JMLR: Workshop and Conference Proceedings vol 52, 380-391, 2016 PGM 2016

Bayesian Networks for Variable Groups

Pekka Parviainen PEKKA.PARVIAINEN@AALTO.FI
Samuel Kaski SAMUEL.KASKI@AALTO.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science
Aalto University, Espoo, Finland

Abstract

Bayesian networks, and especially their structures, are powerful tools for representing conditional independencies and dependencies between random variables. In applications where related variables form *a priori* known groups, chosen to represent different "views" to or aspects of the same entities, one may be more interested in modeling dependencies between groups of variables rather than between individual variables. Motivated by this, we study prospects of representing relationships between variable groups using Bayesian network structures. We show that for dependency structures between groups to be expressible exactly, the data have to satisfy the so-called groupwise faithfulness assumption. We also show that one cannot learn causal relations between groups using only groupwise conditional independencies, but also variable-wise relations are needed. Additionally, we present algorithms for finding the groupwise dependency structures.

1. Introduction

Bayesian networks are representations of joint distributions of random variables. They are powerful tools for modeling dependencies between variables. The dependencies and independencies are implied by the structure of a Bayesian network, which is represented by a directed acyclic graph (DAG).

In practical applications it is common that the analyst does not know the structure of a Bayesian network *a priori*. However, samples from the distribution of interest are commonly available. This has motivated development of algorithms for learning Bayesian networks from observational data. Although the problem is NP-hard (Chickering, 1996), there exist plenty of exact algorithms (Jaakkola et al., 2010; Silander and Myllymäki, 2006) as well as theoretically sound heuristics (Aliferis et al., 2010; Chickering, 2002).

Bayesian networks model dependencies and independencies between individual variables. However, often the relationships between groups of variables are even more interesting. An example is multiple different measurements of expression of the same genes, made with multiple measurement platforms, but the goal being to find relationships between the genes and not of the measurement platforms. The measurements of each gene would here be the groups. Another example is measurements of expression of individual genes, with the goal of the analysis being to understand cross-talk between pathways consisting of multiple genes, or more generally, relationships on a higher level of a hierarchy tree in hierarchically organized data. Here the pathways would be the groups. In both cases, a Bayesian network for variable groups would directly address the analysis problem, and would also have fewer variables and hence be easier to visualize.

More generally, the setup matches multi-view learning where data consist of multiple "views" to the same entity, multiple aspects of the same phenomenon, or multiple phenomena whose relationships we want to study. For these setups, a Bayesian network for variable groups can be seen as a dimensionality reduction technique with which we extract interesting information from a larger, noisy data set.

While the structure learning problem is well-studied for individual variables, knowledge about modeling relationships between variable groups using the Bayesian network framework is scarce. Motivated by this, we study prospects of learning Bayesian networks for variable groups. In summary, while Bayesian networks for variable groups can be learned under some conditions, strong assumptions are required and hence they have limited applicability.

We start by exploring theoretical possibilities and limitations for learning Bayesian networks for variable groups. First, we show that in order to be able to learn a structure that expresses exactly the conditional independencies between variable groups, the distribution and the groups need to together satisfy a condition that we call groupwise faithfulness (Section 3.1); our simulations suggest that this is a rather strong assumption. Then, we study possibilities of finding causal relations between variable groups. It turns out that one can draw only very limited causal conclusions based on only the conditional independencies between groups (Section 3.2), and hence also dependencies between the individual variables are needed.

We introduce methods for learning Bayesian network structures for variable groups. First, it is possible to learn a structure directly using conditional independencies or local scores between groups (Section 4.1). However, this approach suffers from needing lots of data. For the second approach, we observe that if all conditional independencies between individual variables are known, one can infer the conditional independencies between groups. The second approach is to construct a Bayesian network for individual variables and then to infer the structure between groups (Section 4.2). Finally, we evaluate the algorithms in practice (Section 5). Our results suggest that the second approach is more accurate.

1.1 Related Work

We are not aware of any work with close resemblance with this study, but there have been some efforts to solve related problems.

Object-oriented Bayesian networks (Koller and Pfeffer, 1997) are a generalization of Bayesian networks and enable representing groups of variables as objects. Hierarchical Bayesian networks (Gyftodimos and Flach, 2002) are another generalization of Bayesian networks where variables can be aggregations (or Cartesian products) of other variables. Variables form a hierarchical tree structure and a variable's parents are its parent in the tree and possibly some of its siblings. Both of these formalisms are very general and they are capable of representing conditional independencies between variable groups. Therefore, our results may be applied to these models. However, these models are unnecessarily complicated for our analysis and thus we do not consider them.

