entries in B and Ω are generated according to a positive joint density with respect to the Lebesgue measure)[6]. This requirement is analogous to the faithfulness assumption of Spirtes et al. (1993), though faithfulness applies separately to individual conditional independence constraints rather than to the entire model.

2.4. Equivalence properties

There is an important issue when doing structure learning with graphical models: typically the maximizers of (7) will not be unique. This is a fundamental problem for most model classes and a consequence of the model being underdetermined. In general, there are sets of graphs that are statistically indistinguishable (in the sense that they can all parametrize the same joint distributions over the variables). These graphs are called *distributionally equivalent*. For nonparametric DAG models (without non-linearity or non-Gaussianity constraints), for example, the distributional equivalence classes are characterized by conditional independencies and are called Markov equivalence classes. For BAPs, distributional equivalence is not completely characterized yet (see Spirtes et al. (1998) or Williams (2012) for a discussion of the linear Gaussian ADMG case), but we present some necessary and some sufficient conditions, that can be used to simplify structure search in practice. Let us first make precise the different notions of model equivalence.

Definition 2. *Two BAPs G_1, G_2 over a set of nodes V are Markov equivalent if they imply the same m-separation relationships.*

This essentially means they imply the same conditional independencies, and the definition coincides with the classical notion of Markov equivalence when G_1 and G_2 are both DAGs. The following notion of distributional equivalence is stronger.

Definition 3. *Two BAPs G_1, G_2 are distributionally equivalent if $\bar{\mathcal{S}}_{G_1} = \bar{\mathcal{S}}_{G_2}$.*

We now present some sufficient and some necessary conditions for distributional equivalence in BAP models. Note that the Gaussianity assumption (A1) is not required for these to hold.

---

[6]This follows because the models are parametrically defined algebraic varieties, which are therefore irreducible. Any sub-variety of $\bar{\mathcal{S}}_G$, such as that achieved by intersecting with another model is either equal to $\bar{\mathcal{S}}_G$ or has strictly smaller dimension. See, for example, Cox et al. (2007).

2.4.1. Necessary conditions

Spirtes et al. (1998) showed the following global Markov property for general linear path diagrams: if there are nodes $a, b \in V$ and a possibly empty set $C \subset V$ such that $a \perp\!\!\!\perp_m b \mid C$, then the partial correlation of X_a and X_b given X_C is zero. In addition, if such an m-separation does not hold then the partial correlation is non-zero for almost all distributions. As a direct consequence, we get the following first result:

Lemma 1. *If two BAPs G_1, G_2 do not share the same m-separations, they are not distributionally equivalent.*

Unlike for DAGs, the converse is not true, as the counterexample in Figure 2 shows. For DAGs it is trivial to show that having the same skeleton is necessary for Markov equivalence, since a missing edge between two nodes means they can be d-separated, and thus a conditional independency would have to be present in the corresponding distribution. For BAPs a missing edge does not necessarily result in an m-separation, as the counterexample in Figure 2 shows. However, the following result will allow us to improve the necessary condition of same m-separations for BAPs to the same as for DAGs.

Theorem 1. *Let G and G' be distributionally equivalent BAPs on vertices V. Then, for any subset $W \subseteq V$, the induced subgraphs G_W and G'_W are also distributionally equivalent.*

If we in particular look at the induced subgraphs of size two and three we obtain the following necessary conditions for distributional equivalence.

Corollary 1. *Let G and G' be distributionally equivalent BAPs. Then they have the same skeleton and v-structures.*

Since m-separations are not fully determined by the skeleton and the v-structures of a graph, it is also worthwhile to look at larger subgraphs. This leads, for example, to the following result: if two graphs are distributionally equivalent and a particular path is a so-called discriminating path in both graphs, then the discriminated triple will be a collider in both or in neither (see Ali et al., 2009, Section 3.4).

The criteria given above are not complete, in the sense that there exist BAPs that are not distributionally equivalent and yet this cannot be proven by applying these results. For example, the BAPs in Figure 3 are not distributionally equivalent, which can be shown using the results of Shpitser et al. (2014). However, they both have no m-separations. A complete characterization remains an open problem.

2.4.2. Sufficient conditions

To prove sufficient conditions, we first give a characterization of the equivalence class in terms of treks (collider-free paths) using Wright's path tracing formula (Wright, 1960). Wright's formula expresses the covariance between any
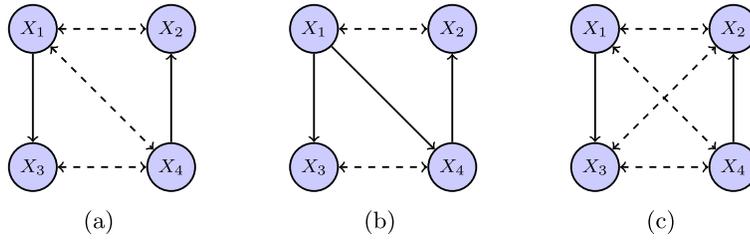
FIG 2. *The two BAPs in (a) and (b) share the same skeleton and v-structures, but in (a) there are no m-separations, whereas in (b) we have $X_2 \perp\!\!\!\perp_m X_3 \mid \{X_1, X_4\}$. BAPs (a) and (c) share the same m-separations (none) but are not distributionally equivalent since they have different skeletons (using Corollary 1).*
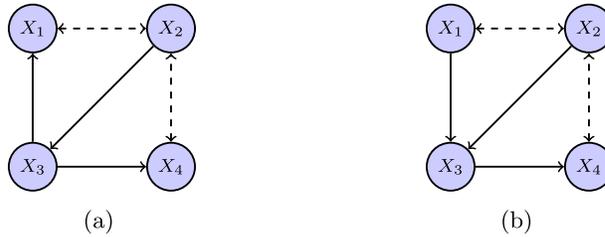

FIG 3. *The two BAPs in (a) and (b) differ only in the direction of the X_1, X_3 edge; both have no m-separations, and every induced subgraph leads to models which are distributionally equivalent. However, by using the results of Shpitser et al. (2014) one can show that these models are not distributionally equivalent.*

two variables in a path diagram as the sum-product over the edge labels of the treks between those variables, as long as all variables are normalized to variance 1. A precise statement as well as a proof of a more general version of Wright's formula can be found in the Appendix (Theorems 3 and 4). As an example, consider the BAP in Figure 1b. There are two treks between X_2 and X_4: $X_2 \to X_3 \to X_4$ and $X_2 \leftrightarrow X_4$. Hence $\text{cov}(X_2, X_4) = B_{32}B_{43} + \Omega_{24}$, assuming normalized parameters. Similarly we have $\text{cov}(X_1, X_4) = B_{21}B_{32}B_{43}$.

As a consequence of Wright's formula, we can show that having the same skeleton and collider triples is sufficient for two acyclic path diagrams to be distributionally equivalent:

Theorem 2. *Let G_1, G_2 be two acyclic path diagrams that have the same skeleton and collider triples. Then G_1 and G_2 are distributionally equivalent.*

Considering Figure 3b, for example, this result shows that if we replace the $X_1 \leftrightarrow X_2$ edge with $X_1 \to X_2$, the resulting graph is distributionally equivalent to the original.

For DAGs, it is known that the weaker condition of having the same skeleton and v-structures is sufficient for being Markov equivalent. For BAPs this is not true, as the counterexample in Figure 2 (together with Lemma 1) shows.

We therefore have that the distributional equivalence class of a BAP G:

- is contained in the set of BAPs with the same skeleton and v-structures as G and
- contains the set of BAPs with the same skeleton and collider triples as G.

We know that the first relation is strict by the counterexample mentioned above and have strong evidence that the second relation is strict as well (Nowzohour, 2015, Appendix B)[7].

3. Greedy search

We aim to find the maximizer of (7) over all graphs over the node set $V = \{1, \ldots, d\}$. Since exhaustive search is infeasible, we use greedy hill-climbing. Starting from some graph G^0, this method obtains increasingly better estimates by exploring the local neighborhood of the current graph. At the end of each exploration, the highest-scoring graph is selected as the next estimate. This approach is also called greedy search and is often used for combinatorial optimization problems. Greedy search converges to a local optimum, although typically not the global one. To alleviate this we repeat it multiple times with different (random) starting points.

We use the following neighborhood relation. A BAP G' is in the *local neighborhood* of G if it differs by exactly one edge, that is, the number of edges differs by at most one, and one of the following holds:

1. $G \subset G'$ (edge addition),
2. $G' \subset G$ (edge deletion), or
3. G and G' have the same skeleton (edge change).

If we only admit the first condition, the procedure is called *forward search*, and it is usually started with the empty graph. Instead of searching through the complete local neighborhood at each step (which can become prohibitive for large graphs), we can also select a random subset of neighbors and only consider those.

In Sections 3.1 and 3.2 we describe some adaptations of this general scheme, that are specific to the problem of BAP learning. In Section 3.3 we describe our greedy equivalence class algorithm.

3.1. Score decomposition

Greedy search becomes much more efficient when the score separates over the nodes or parts of the nodes. For DAGs, for example, the log-likelihood can be written as a sum of components, each of which only depends on one node and its parents. Hence, when considering a neighbor of some given DAG, one

---

[7]These empirical results suggest all 3-node full BAPs to be distributionally equivalent, which would mean there are distributionally equivalent BAPs with different collider triples, implying the strictness of the second inclusion relation above.

only needs to update the components affected by the respective edge change. A similar property holds for BAPs. Here, however, the components are not the nodes themselves, but rather the connected components of the bidirected part of the graph (that is, the partition of V into sets of vertices that are reachable from each other by only traversing the bidirected edges). For example, in Figure 1b the bidirected connected components (sometimes also called districts) are $\{X_1\}, \{X_2, X_4\}, \{X_3\}$. This decomposition property is known (Tian, 2005; Richardson, 2009), but for completeness we give a derivation in appendix B. We write out the special case of the Gaussian parametrization below.

Let us write $p_G^{\mathbf{X}}$ for the joint density of $\mathbf{X}$ under the model (2), and $p_G^{\boldsymbol{\epsilon}}$ for the corresponding joint density of $\boldsymbol{\epsilon}$. Let $C_1, \ldots, C_K$ be the connected components of the bidirected part of G. We separate the model G into submodels $G_1, \ldots, G_K$ of the full SEM (2), where each G_k consists only of nodes in $V_k = C_k \cup \mathrm{pa}(C_k)$ and without any edges between nodes in $\mathrm{pa}(C_k) \setminus C_k$. Then, as we show in appendix B, the log-likelihood of the model with joint density $p_G^{\mathbf{X}}$ given data $\mathcal{D} = \{x_i^{(s)}\}$ (with $1 \leq i \leq d$ and $1 \leq s \leq n$) can be written as:

$$
\begin{aligned}
l(p_G^{\mathbf{X}}; \mathcal{D}) &= \sum_{s=1}^{n} \log p_G^{\mathbf{X}}(x_1^{(s)}, \ldots, x_p^{(s)}) \\
&= \sum_k \left(l(p_{G_k}^{\mathbf{X}}; \{x_i^{(s)}\}_{i \in V_k}^{s=1,\ldots,n}) - \sum_{j \in \mathrm{pa}(C_k) \setminus C_k} l(p_{G_k}^{\mathbf{X}}; \{x_j^{(s)}\}^{s=1,\ldots,n}) \right),
\end{aligned}
$$

where $l(p_{G_k}^{\mathbf{X}}; \{x_j^{(s)}\}^{s=1,\ldots,n})$ refers to the likelihood of the X_j-marginal of $p_{G_k}^{\mathbf{X}}$. For our Gaussian parametrization, using (5), this becomes

$$
\begin{aligned}
l(\Sigma_{G_1}, \ldots, \Sigma_{G_K}; S) = \\
-\frac{n}{2} \sum_k \Bigg(|C_k| \log 2\pi + \log \frac{|\Sigma_{G_k}|}{\prod_{j \in \mathrm{pa}(C_k) \setminus C_k} \sigma_{kj}^2} + \\
\frac{n-1}{n} \mathrm{tr}\big(\Sigma_{G_k}^{-1} S_{G_k} - |\mathrm{pa}(C_k) \setminus C_k|\big) \Bigg),
\end{aligned}
$$

where S_{G_k} is the restriction of S to the rows and columns corresponding to C_k, and σ_{kj}^2 is the diagonal entry of Σ_{G_k} corresponding to parent node j. Note that now the log-likelihood depends on $\{x_i^{(s)}\}$ and $p_G^{\mathbf{X}}$ only via S and $\Sigma_{G_1}, \ldots, \Sigma_{G_K}$. Furthermore, the log-likelihood is now a sum of contributions from the submodels G_k. This means we only need to re-compute the likelihood of the submodels that are affected by an edge change when scoring the local neighborhood. In practice, we also cache the submodel scores, that is, we assign each encountered submodel a unique hash and store the respective scores, so they can be re-used.
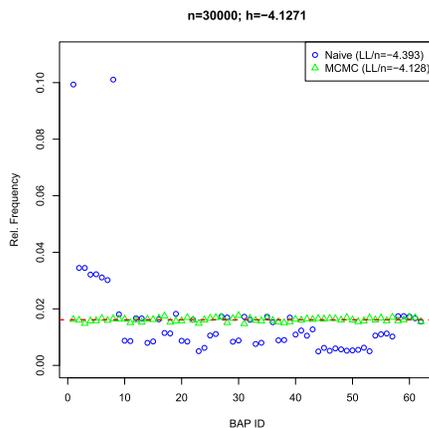
FIG 4. *Relative frequencies of the 62 3-node BAPs when sampled 30000 times with the "naive" (triangular matrix sampling) and the MCMC method.*

3.2. Uniformly random restarts

To restart the greedy search we need random starting points (BAPs), and it seems desirable to sample them *uniformly* at random[8]. Just like for DAGs, it is not straightforward to achieve this. What is often done in practice is uniform sampling of triangular (adjacency) matrices and subsequent uniform permutation of the nodes. However, this does not result in uniformly distributed graphs, since for some triangular matrices many permutations yield the same graph (the empty graph is an extreme example). The consequence is a shift of weight to more symmetric graphs, that are invariant under some permutations of their adjacency matrices. A simple example with BAPs for $d = 3$ is shown in Figure 4. One way around this is to design a random process with graphs as states and a uniform limiting distribution. A corresponding Markov Chain Monte Carlo (MCMC) approach is described for example in Melançon et al. (2001) for the case of DAGs. See also Kuipers and Moffa (2015) for an overview of different sampling schemes.

We adapted the MCMC algorithm for BAPs as described below.

Algorithm 1. *Let $G_k = (V, E_D, E_B)$ be the BAP of the current MCMC iteration. Let $(i, j) \in V \times V \setminus \{(i, i) \mid i \in V\}$ be a position sampled uniformly at random and let $\sigma \in Bernoulli(0.5)$ be a single Bernoulli draw. We then form $G_{k+1} = (V, E'_D, E'_B)$ by applying the following rules.*

1. *If there is an edge at (i, j) (i.e. if $(i, j) \in E_D$ or $(j, i) \in E_D$ or $(i, j) \in E_B$), and*

---

[8] Another motivation for uniform BAP generation is generating ground truths for simulations.

(a) *if $\sigma = 0$: remove the edge (i.e. $E'_D = E_D \setminus \{(i,j),(j,i)\}$, $E'_B = E_B \setminus \{(i,j),(j,i)\}$).*

(b) *if $\sigma = 1$: do nothing.*

2. *If there is no edge at (i,j) (i.e. if $(i,j) \notin E_D$ and $(j,i) \notin E_D$ and $(i,j) \notin E_B$), and*

   (a) *if $\sigma = 0$: add $i \rightarrow j$ (i.e. $E'_D = E_D \cup \{(i,j)\}$, $E'_B = E_B$) as long as this does not create a directed cycle, otherwise do nothing;*

   (b) *if $\sigma = 1$: add $i \leftrightarrow j$ (i.e. $E'_D = E_D$, $E'_B = E_B \cup \{(i,j),(j,i)\}$).*

It is easy to check that the resulting transition matrix is irreducible and symmetric (see Appendix C) and hence the Markov chain has a (unique) uniform stationary distribution. Thus, starting from any graph, after an initial burn-in period, the distribution of the visited states will be approximately uniform over the set of all BAPs. In practice, we start the process from the empty graph and sample after taking $O(d^4)$ steps (c.f. Kuipers and Moffa (2015)).

It is straightforward to adapt this sampling scheme to a number of constraints, for example uniform sampling over all BAPs with a given maximal in-degree.

3.3. Greedy equivalence class construction

We propose the following recursive algorithm to greedily estimate the distributional equivalence class $EC(G)$ of a given BAP G with score ζ. We start by populating the *empirical equivalence class* $\widehat{EC}(G)$ with graphs that have the same skeleton and collider triples as G, since these are guaranteed to be equivalent by Theorem 2. This is a significant computational shortcut, since these graphs do not have to be found greedily anymore. Then, starting once from each of the graphs in $\widehat{EC}(G)$ found above, at each recursion level we search all edge-change neighbors of the current BAP for BAPs that have a score within ϵ of ζ (edge additions or deletions would result in non-equivalent graphs by Corollary 1). For each such BAP, we start a new recursive search until a maximum depth of $d(d-1)/2$ (corresponding to the maximum number of possible edges) is reached, and always comparing against the original score ζ. Already visited states are stored and ignored. Finally, all found graphs are added to $\widehat{EC}(G)$. The main tuning parameter here is ϵ, essentially specifying the threshold for numerical error, as well as statistically indistinguishable graphs. This results in conservative estimates for total causal effects using the methods discussed in Section 4.1 by also including neighboring equivalence classes, that are statistically indistiguishable from the given data.