Module networks (Segal et al., 2005) have been designed to handle large data sets. The variables are partitioned into modules where the variables in the same module share parents and parameters. Module networks are particularly good for approximate density estimation. However, their structural limitations make them unsuitable for analysing conditional independencies between variable groups.

Burge and Lane (2006) have presented Bayesian networks for aggregation hierarchies which are related to hierarchical Bayesian networks. Groups of variables are aggregated by, for example, taking a maximum or mean and then networks are learned between the aggregated variables. From our point of view, the downside of this approach is that conditional independencies between aggregated variables do not necessarily correspond to conditional independencies between groups.

Entner and Hoyer (2012) have presented an algorithm for finding causal structures among groups of continuous variables. Their model works under the assumptions that variables are linearly related and associated with non-Gaussian noise.

2. Preliminaries

2.1 Conditional Independencies

Two random variables x and y are *conditionally independent* given a set S of random variables if $P(x, y|S) = P(x|S)P(y|S)$. If the set S is empty, variables x and y are marginally independent. We use $x \perp\!\!\!\perp y|S$ to denote that x and y are conditionally independent given S.

Conditional independence can be generalized to sets of random variables. Two sets of random variables X and Y are conditionally independent given a set S of random variables if $P(X, Y|S) = P(X|S)P(Y|S)$.

2.2 Bayesian Networks

A *Bayesian network* is a representation of a joint distribution of random variables. A Bayesian network consists of two parts: a structure and parameters. The structure of a Bayesian network is a directed acyclic graph (DAG) which expresses the conditional independencies and the parameters determine the conditional distributions.

Formally, a DAG is a pair (N, A) where N is the node set and A is the arc set. If there is an arc from u to v, that is, $uv \in A$ then we say that u is a *parent* of v and v is a *child* of u. The set of parents of v in A is denoted by A_v. Nodes v and u are said to be *spouses* of each other if they have a common child and there is no arc between v and u. Further, if there is a directed path from u to v we say that u is an *ancestor* of v and v is a *descendant* of u. The cardinality of N is denoted by n. When there is no ambiguity on the node set N, we identify a DAG by its arc set A.

Each node in a Bayesian network is associated with a conditional probability distribution of the node given its parents. The conditional probability distribution is specified by the parameters. A DAG represents a joint probability distribution over a set of random variables if the joint distribution satisfies the *local Markov condition*, that is, every node is conditionally independent of its non-descendants given its parents. Then the joint distribution over a node set N can be written as $P(N) = \prod_{v \in N} P(v|A_v)$.

The conditional independencies implied by a DAG can be extracted using a d-separation criterion. The *skeleton* of a DAG A is an undirected graph that is obtained by replacing all directed arcs $uv \in A$ with undirected edges between u and v. A *path* in a DAG is a cycle-free sequence of edges in the corresponding skeleton. A node v is a *head-to-head node* along a path if there are two consecutive arcs uv and wv on that path. Nodes v and u are *d-connected* by nodes Z along a path from v to u if every head-to-head node along the path is in Z or has a descendant in Z and none of the other nodes along the path is in Z. Nodes v and u are *d-separated* by nodes Z if they are not d-connected by Z along any path from v to u.

Nodes s, t, and u form a *v-structure* in a DAG if s and t are spouses and u is their common child. Two DAGs are said to be *Markov equivalent* if they imply the same set of conditional independence statements. It can be shown that two DAGs are Markov equivalent if and only if they have the same skeleton and same v-structures (Verma and Pearl, 1990).

A distribution p is said to be *faithful* to a DAG A if A and p imply exactly the same set of conditional independencies. If p is faithful to A then v and u are conditionally independent given Z in p if and only if v and u are d-separated by Z in A. This generalizes to variable sets. That is, if p is faithful to A then variable sets T and U are conditionally independent given Z in p if and only if t and u are d-separated by Z in A for all $t \in T$ and $u \in U$.