3.4. Implementation

Our implementation is done in the statistical computing language R (R Core Team, 2015), and the code is available as an R package on github (Nowzohour,

[2017](). We make heavy use of the RICF implementation `fitAncestralGraph`[9] in the `ggm` package (Marchetti et al., [2015]). We noted that there are sometimes convergence issues, so we adapted the implementation of RICF to include a maximal iteration limit (which we set to 10 by default).

4. Empirical results

In this section we present some empirical results to show the effectiveness of our method. First, we consider a simulation setting where we can compare against the ground truth. Then we turn to a well known genomic data set, where the ground truth is unknown, but the likelihood of the fitted models can be compared against other methods.

4.1. Causal effects discovery on simulated data

To validate our method, we randomly generate ground truths, simulate data from them, and try to recover the true total causal effects from the generated datasets. This procedure is repeated $N = 100$ times and the results are averaged. We now discuss each step in more detail.

Randomly generate a BAP G.
We do this *uniformly* at random (for a fixed model size $d = 10$ and maximal in-degree $\alpha = 2$). The sampling procedure is described in Section [3.2].

Randomly generate parameters θ_G.
We sample the directed edge labels in B independently from a standard Normal distribution. We do the same for the bidirected edge labels in Ω, and set the error variances (diagonal entries of Ω) to the respective row-sums of absolute values plus an independently sampled $\chi^2(1)$ value[10].

Simulate data $\{x_i^{(s)}\}$ from θ_G.
This is straightforward, since we just need to sample from a multivariate Normal distribution with mean $\mathbf{0}$ and covariance $\phi(\theta_G)$. We use the function `rmvnorm()` from the package `mvtnorm` (Genz et al., [2014]).

Find an estimate $\hat{G}$ from $\{x_i^{(s)}\}$.
We use greedy search with $R = 100$ uniformly random restarts (as outlined in Section [3]), as well as one greedy forward search starting from the empty model.

Compare G and $\hat{G}$.
A direct comparison of the graphs does not make sense since they could be different but in the same equivalence class. We therefore estimate the equivalence

---

[9]Despite the function name the implementation is not restricted to ancestral graphs.

[10]By Gershgorin's circle theorem, this is guaranteed to result in a positive definite matrix. To increase stability, we also repeat the sampling of Ω if its minimal eigenvalue is less then 10^{-6}.

classes of both G and $\hat{G}$ using the greedy approach described in Section 3.3 with $\epsilon = 10^{-10}$ to get $\widehat{EC}(G)$ and $\widehat{EC}(\hat{G})$.

Since the estimated equivalence classes are empirical, it is not straightforward to compare them. For one, they might be intersecting, but not equal (if the recursion level was set too low and they were started from different graphs for example). More relevantly, they might be entirely different, but still agree in large areas of the graph. We therefore chose to evaluate not the graph structure but the identifiability of causal effects. Often this is also more relevant in practice. Maathuis et al. (2009) developed a method (which they called IDA) to find identifiable causal effects in a multiset of DAGs. We apply the same idea in our setting. Specifically, this means we estimate the causal effects matrix $\hat{E}$ for each graph $G' \in \widehat{EC}(G)$ (using the estimated parameters $\hat{\theta}_{G'} = (\hat{B}', \hat{\Omega}')$ and (4)). We then take absolute values and take the entry-wise minima over all $\hat{E}$ to obtain $\hat{E}_G^{min}$, the minimal absolute causal effects matrix (if an entry E_{ij} is nonzero, there is a nonzero causal effect from X_i to X_j for every graph in the equivalence class). We do the same for $\hat{G}$ to get $\hat{E}_{\hat{G}}^{min}$.

What is left is to compare the minimal absolute causal effects matrix $\hat{E}_G^{min}$ of the ground truth to the minimal absolute causal effects matrix $\hat{E}_{\hat{G}}^{min}$ of the estimate. Thus, our target set consists of all pairs (i, j), such that $(\hat{E}_G^{min})_{ij} > 0$. We score the pairs according to our estimated $\hat{E}_{\hat{G}}^{min}$ values, and we report the area under the ROC curve (AUC, see Hanley and McNeil (1983)). The AUC ranges from 0 to 1, with 1 meaning perfect classification and 0.5 being equivalent to random guessing[11]. In our case, we have a separate ROC curve for each graph. The points on the curve correspond to the thresholding on the estimated absolute value of the causal effects; the k-th point shows the situation when we classify the largest $k - 1$ values as causal, and the rest as non-causal.

The results for 100 simulations can be seen in Figure 5; the average AUC is 0.75. While this suggests that perfect graph discovery is usually not achieved, causal effects can be identified to some extent. We also note that our simulation setting is challenging, in the sense that non-zero edge weights can be arbitrarily close to zero. The computations took 2.5 hours on an AMD Opteron 6174 processor using 20 cores.

4.2. Genomic data

We also applied our method to a well-known genomics data set (Sachs et al., 2005), where the expression of 11 proteins in human T-cells was measured under 14 different experimental conditions (the sample size varies between 707 and 927). There are likely hidden confounders, which makes this setting suitable for hidden variable models. However, it is questionable whether the bow-freeness, linearity, and Gaussianity assumptions hold to a reasonable approximation (in

---

[11]Some care has to be taken because of the fact that the cases $(\hat{E}_G^{min})_{ij} > 0$ and $(\hat{E}_G^{min})_{ji} > 0$ exclude each other, but we took this into account when computing the false positive rate.
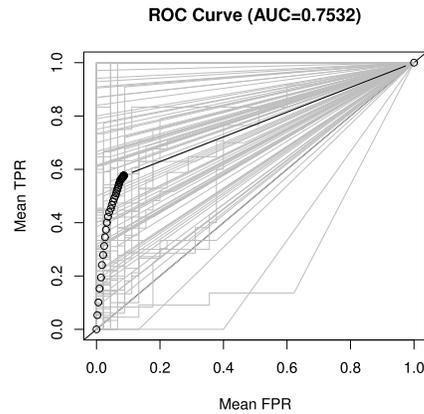
ROC Curve (AUC=0.7532)

FIG 5. *ROC curves for causal effect discovery for $N = 100$ simulation runs of BAPs with $d = 10$ nodes and a maximal in-degree of $\alpha = 2$. Sample size was $n = 1000$, greedy search was repeated $R = 100$ times at uniformly random starting points. The average area under the ROC curves (AUC) is 0.75. The thick curve is the point-wise average of the individual ROC curves.*

fact the data seem not to be multivariate normal). Furthermore, there does not exist a ground truth network (although some individual links between pairs of proteins are reported as known in the original paper). So we abstain from comparing a "best" network with reported links in literature, but instead use this as an example for comparing highest-scoring BAPs and DAGs.

To do this, we first log-transform all variables since they are heavily skewed. We then run two sets of greedy searches for each of the 14 data sets: one with BAPs and one with DAGs. We use 100 random restarts in both cases. The results can be seen in Figures 6 and 7. The computations took 4 hours for the BAP models and 1.5 hours for the DAG models on an AMD Opteron 6174 processor using 20 cores.

Note that while the BAPs and DAGs look very similar in many cases, the BAPs are more conservative in identifying causal effects. Eg for dataset 4 there is a v-structure at `pip3` (with `pip2` and plcg) in both the highest-scoring BAP and DAG. However, by Theorem 2, this part of the BAP is equivalent to versions with different edge directions (as long as the collider is preserved). This is not the case for the DAG. Hence, in the DAG model these edges are identifiable, but this identifiability disappears in the presence of potential hidden confounders in BAPs. This exemplifies the more conservative nature of BAP models. Another example is the v-structure at `pakts473` (with `pka` and `pkc`) in dataset 8.

5. Conclusions

We have presented a structure learning method for BAPs, which can be viewed as a generalization of Gaussian linear DAG models that allow for certain latent variables. Our method is computationally feasible and the first of its kind. The
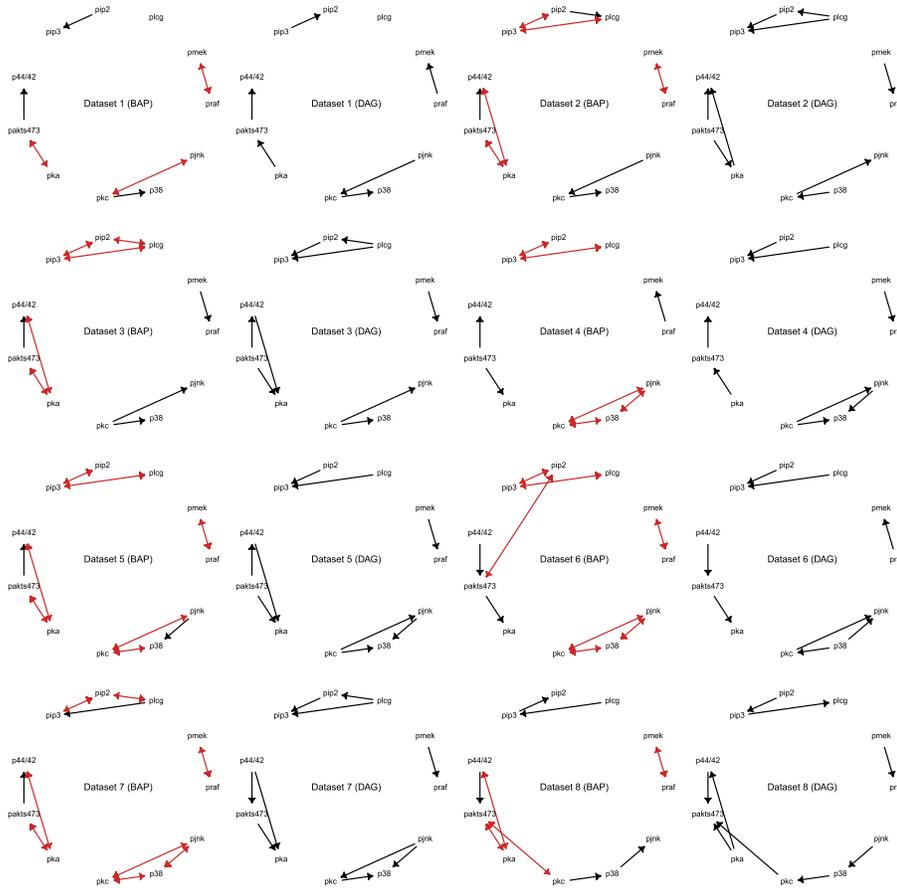
FIG 6. *Highest-scoring BAPs and DAGs found by greedy search for 8 of the 14 genomic datasets in Sachs et al. (2005) (continued in Figure 7). For simplicity only one highest-scoring graph is shown per example while equivalent and equally high-scoring graphs are omitted. Note that the equivalence classes in the corresponding BAPs and DAGs are similar but some v-structures lead to identifiablity in DAGs but not in BAPs.*

results on simulated data are promising, keeping in mind that structure learning and inferring causal effects are difficult, even for the easier case with DAGs. The main sources of errors (given the model assumptions are fulfilled) are sampling variability, finding a local optimum only, and not knowing the equivalence classes. Local optima are a general weakness of many structure learning methods in Bayesian networks since this problem is NP-hard in general (Chickering, 1996). In our simulations, overestimating the equivalence class leads to too few causal effects, while the opposite happens if we underestimate it. On the other hand, our approach of greedily approximating the empirical equivalence class is building on the idea that some models are statistically indistinguishable, due to limited sample size and estimation error. Therefore, our approach has the
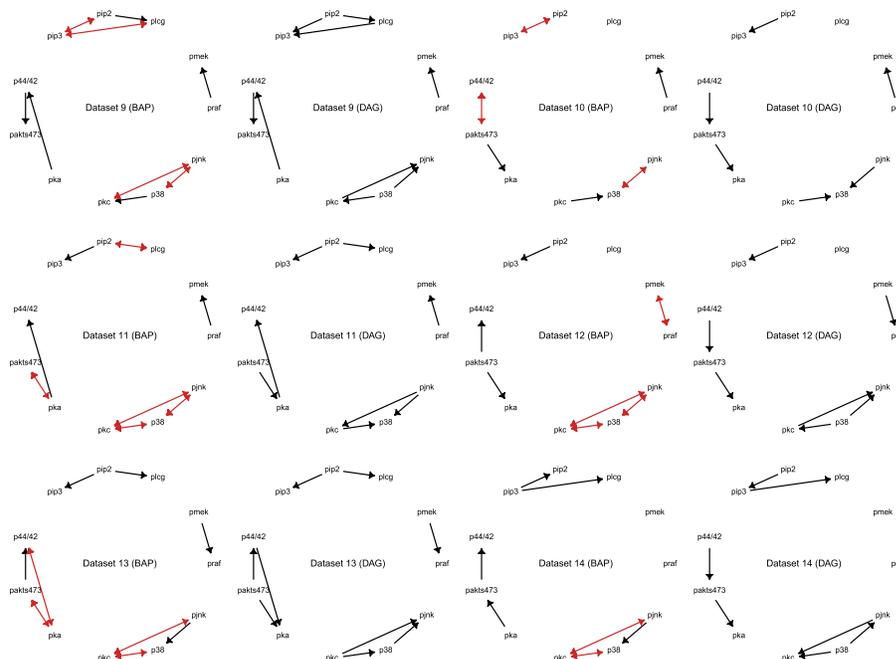
FIG 7. *Highest-scoring BAPs and DAGs found by greedy search for the remaining 6 genomic datasets in Sachs et al. (2005) (continuation of Figure 6). Dataset Names: 1:* cd3cd28, *2:* cd3cd28icam2+aktinhib, *3:* cd3cd28icam2+g0076, *4:* cd3cd28icam2+psit, *5:* cd3cd28icam2+u0126, *6:* cd3cd28icam2+ly, *7:* cd3cd28icam2, *8:* cd3cd28+aktinhib, *9:* cd3cd28+g0076, *10:* cd3cd28+psitect, *11:* cd3cd28+u0126, *12:* cd3cd28+ly, *13:* pma, *14:* b2camp

advantage that it can include neighboring equivalence classes, that score almost as well, which is desirable from a statistical point of view. Our theoretical results about model equivalence go some way towards characterizing the distributional equivalence classes in BAP models and allow us to efficiently approximate them empirically.

In many applications, not all relevant variables are observed, calling for hidden variable models. While there have been structure learning methods for general hidden variable models for many years (FCI, RFCI, FCI+, see Spirtes et al. (1993); Colombo et al. (2012); Claassen et al. (2013)), causal inference based on these models is very conservative (Malinsky and Spirtes, 2017). BAP models are restricted hidden variable models, where the restriction comes from the bow-freeness constaint. As such, they form an interesting middle ground between general hidden variable models and models that do not allow any hidden variables. In particular, the bow-freeness constraint leads to improved identifiability of causal effects when compared to general hidden variable models, while being more conservative than models without hidden variables. This makes our structure learning algorithm for BAPs a useful addition to existing structure

learning methods. Structure learning for a different type of restricted hidden variable models is considered in Frot et al. (2017), and it will be interesting to compare our results with this method.

Appendix A: Distributional equivalence

A.1. Necessary conditions

The following Lemma shows that the point $(B, \Omega) = (0, I)$ is non-singular for the map $\phi : \Theta_G \to \mathcal{S}_G$ for any BAP G. The result also appears in Brito and Pearl (2002) and Drton et al. (2011).

Lemma 2. *Let G be a BAP with parameters (B, Ω), and let ϕ be as in (3). Then $\phi^{-1}(I) = \{(0, I)\}$; that is, the parameters are uniquely identifiable at $\Sigma = I$ (or indeed at any diagonal Σ).*

Proof. We proceed by induction on d, the number of vertices in G. If $d = 1$ then the result is trivial since $B = 0$.

Otherwise assume without loss of generality that the last vertex d has no children. The result holds for the subgraph of the remaining vertices by the induction hypothesis and the fact that the distribution of $\mathbf{X}$ is defined recursively. We know that Σ is of the form

$$\Sigma = (I_d - B)^{-1} \Omega (I_d - B)^{-T},$$

and we may deduce by the induction hypothesis that the first $d - 1$ rows of B are zero, and the upper $(d-1) \times (d-1)$-sub matrix of Ω is the identity matrix. Hence

$$\Omega = \begin{pmatrix} I_{d-1} & \boldsymbol{\omega} \\ \boldsymbol{\omega}^T & \omega_{dd} \end{pmatrix},$$

where $\boldsymbol{\omega}^T = (\omega_{1d}, \ldots, \omega_{d-1,d})$, and

$$(I_d - B)^{-1} = \begin{pmatrix} I_{d-1} & 0 \\ -\boldsymbol{\beta}^T & 1 \end{pmatrix}^{-1} = \begin{pmatrix} I_{d-1} & 0 \\ \boldsymbol{\beta}^T & 1 \end{pmatrix},$$