3. Groupwise Independencies

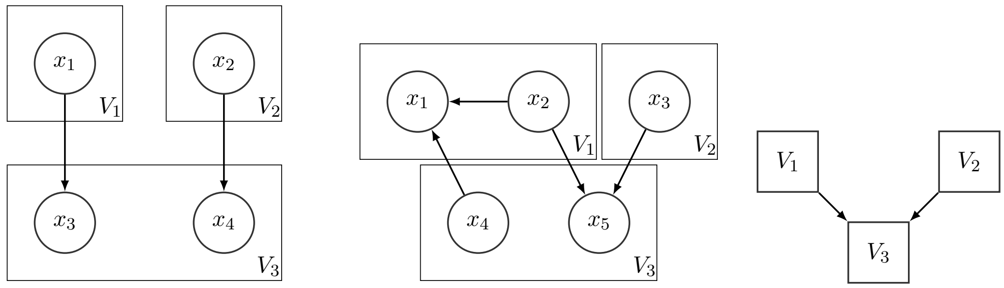
In this section we introduce a new assumption, groupwise faithfulness, that is necessary for principled learning of DAGs for variable groups. We will also show that groupwise conditional independencies have a limited role in learning causal relations between groups.

3.1 Groupwise Faithfulness

First, let us introduce some terminology. Recall that N is our node set. Let $W = \{W_1, \ldots, W_k\}$ be a collection of nonempty sets where $W_i \subseteq N \, \forall i$, and W forms a partition of N. We call W a *grouping*. We call a DAG on N a *variable DAG* and a DAG on W a *group DAG*; Note that the nodes of the group DAG are subsets of N. We try to solve the following computational problem. We are given a grouping W and data D from a distribution p on variables N that is faithful to a variable DAG G. The task is to learn a group DAG H on W such that for all $W_i, W_j \in W$ and $S = \cup_l T_l$, with $T = \{T_1, \ldots, T_k\} \subseteq W \setminus \{W_i, W_j\}$, it holds that W_i and W_j are d-separated by S in H if and only if $W_i \perp\!\!\!\perp W_j | S$ in p.

It is well-known that DAGs are not closed under marginalization. That is, even though the data-generating distribution is faithful to a DAG on a node set N, it is possible that the conditional independencies on some subset of N are not exactly representable by any DAG. We note that DAGs are not closed under aggregation, either. By aggregation we mean representing conditional independencies among groups using a group DAG. We show that by presenting an example. Consider a distribution that is faithful to the DAG in Figure 1(a). We want to express conditional independencies between groups V_1, V_2, and V_3. By inferring conditional independencies from the variable DAG, we get that $V_1 \perp\!\!\!\perp V_2$ and $V_1 \perp\!\!\!\perp V_2 | V_3$. There does not exist a DAG that expresses this set of conditional independencies exactly.

To avoid cases where conditional independencies are not representable by any group DAG, we introduce a new assumption: groupwise faithfulness. Formally, we define groupwise faithfulness as follows.



(a) (b) (c)

Figure 1: (a) A variable DAG where conditional independencies among groups V_1, V_2, and V_3 cannot be expressed exactly using any DAG. (b) A causal variable DAG where conditional independencies among groups V_1, V_2, and V_3 lead to a group DAG in which v-structures cannot be interpreted causally. (c) A group DAG corresponding to causal variable DAG in (b).

Definition 1 (Groupwise faithfulness) *A distribution p is groupwise faithful to a group DAG H given a grouping W, if H implies the exactly same set of conditional independencies as p over the groups W.*

Note that this assumption is analogous with the faithfulness assumption in the sense that in both cases there exists a DAG that expresses exactly the independencies in the distribution.

Sometimes it is convenient to investigate whether conditional independencies implied by a variable DAG given a grouping are equivalent to the conditional independencies implied by a group DAG.

Definition 2 (Groupwise Markov equivalence) *A variable DAG G is groupwise Markov equivalent to a group DAG H given a grouping W, if H implies the exactly same set of conditional independencies as G over groups W.*

We note that if a distribution p is faithful to a DAG G, and G is groupwise Markov equivalent to a DAG H given a grouping W, then p is groupwise faithful to H given W. This shows that faithfulness and groupwise Markov equivalence together imply groupwise faithfulness. However, neither faithfulness nor groupwise Markov equivalence alone is necessary or sufficient for groupwise faithfulness.