where $\boldsymbol{\beta}^T = (\beta_{d1}, \ldots, \beta_{d,d-1})$. Hence

$$\begin{aligned} \Sigma &= (I - B)^{-1} \Omega (I - B)^{-T} \\ &= \begin{pmatrix} I_{d-1} & 0 \\ \boldsymbol{\beta}^T & 1 \end{pmatrix} \begin{pmatrix} I_{d-1} & \boldsymbol{\omega} \\ \boldsymbol{\omega}^T & \omega_{dd} \end{pmatrix} \begin{pmatrix} I_{d-1} & \boldsymbol{\beta} \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} I_{d-1} & 0 \\ \boldsymbol{\beta}^T & 1 \end{pmatrix} \begin{pmatrix} I_{d-1} & \boldsymbol{\beta} + \boldsymbol{\omega} \\ \boldsymbol{\omega}^T & \boldsymbol{\beta}^T \boldsymbol{\omega} + \omega_{dd} \end{pmatrix} \end{aligned}$$

but note that $\boldsymbol{\beta}^T \boldsymbol{\omega} = 0$ by the bow-free assumption, so we get

$$\Sigma = \begin{pmatrix} I_{d-1} & \boldsymbol{\beta} + \boldsymbol{\omega} \\ \boldsymbol{\beta}^T + \boldsymbol{\omega}^T & \|\boldsymbol{\beta}\|^2 + \omega_{dd} \end{pmatrix},$$

and hence $\boldsymbol{\beta} + \boldsymbol{\omega} = \mathbf{0}$. Now note that for each j, either $\beta_{dj} = 0$ or $\omega_{jd} = 0$ by the bow-free assumption; hence $\boldsymbol{\beta} + \boldsymbol{\omega} = \mathbf{0}$ implies that $\boldsymbol{\beta} = \boldsymbol{\omega} = \mathbf{0}$, leaving

$$\Sigma = \left(\begin{array}{cc} I_{d-1} & \mathbf{0} \\ \mathbf{0}^T & \omega_{dd} \end{array} \right)$$

and hence $\omega_{dd} = 1$. This completes the result. $\qquad\square$

Corollary 2. *Let G be a BAP. For some neighborhood U of the set of covariance matrices containing I, if $\Sigma \in \mathcal{S}_G \cap U$ with $\sigma_{ij} = 0$ (for $i \neq j$), then this implies that $\omega_{ij} = \beta_{ij} = 0$.*

Proof. Since ϕ is nonsingular and differentiable at $\theta_0 = (O, I)$, its partial derivatives are defined and given by $\frac{\partial \phi}{\partial \omega_{ij}}(\theta_0) = 1$ and $\frac{\partial \phi}{\partial \beta ij}(\theta_0) = 1$ (this can be shown via a Taylor expansion for example). Therefore, in a small neighborhood around $\phi(\theta_0)$ we have $\sigma_{ij} = 0$ only if $\omega_{ij} = \omega_{ji} = \beta_{ij} = 0$. $\qquad\square$

Note that Lemma 2 allows a direct proof of the fact that having the same skeleton is necessary for BAPs to be distributionally equivalent by looking at the tangent spaces of the models at $\Sigma = I$ and showing that they are determined by the skeletons of the graphs.

In the proof of Theorem 1 we make use of the language of polynomial varieties (see Cox et al. (2007) for an overview). A variety is a set defined by the zeros of some collection of polynomials (in our case polynomials in the entries of Σ), and all SEM models are varieties.

Let G be a BAP with vertices $V = W \dot\cup \overline{W}$ where $W \cap \overline{W} = \emptyset$. Let $\mathcal{B}_G^W$ be the set of matrices $B \in \mathcal{B}_G$ such that only entries corresponding to directed edges in G between vertices in W have non-zero coefficients. Similarly, let $\mathcal{O}_G^W$ be the set of $\Omega \in \mathcal{O}_G$ such that entries corresponding to edges outside W are zero and diagonal entries outside W are 1.

Define a model $\tilde{\mathcal{S}}_G^W$ as the image of the map ϕ applied to $(\mathcal{B}_G^W, \mathcal{O}_G^W)$. So in other words, we only manipulate parameters in G that correspond to vertices and edges in G_W. The resulting model is canonically isomorphic to $\mathcal{S}_{G_W}$ via a simple projection, since this is the same setup as for the BAP G_W, but with the matrices extended to include independent vertices in $\overline{W} \equiv V \setminus W$.

Let $\mathcal{T}_W$ be the set of covariance matrices Σ on V such that $\Sigma_{\overline{W}\overline{W}} = I$ and $\Sigma_{W\overline{W}} = 0$ (i.e. so that vertices outside W are completely independent).

We will show that looking at the set of covariance matrices in $\mathcal{S}_G$ that are also in $\mathcal{T}_W$ is essentially the same as the set $\tilde{\mathcal{S}}_G^W$. Since the first set is a property of the full model, and the second set is determined by the subgraph G_W, this will be enough to prove Theorem 1.

Proof of Theorem 1. What we need to prove is that $\mathcal{S}_G = \mathcal{S}_G'$ implies $\mathcal{S}_{G_W} = \mathcal{S}_{G'_W}$. Consider again the variety $\mathcal{T}_W$ defined above.

Clearly $\mathcal{S}_G = \mathcal{S}_{G'}$ implies $\mathcal{T}_W \cap \mathcal{S}_G = \mathcal{T}_W \cap \mathcal{S}_{G'}$. We will show that the irreducible component of $\mathcal{T}_W \cap \mathcal{S}_G$ which contains $\Sigma = I$ is the same as $\tilde{\mathcal{S}}_G^W$; since this last quantity is isomorphic to $\mathcal{S}_{G_W}$ we will prove the result.

First, note that $\tilde{\mathcal{S}}_G^W \subseteq \mathcal{T}_W$ and $\tilde{\mathcal{S}}_G^W \subseteq \mathcal{S}_G$, so clearly $\tilde{\mathcal{S}}_G^W \subseteq \mathcal{S}_G \cap \mathcal{T}_W$. In addition, note that by Corollary 2, in a neighborhood of $\Sigma = I$ every element of $\mathcal{T}_W \cap \mathcal{S}_G$ is also contained in $\tilde{\mathcal{S}}_G^W$. It follows that the entire irreducible component

of $\mathcal{T}_W \cap \mathcal{S}_G$ containing $\Sigma = I$ is contained within $\tilde{\mathcal{S}}_G^W$, and therefore that the irreducible component of $\mathcal{T}_W \cap \mathcal{S}_G$ containing $\Sigma = I$ is $\tilde{\mathcal{S}}_G^W$. □

We can now prove the Theorem giving necessary conditions for BAP equivalence.

Proof of Corollary 1. Let us first consider vertex pairs, i.e. $W = \{i, j\}$. By Theorem 1 we have G_W being distributionally equivalent to G'_W. If $G_W \neq G'_W$ we would have $i \perp\!\!\!\perp_m j$ in one of the graphs but not the other, and using Lemma 1 this would lead to a contradiction. Hence $G_W = G'_W$ for any vertex pair, and hence G and G' must have the same skeleton.

Let us now consider vertex triplets $W = \{i, j, k\}$, such that (without loss of generality) there is a v-structure at j in G_W. Then $i \perp\!\!\!\perp_m k$ in G_W and by the same argument as above we must have $i \perp\!\!\!\perp_m k$ also in G'_W. This is only possible if there is a v-structure at j in G'_W. Hence G and G' must have the same v-structures. □

A.2. Sufficient conditions

We first make precise the definition of an important class of paths: *treks*. These are paths that do not contain colliders. We adopt the notation of Foygel et al. (2012). A trek τ from i to j can have one of the following forms:

$$v_l^L \leftarrow \cdots \leftarrow v_0^L \longleftrightarrow v_0^R \rightarrow \cdots \rightarrow v_r^R$$

or

$$v_l^L \leftarrow \cdots \leftarrow v_0 \rightarrow \cdots \rightarrow v_r^R,$$

where $v_l^L = i$ and $v_r^R = j$ and in the second case $v_0 = v_0^L = v_0^R$. Accordingly, we define the *left-hand side* of τ as $\mathrm{Left}(\tau) = v_l^L \leftarrow \cdots \leftarrow v_0^L$ and the *right-hand side* of τ as $\mathrm{Right}(\tau) = v_0^R \rightarrow \cdots \rightarrow v_r^R$. Note that there is nothing inherently directional about a trek other that the (arbitrary) definition which end node is on the left. That is, every trek from i to j is also a trek from j to i just with the left and right sides switched. We denote the lengths of the left- and right-hand sides of a trek τ by $\lambda_L(\tau)$ and $\lambda_R(\tau)$ respectively. If τ does not contain a bidirected edge, we define its *head* to be $H_\tau = v_0$. If the left- and right-hand sides of τ do not intersect (except possibly at H_τ), we call τ *simple*[12]. We define the following sets that will be useful later:

$$\mathcal{D}_G^{ij} = \{\pi \mid \pi \text{ is a directed path from } i \text{ to } j \text{ in } G\},$$
$$\mathcal{T}_G^{ij} = \{\tau \mid \tau \text{ is a trek from } i \text{ to } j \text{ in } G\},$$
$$\mathcal{S}_G^{ij} = \{\tau \mid \tau \text{ is a simple trek from } i \text{ to } j \text{ in } G\}.$$

We will usually drop the subscript if it is clear from context which is the reference graph.