To see this, let us consider the following examples. First, to see that faithfulness is not sufficient for groupwise faithfulness, assume that we have a distribution that is faithful to the DAG in Figure 1(a). Given groups V_1, V_2, and V_3, the distribution is groupwise unfaithful. Second, consider a distribution over the variable set x_1, x_2, x_3, x_4, and x_5. Let us assume that the groups are $V_1 = \{x_1, x_2\}$, $V_2 = \{x_3\}$, and $V_3 = \{x_4, x_5\}$ and the Bayesian network factorizes according to the variable DAG in Figure 1(b). Now, it is possible to construct a distribution such that the local conditional distribution at node x_1 is exclusive or (XOR), and thus the variable DAG is unfaithful. If the other local conditional distributions do not introduce any additional independencies then the distribution is groupwise faithful. This shows that faithfulness is not necessary for groupwise faithfulness. Next, let us consider the same structure but let us assume that both x_1 and x_5 are associated with XOR distributions. In this case the variable DAG is groupwise Markov equivalent to the group DAG but the distribution is not groupwise faithful which shows that groupwise Markov equivalence is not sufficient for groupwise faithfulness. Finally, consider the variable DAG and the grouping in Figure 1(a). This variable DAG is not groupwise Markov equivalent to the group DAG given the grouping. However, if the distribution is unfaithful to the DAG and the variables x_1 and x_3 are independent then the distribution is groupwise faithful. This shows that groupwise Markov equivalence is not necessary for groupwise faithfulness. As neither faithfulness nor groupwise Markov equivalence is sufficient or necessary for groupwise faithfulness, it follows that groupwise faithfulness implies neither faithfulness or groupwise Markov equivalence.

Next, we will explore how strong the groupwise faithfulness assumption is. That is, how likely we are to encounter groupwise faithful distributions. To this end, we consider distributions that are faithful to variable DAGs. The joint space of DAGs and groupings is too large to be enumerated and we are not aware of any formula for assessing the number of groupwise unfaithful networks. Therefore, we analyze the prevalence of groupwise faithfulness by an empirical evaluation using simulations.

In simulations, a key question is how to check groupwise faithfulness. That is, given a variable DAG and a grouping, how to check whether the conditional independencies entailed by the variable DAG over groups can be represented exactly using a group DAG. This can be done by first using the PC algorithm (Spirtes et al., 2000) to construct a group DAG; here we use d-separation in the variable DAG as our independence test. Once the group DAG has been constructed we can check that the set of conditional independencies entailed by the group DAG is exactly the set of groupwise conditional independencies implied by the variable DAG and the grouping. The PC algorithm is sound and complete so if there exists a DAG that implies exactly the set of given conditional independencies, then the PC algorithm returns (the equivalence class of) that DAG. Thus, the conditional independencies match if and only if the variable DAG and the grouping are faithful to a group DAG.

We used the Erdős-Rényi model to generate random DAGs. A DAG from model $G(n, p)$ has n nodes and each arc is included with probability p independently of all other arcs; to get an acyclic directed graph, we fix the order of nodes. We generated random DAGs with $n = 20$ by varying the parameter p from 0.1 to 0.9. We also generated random groupings where group size was fixed to 2, 3, 4, or 5 (20 is not divisible by 3, so in this case one group is smaller than the others). For each value of p, we generated 100 random graphs. Then, we generated 10 groupings for each graph for each group size and counted the proportion of groupwise faithful DAG-grouping pairs. The results are shown in Figure 2. It can be seen that groupwise unfaithfulness is probable with sparse graphs and small group sizes. One should, however, note that the simulation results are for random graphs and groupings, and real life graphs and groupings may or may not follow this pattern.

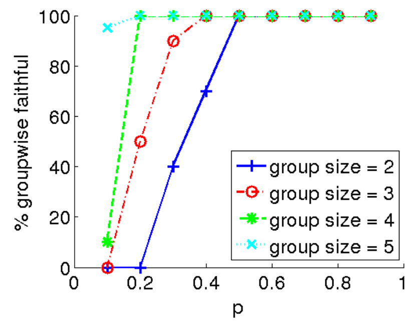


Figure 2: Proportion of DAG-grouping pairs that are groupwise faithful in random graphs of 20 nodes. Parameter p is the probability that an arc is present.

3.2 Causal Interpretation

Probabilistic causation between variables is typically defined to concern predicting effects of interventions. This means that an external manipulator intervenes the system and forces certain variables to take certain values. A DAG is called *causal* if it satisfies the *causal Markov condition*, that is, all variables are conditionally independent of their non-effects given their direct causes. Assuming faithfulness and causal sufficiency (if any pair of observed variables has a common cause then it is observed), it is possible to identify causal effects using the *do*-operator (Pearl, 2000). The *do*-operator $do(v = v_1)$ sets the value of the variable v to be v_1. The probability $P(u|do(v = v_1))$ is the conditional probability distribution of u given that the variable v has been forced to take value v_1. In other words, one takes the original joint distribution, removes all arcs that head to v and sets $v = v_1$; then one computes the probability $P(u|v = v_1)$ in the new distribution. We define a cause using the so-called operational criterion for causality (Aliferis et al., 2010), that is, we say that a variable v is a *cause* (direct or indirect) of a variable u if and only if $P(u|do(v = v_1)) \neq P(u|do(v = v_2))$ for some values v_1 and v_2. A straightforward generalization leads to the following definition of causality for variable groups.