---

[12]Note that each side might well be self-intersecting, if the corresponding graph is cyclic.

We now show some intermediate results that are well-known, but we prove them here for completeness nevertheless. All of these apply more generally to path diagrams (possibly cyclic and with bows).

Lemma 3. *Let $B \in \mathbb{R}^{d \times d}$ such that every eigenvalue λ of B satisfies $|\lambda| < 1$. Then $(I - B)^{-1}$ exists and is equal to $\sum_{s=0}^{\infty} B^s$.*

Proof. First note that $\det(\lambda I - B) = 0$ only if $|\lambda| < 1$, hence $\det(I - B) \neq 0$, and therefore $(I - B)^{-1}$ exists. The eigenvalue condition also implies $\lim_{l \to \infty} B^l = 0$, therefore

$$(I - B) \sum_{s=0}^{\infty} B^s = \lim_{l \to \infty} \sum_{s=0}^{l} (B^s - B^{s+1}) = \lim_{l \to \infty} (I - B^{l+1}) = I,$$

and the result follows. □

Lemma 4. *Let G be a path diagram over d nodes and $B \in \mathcal{B}_G$. Then*

$$(B^l)_{ij} = \sum_{\substack{\pi \in \mathcal{D}^{ji} \\ \lambda(\pi) = l}} \prod_{s \to t \in \pi} B_{ts}.$$

Proof. By induction on l. For $l = 1$ the claim follows from the definition of $\mathcal{B}_G$. Using the inductive hypothesis we get

$$(B^l)_{ij} = (BB^{l-1})_{ij} = \sum_{k=1}^{d} B_{ik}(B^{l-1})_{kj} = \sum_{k=1}^{d} B_{ik} \sum_{\substack{\pi \in \mathcal{D}^{jk} \\ \lambda(\pi) = l-1}} \prod_{s \to t \in \pi} B_{ts},$$

and the claim follows, since every directed path from j to i of length l can be decomposed into a directed path π of length $l - 1$ from j to some node k and the edge $k \to i$. □

Lemma 5. *Let G be an acyclic path diagram over d nodes and $B \in \mathcal{B}_G$. Then $(I - B)^{-1} = I + B + \ldots + B^{d-1}$.*

Proof. Since G is acyclic, there is an ordering of the nodes, such that B is strictly lower triangular and hence all its eigenvalues are zero. Furthermore, the longest directed path in G has length $d - 1$. Therefore the result follows from Lemma 3 and Lemma 4. □

The following theorem is a version of Wright's theorem that applies to non-standardized variables. It does not require a proper parametrization (in the sense that Ω needs to be positive definite). This result is probably known to experts, but we could not find a proof in the literature.

Theorem 3. *Let G be a (possibly cyclic) path diagram over d nodes, $B \in \mathcal{B}_G$, and $\Omega \in \mathbb{R}^{d \times d}$ such that Ω is symmetric (but not necessarily positive definite)*

and $\Omega_{ij} = 0$ if $i \leftrightarrow j$ is not an edge in G. Then the entries of the matrix $\phi = (I - B)^{-1}\Omega(I - B)^{-T}$ are given by

$$\phi_{ij} = \sum_{\substack{\tau \in \mathcal{S}^{ij} \\ \leftrightarrow \in \tau}} \prod_{s \to t \in \tau} B_{ts} \prod_{s \leftrightarrow t \in \tau} \Omega_{st} + \sum_{\substack{\tau \in \mathcal{S}^{ij} \\ \leftrightarrow \notin \tau}} \prod_{s \to t \in \tau} B_{ts} \cdot \phi_{H_\tau H_\tau} \qquad (i \neq j),$$

$$\phi_{ii} = \sum_{\substack{\tau \in \mathcal{T}^{ii} \\ \leftrightarrow \in \tau}} \prod_{s \to t \in \tau} B_{ts} \prod_{s \leftrightarrow t \in \tau} \Omega_{st} + \sum_{\substack{\tau \in \mathcal{T}^{ii} \\ \leftrightarrow \notin \tau}} \prod_{s \to t \in \tau} B_{ts} \cdot \Omega_{H_\tau H_\tau} + \Omega_{ii}.$$

Proof. Let us write

$$c_e(\tau; B, \Omega) = \prod_{s \to t \in \tau} B_{ts} \prod_{s \leftrightarrow t \in \tau} \Omega_{st}$$

as a shorthand for the edge contribution[13] of a trek τ given parameter matrices B and Ω. We write $c(\tau; B, \Omega) = c_e(\tau; B, \Omega) \cdot \Omega_{H_\tau H_\tau}$ for the total contribution of τ (where we define $\Omega_{H_\tau H_\tau}$ to be 1 if τ contains a bidirected edge and therefore $H_\tau = \varnothing$).

Using Lemma 3, we can expand ϕ as $\phi = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} B^k \Omega (B^l)^T$. We now first show the following intermediate result, which interprets the entries of these matrices as contributions of certain treks:

$$(B^k \Omega (B^l)^T)_{ij} = \sum_{\substack{\tau \in \mathcal{T}^{ij} \\ \lambda_L(\tau) = k \\ \lambda_R(\tau) = l}} c(\tau; B, \Omega) + \Omega_{ii} \mathbb{1}\{i = j\}, \qquad (8)$$

for integers $k \geq 0$, $l \geq 0$. To see this, we expand the double matrix product and use Lemma 4 to get

$$(B^k \Omega (B^l)^T)_{ij} = \sum_{a=1}^{d} \sum_{b=1}^{d} (B^k)_{ia} \Omega_{ab} (B^l)_{jb}$$

$$= \sum_{a=1}^{d} \sum_{b=1}^{d} \left(\sum_{\substack{\pi \in \mathcal{D}^{ai} \\ \lambda(\pi) = k}} \prod_{s \to t \in \pi} B_{ts} \right) \Omega_{ab} \left(\sum_{\substack{\pi \in \mathcal{D}^{bj} \\ \lambda(\pi) = l}} \prod_{s \to t \in \pi} B_{ts} \right),$$

and (8) follows since each bracketed expression corresponds to one side of the trek from i to j via a and b (and the diagonal entries of Ω do not correspond to a trek, so they are separate). Now summing over k and l gives the following

$$\phi_{ij} = \sum_{\tau \in \mathcal{T}^{ij}} c(\tau; B, \Omega) + \Omega_{ii} \mathbb{1}\{i = j\}, \qquad (9)$$

which gives the result for the diagonal entries ϕ_{ii}.

---

[13] That is, the contribution depending only on the edge labels and not the diagonal elements of Ω.

For the off-diagonal entries ϕ_{ij}, we can get a simpler expression involving only simple treks and the diagonal entries ϕ_{ii}. Note that every trek τ can be uniquely decomposed into a simple part $\xi(\tau)$ and a (possibly empty) non-simple part $\rho(\tau)$ (we just split at the point, where the right- and the left-hand sides of τ first intersect). Since

$$c(\tau) = \begin{cases} c_e(\xi(\tau)) \cdot \Omega_{H_{\xi(\tau)} H_{\xi(\tau)}} & \text{if } \rho(\tau) = \varnothing \\ c_e(\xi(\tau)) \cdot c(\rho(\tau)) & \text{otherwise,} \end{cases}$$

(dropping the parameter matrices B and Ω in our notation), we can factor out the contributions of the simple parts. Note that if the simple part $\xi(\tau)$ contains a bidirected edge, then $\rho(\tau)$ must be empty and $\Omega_{H_\xi(\tau) H_\xi(\tau)} = 1$. Hence (9) becomes

$$
\phi_{ij} = \sum_{\substack{\tau \in \mathcal{T}^{ij} \\ \leftrightarrow \in \xi(\tau)}} c(\tau) + \sum_{\substack{\tau \in \mathcal{T}^{ij} \\ \leftrightarrow \notin \xi(\tau) \\ \rho(\tau) \neq \varnothing}} c(\tau) + \sum_{\substack{\tau \in \mathcal{T}^{ij} \\ \leftrightarrow \notin \xi(\tau) \\ \rho(\tau) = \varnothing}} c(\tau)
$$

$$
= \sum_{\substack{\tau \in \mathcal{T}^{ij} \\ \leftrightarrow \in \xi(\tau)}} c_e(\xi(\tau)) + \sum_{\substack{\tau \in \mathcal{T}^{ij} \\ \leftrightarrow \notin \xi(\tau) \\ \rho(\tau) \neq \varnothing}} c_e(\xi(\tau)) \cdot c(\rho(\tau)) + \sum_{\substack{\tau \in \mathcal{T}^{ij} \\ \leftrightarrow \notin \xi(\tau) \\ \rho(\tau) = \varnothing}} c_e(\xi(\tau)) \cdot \Omega_{H_{\xi(\tau)} H_{\xi(\tau)}}
$$

$$
= \sum_{\substack{\xi \in \mathcal{S}^{ij} \\ \leftrightarrow \in \xi}} c_e(\xi) + \sum_{\substack{\xi \in \mathcal{S}^{ij} \\ \leftrightarrow \notin \xi}} c_e(\xi) \sum_{\rho \in \mathcal{T}^{H_\xi H_\xi}} c(\rho) + \sum_{\substack{\xi \in \mathcal{S}^{ij} \\ \leftrightarrow \notin \xi}} c_e(\xi) \cdot \Omega_{H_\xi H_\xi}
$$

$$
= \sum_{\substack{\xi \in \mathcal{S}^{ij} \\ \leftrightarrow \in \xi}} c_e(\xi) + \sum_{\substack{\xi \in \mathcal{S}^{ij} \\ \leftrightarrow \notin \xi}} c_e(\xi) \left(\sum_{\rho \in \mathcal{T}^{H_\xi H_\xi}} c(\rho) + \Omega_{H_\xi H_\xi} \right),
$$

and the result follows. □

The following version for standardized parameters is often quoted as Wright's theorem.

Theorem 4. *Let G be a (not necessarily acyclic) path diagram over d nodes, $B \in \mathcal{B}_G$, and $\Omega \in \mathbb{R}^{d \times d}$ such that Ω is symmetric (but not necessarily positive definite) and $\Omega_{ij} = 0$ if $i \leftrightarrow j$ is not an edge in G. Furthermore assume that we have standardized parameters B, Ω such that $(\phi(B, \Omega))_{ii} = 1$ for all i. Then the off-diagonal entries of $\phi(B, \Omega)$ are given by*

$$(\phi(B, \Omega))_{ij} = \sum_{\tau \in \mathcal{S}^{ij}} \prod_{s \to t \in \tau} B_{ts} \prod_{s \leftrightarrow t \in \tau} \Omega_{st}.$$

Proof. This is a direct consequence of Theorem 3. □

We can now prove Theorem 2, which is a consequence of Wright's formula.

Proof of Theorem 2. Let $\theta_{G_1} \in \bar{\Theta}_{G_1}$ and choose $\theta_{G_2} = (B_2, \Omega_2)$ such that their edge labels agree, that is,

$$(B_2)_{ij} = \begin{cases} (B_1)_{ij} & \text{if } i \leftarrow j \in G_1, i \leftarrow j \in G_2, \\ (B_1)_{ji} & \text{if } i \rightarrow j \in G_1, i \leftarrow j \in G_2, \\ (\Omega_1)_{ij} & \text{if } i \leftrightarrow j \in G_1, i \leftarrow j \in G_2, \\ 0 & \text{if } i \leftarrow j \notin G_2, \end{cases}$$

and

$$(\Omega_2)_{ij} = \begin{cases} (B_1)_{ij} & \text{if } i \leftarrow j \in G_1, i \leftrightarrow j \in G_2, \\ (B_1)_{ji} & \text{if } i \rightarrow j \in G_1, i \leftrightarrow j \in G_2, \\ (\Omega_1)_{ij} & \text{if } i \leftrightarrow j \in G_1, i \leftrightarrow j \in G_2, \\ 0 & \text{if } i \leftrightarrow j \notin G_2. \end{cases}$$

This is possible since G_1 and G_2 have the same skeleton: we just assign the edge labels of G_1 to G_2, irrespective of the edge type. The diagonal entries of Ω_2 are still free—we now show that they can be used to enforce

$$(\phi(B_2, \Omega_2))_{ii} = 1 \tag{10}$$

for all i, which defines a linear system for the diagonal entries of Ω_2. Let $\mathbf{d} = \text{diag}(\Omega_2)$ be the vector consisting of the diagonal elements of Ω_2, and write (10) as $M\mathbf{d} + \mathbf{c} = \mathbf{1}$, where M is the coefficient matrix of the linear system, and $\mathbf{c}$ is constant. To show that (10) always has a solution, we need to show that $\det(M) \neq 0$. Without loss of generality, assume that the nodes are topologically ordered according to G_2 (this is possible since G_2 is assumed to be acyclic), that is, there is no directed path from i to j if $i > j$. Then we have $H_\tau < i$ (or $H_\tau = \varnothing$) for all $\tau \in \mathcal{T}^{ii}$, and using the expression for ϕ_{ii} in Theorem 3 we see that M must be lower triangular with diagonal equal to 1. Thus $\det(M) = 1$, and we can enforce (10).

Since G_1 and G_2 share the same collider triples, the sets of simple treks between any two nodes are the same in both graphs: $\mathcal{S}_{G_1}^{ij} = \mathcal{S}_{G_2}^{ij} \; \forall i, j$. Together with Theorem 4 and the fact that the edge labels agree this shows that

$$\phi(\theta_{G_1}) = \phi(\theta_{G_2}). \tag{11}$$

What is left to show is that Ω_2 is a valid covariance matrix, that is, it is positive semi-definite. By (3) and (11) we have that

$$\Omega_2 = (I - B_2)\Sigma_1(I - B_2)^T,$$

where $\Sigma_1 = \phi(\theta_{G_1})$. Since Σ_1 is positive semidefinite, so is Ω_2. $\qquad\square$

Appendix B: Likelihood separation

Since we can write $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(\mathbf{X})$ as a function of $\mathbf{X} = (X_1, \ldots, X_d)$, we have that their densities satisfy

$$p_G^{\mathbf{X}}(X_1, \ldots, X_d) = p_G^{\boldsymbol{\epsilon}}(\epsilon_1(\mathbf{X}), \ldots, \epsilon_d(\mathbf{X})). \tag{12}$$

The joint density for the errors $\boldsymbol{\epsilon}$ can be factorized according to the independence structure implied by Ω. Let us adopt the notation $\{X\}_I := \{X_i\}_{i \in I}$ and $\{\epsilon\}_I := \{\epsilon_i\}_{i \in I}$ for some index set I. Then we have $\{\epsilon\}_{C_k} \perp\!\!\!\perp \{\epsilon\}_{C_l} \ \forall k \neq l$. Furthermore, we implicitly refer to marginalized densities via the arguments, i.e. we write $p_G^{\boldsymbol{\epsilon}}(\{\epsilon\}_{C_1})$ for the marginal density of $\{\epsilon\}_{C_1}$. We can thus write

$$p_G^{\boldsymbol{\epsilon}}(\epsilon_1, \ldots, \epsilon_d) = p_G^{\boldsymbol{\epsilon}}(\{\epsilon\}_{C_1}) \cdots p_G^{\boldsymbol{\epsilon}}(\{\epsilon\}_{C_K}).$$

Hence (12) becomes

$$p_G^{\mathbf{X}}(X_1, \ldots, X_d) = \prod_k p_G^{\boldsymbol{\epsilon}}\left(\left\{X_i - \sum_{j \in \mathrm{pa}(i)} B_{ij} X_j\right\}_{i \in C_k}\right). \tag{13}$$

Each factor depends only on the nodes in the respective component C_k and the parents of that component $\mathrm{pa}(C_k)$. By the same argument the joint density of the submodel G_k is

$$p_{G_k}^{\mathbf{X}}(\{X\}_{V_k}) = p_{G_k}^{\boldsymbol{\epsilon}}(\{\epsilon\}_{C_k}) \prod_{j \in \mathrm{pa}(C_k) \backslash C_k} p_{G_k}^{\boldsymbol{\epsilon}}(\epsilon_j)$$

$$= p_{G_k}^{\boldsymbol{\epsilon}}\left(\left\{X_i - \sum_{j \in \mathrm{pa}(i)} B_{ij} X_j\right\}_{i \in C_k}\right) \prod_{j \in \mathrm{pa}(C_k) \backslash C_k} p_{G_k}^{\boldsymbol{\epsilon}}(X_j).$$

This factorization is symbolic, since the parents $\{X_j\}_{j \in \mathrm{pa}(C_k) \backslash C_k}$ will not be independent in general. This does not matter, however, since these terms cancel when reconstructing the full density $p_G^X(X_1, \ldots, X_d)$ later. The advantage of this symbolic factorization is that we can still *fit* the (wrong) submodel and then use the easier to compute product of marginal densities to reconstruct the full density, rather than doing the same with the actual submodel factorization and the joint density of the component parents.