Definition 3 (Group causality) *Given variable groups V and U, V is a cause of U if $P(U|do(V = V_1)) \neq P(U|do(V = V_2))$ for some instantiations V_1 and V_2 of values of V.*

Note that the above definition allows causal cycles between groups. To see this, consider a causal DAG on $\{v_1, v_2, v_3, v_4\}$ which has arcs v_1v_3 and v_4v_2. If there are two groups $W_1 = \{v_1, v_2\}$ and $W_2 = \{v_3, v_4\}$ then W_1 is a cause of W_2 (because there is a causal arc v_1v_3) and W_2 is a cause of W_1 (because of a causal arc v_4v_2).

Next, we will study to what extent causality between variable groups can be detected from observational data using only conditional independencies among groups. We assume that the data come from a distribution that is faithful to a causal variable DAG. Further, we assume that we have no access to the raw data but only to an oracle that conducts conditional independence tests. Formally, we assume that we have access to an oracle $\mathcal{O}_G$ that answers queries $W_i \perp\!\!\!\perp W_j | S$, where $W_i, W_j \in W$ and $S = \cup_l T_l$ with $T = \{T_1, \ldots T_m\} \subseteq W \setminus \{W_i, W_j\}$. Note that in the standard scenario with conditional independencies between variables, we have an oracle $\mathcal{O}_V$ that answers queries $X \perp\!\!\!\perp Y | Z$, where $X, Y \in N$ and $Z \subseteq N \setminus \{X, Y\}$; If $\max_i |W_i| > 1$ then the oracle $\mathcal{O}_V$ is strictly more powerful than $\mathcal{O}_G$.

It is well-known that, under standard assumptions, a causal variable DAG can be learned up to the Markov equivalence class. A Markov equivalence class can be represented by a completed partial DAG (CPDAG) where we have both directed and undirected edges. Directed edges or arcs are the edges that point to the same direction in every member of the equivalence class whereas undirected edges express cases where the edge is not directed to the same direction in all members of the equivalence class. If there is a directed path from a variable v to a variable u in the CPDAG then v is a cause of u. In other words, existence of such a path is a sufficient condition for causality. However, it is not a necessary condition and it is possible that v is a cause of u even when there is no directed path from v to u in the CPDAG.

Next, we consider causality in the group context. Manipulating an ancestor of a node affects its distribution and thus the ancestor is a cause of its descendant. It is easy to see that given a causal variable DAG G, a group W_i is a group cause of a group W_j if and only if there is at least one directed path from W_i to W_j in G, that is, there exists $v \in W_i$ and $u \in W_j$ such that there is a directed path from v to u. It is clear from the above that a sufficient condition for a group W_i to be a group cause of a group W_j is that there is at least one directed path from W_i to W_j in the CPDAG.

Standard constraint-based algorithms for causal learning start by constructing a skeleton and then directing arcs based on a set of rules. So let us take a look on these rules in the group context. The first rule is to direct v-structures. The following theorem shows that arcs that are part of a v-structure in a group DAG imply group causality.

Theorem 4 *Let N be a node set and W a grouping on N. Let p be a distribution that is groupwise faithful to some group DAG H given the grouping W. If there exist groups $W_i, W_j, W_k \in W$ such that (i) $W_i \perp\!\!\!\perp W_k | S$ for some $S \subseteq W \setminus \{W_i, W_j, W_k\}$ and (ii) $W_i \not\perp\!\!\!\perp W_k | (\cup_l T_l) \cup W_j$ for all $T = \{T_1, \ldots, T_m\} \subseteq W \setminus \{W_i, W_j, W_k\}$ then W_i is a group cause of W_j.*

Proof It is sufficient to show that there exists a pair $w_i \in W_i$ and $w_j \in W_j$ such that w_i is an ancestor of w_j in the variable DAG.