Note that

$$p_{G_k}^{\boldsymbol{\epsilon}}\left(\left\{X_i - \sum_{j \in \mathrm{pa}(i)} B_{ij} X_j\right\}_{i \in C_k}\right) = p_G^{\boldsymbol{\epsilon}}\left(\left\{X_i - \sum_{j \in \mathrm{pa}(i)} B_{ij} X_j\right\}_{i \in C_k}\right),$$

that is, the conditionals $\{X\}_{C_k} | \{X\}_{\mathrm{pa}(C_k) \backslash C_k}$ are the same in models G and G_k. This is because the structural equations of $\{X\}_{C_k}$ are the same in these models. Note also that $p_{G_k}^{\boldsymbol{\epsilon}}(X_j) = p_{G_k}^{\mathbf{X}}(X_j)$ for all $j \in \mathrm{pa}(C_k) \backslash C_k$ and all k, since $\mathrm{pa}(C_k) \backslash C_k$ are source nodes in model G_k (all edges between them were removed).

Thus we can reconstruct the full joint density (13) from joint densities of the connected component submodels and marginal densities of the parent variables:

$$p_G^{\mathbf{X}}(X_1, \ldots, X_d) = \prod_k p_{G_k}^{\mathbf{X}}(\{X\}_{V_k}) \cdot \left(\prod_{j \in \mathrm{pa}(C_k) \backslash C_k} p_{G_k}^{\mathbf{X}}(X_j)\right)^{-1}.$$

Writing $\mathcal{D}$ for the observed data $\{x_i^s\}$ (with $1 \leq i \leq d$ and $1 \leq s \leq n$), the log-likelihood can then be written as

$$
\begin{aligned}
l(p_G^{\mathbf{X}}; \mathcal{D}) &= \sum_{s=1}^{n} \log p_G^{\mathbf{X}}(x_1^{(s)}, \ldots, x_p^{(s)}) \\
&= \sum_{s=1}^{n} \sum_{k} \left(\log p_{G_k}^{\mathbf{X}}(\{x_i^{(s)}\}_{i \in V_k}) - \sum_{j \in \mathrm{pa}(C_k) \setminus C_k} \log p_{G_k}^{\mathbf{X}}(x_j^{(s)}) \right) \\
&= \sum_{k} \left(l(p_{G_k}^{\mathbf{X}}; \{x_i^{(s)}\}_{i \in V_k}^{s=1,\ldots,n}) - \sum_{j \in \mathrm{pa}(C_k) \setminus C_k} l(p_{G_k}^{\mathbf{X}}; \{x_j^{(s)}\}^{s=1,\ldots,n}) \right),
\end{aligned}
$$

where $l(p_{G_k}^{\mathbf{X}}; \{x_j^{(s)}\}^{s=1,\ldots,n})$ refers to the likelihood of the X_j-marginal of $p_{G_k}^{\mathbf{X}}$.

Appendix C: Symmetry and irreducibility of Markov Chain

We show that the transition matrix of the Markov Chain described in Algorithm 1 is symmetric and irreducible. For two BAPs G, G', let $P(G, G')$ be the probability of a single step transition from G to G'.

Theorem 5. *We have*

1. *Symmetry: $P(G, G') = P(G', G)$.*
2. *Irreducibility: $\exists G_1, \ldots, G_n$ such that*

$$
P(G, G_1) \left(\prod_{i=1}^{n-1} P(G_i, G_{i+1}) \right) P(G_{n-1}, G') > 0.
$$

Proof. Let p be the probability of sampling one position (i, j), i.e. $p = 1/(d(d-1))$. Let us first consider the case where G and G' only differ by one edge addition, i.e. WLOG either

- $G = G' \cup i \to j$.
- $G = G' \cup i \leftrightarrow j$.

In both cases we get $P(G, G') = p/2 = P(G', G)$. In the first case, by multiplying the probabilities along the branches of 1a and 2a in Algorithm 1 respectively, and since G has no cycles. In the second case, by multiplying the probabilities along the branches of 1a and 2b respectively. Hence symmetry holds, and irreducibility is trivially true in this case.

For the general case, note that the transitions described in Algorithm 1 involve either edge additions or deletions, so if G, G' do not differ by only one edge addition, we have $P(G, G') = P(G', G) = 0$. Furthermore, we can always find a collection of graphs $G_1, \ldots, G_n$, such that irreducibility holds, e.g. by successively removing edges from G until the graph is empty and then successively adding edges until we arrive at G'. Then we have $P(G_i, G_{i+1}) = p/2 > 0$ for all $1 \leq i < n$ by the case considered above, and the claim follows. $\square$

References

Ali, R. A., Richardson, T. S., and Spirtes, P. Markov equivalence for ancestral graphs. *Annals of Statistics*, 37(5B):2808–2837, 2009. MR2541448

Brito, C. and Pearl, J. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002. MR1930449

Chickering, D. M. Learning Bayesian networks is NP-complete. In *Learning from Data*, volume 112 of *Lecture Notes in Statistics*, pages 121–130. 1996. MR1473013

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002. MR1991085

Claassen, T., Mooij, J. M., and Heskes, T. Learning sparse causal models is not NP-hard. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 172–181, 2013.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012. MR3014308

Cox, D. A., Little, J., and O'Shea, D. *Ideals, Varieties, and Algorithms*. Springer Verlag, 2007. MR2290010

Drton, M., Eichler, M., and Richardson, T. S. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10:2329–2348, 2009. MR2563984

Drton, M., Foygel, R., and Sullivant, S. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011. MR2816341

Duncan, O. T. *Introduction to Structural Equation Research*. Academic Press, 1975. MR0398558

Evans, R. J. Graphs for margins of bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016. MR3543314

Fox, C. J., Käufl, A., and Drton, M. On the causal interpretation of acyclic mixed graphs under multivariate normality. *Linear Algebra and Its Applications*, 473: 93–113, 2015. MR3338327

Foygel, R., Draisma, J., and Drton, M. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, 40(3): 1682–1713, 2012. MR3015040

Frot, B., Nandy, P., and Maathuis, M. H. Learning directed acyclic graphs with hidden variables via latent gaussian graphical model selection. 2017. URL https://arxiv.org/abs/1708.01151. Preprint.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. *mvtnorm: Multivariate Normal and t Distributions*, 2014. URL http://CRAN.R-project.org/package=mvtnorm. R package version 1.0–2.

Glymour, C. and Scheines, R. Causal modeling with the TETRAD program. *Synthese*, 68(1):37–63, 1986.

Hanley, J. A. and McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148 (3):839–843, 1983.

Jöreskog, K. G. A general method for analysis of covariance structures. *Biometrika*, 57(2):239–251, 1970. MR0269024

Jöreskog, K. G. *LISREL 8: User's Reference Guide*. Scientific Software International, 2001.

Kuipers, J. and Moffa, G. Uniform random generation of large acyclic digraphs. *Statistics and Computing*, 25(2):227–242, 2015. MR3306703

Maathuis, M. H., Kalisch, M., and Bühlmann, P. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A): 3133–3164, 2009. MR2549555

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4), 2010.

Malinsky, D. and Spirtes, P. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*, 88:371–384, 2017. MR3679153

Marchetti, G. M., Drton, M., and Sadeghi, K. *ggm: Functions for graphical Markov models*, 2015. URL http://CRAN.R-project.org/package=ggm. R package version 2.3.

Mardia, K. V., Kent, J. T., and Bibby, J. M. *Multivariate Analysis*. Academic Press, 1979. MR0560319

Melançon, G., Dutout, I., and Bousquet-Mélou, M. Random generation of directed acyclic graphs. *Electronic Notes in Discrete Mathematics*, 10:202–207, 2001. MR2154510

Nowzohour, C. *Estimating Causal Networks from Multivariate Observational Data*. PhD thesis, ETH Zürich, 2015. URL http://e-collection.library.ethz.ch/view/eth:48348.

Nowzohour, C. greedyBAPs, 2017. URL https://github.com/cnowzohour/greedyBAPs.

Pearl, J. *Causality*. Cambridge University Press, 2000. MR1744773

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

Richardson, T. A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI-09)*, 2009.

Richardson, T. and Spirtes, P. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002. MR1926166

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Shpitser, I., Richardson, T. S., Robins, J. M., and Evans, R. Parameter and structure learning in nested markov models. In *UAI Workshop on Causal Structure Learning*, 2012.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Silander, T. and Myllymäki, P. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the Twenty-Second Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452, 2006.

Silva, R. and Ghahramani, Z. Bayesian inference for gaussian mixed graph models. In *Proceedings of the Twenty-Second Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 453–460, 2006.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. Springer Verlag, 1993. MR1227558

Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. Using path diagrams as a structural equation modelling tool. *Sociological Methods & Research*, 27(2):182–225, 1998.

Tian, J. Identifying direct causal effects in linear models. In *AAAI'05 Proceedings of the 20th national conference on Artificial intelligence*, 2005.

Verma, T. S. and Pearl, J. Equivalence and synthesis of causal models. In *Proceedings of the sixth annual Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 220–227, 1991.

Williams, L. Equivalent models: Concepts, problems, and alternatives. In *The Handbook of Structural Equation Modeling*, pages 247–260, 2012.

Wright, S. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(2):161–215, 1934.

Wright, S. Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, 16(2):189–202, 1960.