Due to (i), all paths that go from W_i to W_k without visiting S must have a head-to-head node. Due to (ii) there has to exist at least one path between W_i and W_k such that there are no non-head-to-head nodes in $W \setminus \{W_i, W_k\}$ and all head-to-head nodes are unblocked by W_j; let us denote one such a path by R. Without loss of generality, we can assume that all nodes in R except the endpoints are in $W \setminus \{W_i, W_k\}$. Let $s, t, u \in N$ be three consecutive nodes in path R such that there are edges st and ut. Nodes s and u cannot be head-to-head nodes along R and therefore $s, u \in W_i \cup W_k$. Node t is a head-to-head node and therefore either $t \in W_j$ or t has a descendant in W_j. In both cases there is a directed path from both s and u to the set W_j. The path R has one end-point in W_i and another in W_k. Thus, there is a directed path from W_i to W_j in the variable DAG. ∎

Note that the proof of the previous theorem implies that there is a v-structure $W_i \to W_j \leftarrow W_k$ in the group DAG only if there exists $w_i \in W_i$, $w_j \in W_j$, and $w_k \in W_k$ such that there exists a v-structure $w_i \to w_j \leftarrow w_k$ in the variable DAG.

After v-structures have been directed, one can direct the rest of the edges that point to the same direction in every DAG of the Markov equivalence class using four local rules often referred to as the Meek rules (Meek, 1995). The rules are (Pearl, 2000):

R1: Orient $v - s$ into $v \to s$ if there is an arrow $u \to v$ such that u and s are nonadjacent.

R2: Orient $u - v$ into $u \to v$ if there is a chain $u \to s \to v$.

R3: Orient $u - v$ into $u \to v$ if there are two chains $u - s \to v$ and $u - t \to v$ such that s and t are nonadjacent.

R4: Orient $u - v$ into $u \to v$ if there are two chains $u - s \to t$ and $s \to t \to v$ such that s and v are nonadjacent and u and t are adjacent.

We would like to generalize these rules for variable groups. However, these rules are not sufficient to infer group causality if one does have access only to the groupwise conditional independencies (and to nothing else). To see this, consider a group DAG $H = (W, E)$ where $W = \{S, T, U, V\}$ and $E = \{SU, TU, UV\}$. Now, Theorem 4 says that S and T are causes of U. The rule R1 suggest that we could claim that U is a cause of V. However, we can construct a causal variable DAG $G = (N, F)$ with $N = \{s_1, s_2, t_1, t_2, u_1, u_2, u_3, v_1, v_2\}$ and $F = \{s_1u_1, t_1u_1, v_2u_2, u_2t_2, v_1u_3, u_3s_2\}$ and $S = \{s_1, s_2\}$, $T = \{t_1, t_2\}$, $U = \{u_1, u_2, u_3\}$, and $V = \{v_1, v_2\}$. Clearly, G implies the same conditional independencies on W as does H and there is no directed path from U to V in G. Thus, U is not a cause of V in G.

The above observation implies that the Meek rules cannot be used to infer causality in group DAGs. However, it is not known whether there are some special conditions under which the Meek rules would apply in this context. Note that the above applies only when the conditional independencies between individual variables are not known; when the variable DAG is known, this information can be used to help to infer more causal relations.

4. Algorithms

Next, we will introduce two approaches for learning group DAGs.

4.1 Direct Learning

The most straightforward approach is to learn a group DAG directly, that is, either using conditional independencies or local scores on a grouping W. In other words, we can consider each group as a variable. Assuming that the variables are discrete, the possible states of the new variable w_i, corresponding to the group W_i, are the Cartesian product of the states of the variables in W_i. Now there is a bijective mapping between joint configurations of variables in W_i and states of w_i. Thus $W_i \perp\!\!\!\perp W_j | S_1$ if and only if $w_i \perp\!\!\!\perp w_j | S_2$ where $W_l \subseteq S_1$ if and only if $w_l \in S_2$. This leads to a simple procedure described in Algorithm 1. The procedure

FINDVARIABLEDAG in the second step is an exact algorithm for finding a DAG; it can use either the constraint-based or score-based approach.

Algorithm 1 FINDGROUPDAG1

Input: Data D on a node set N, a grouping W on N.
Output: Group DAG G

1: Convert variables $x_i \in N$ into new variables y_j on W such that $y_j = \times_{x_i \in W_j} x_i$.
2: Learn a DAG G on the new variables on W using procedure FINDVARIABLEDAG.
3: **return** G

4.2 Learning via Variable DAGs

We note that a DAG over individual variables specifies also all the conditional independencies and dependencies between groups. Thus, it is possible to learn a group DAG by first learning a variable DAG and then inferring the group DAG. Algorithm 2 summarizes this approach.

Algorithm 2 FINDGROUPDAG2

Input: Data D on a node set N, a grouping W on N.
Output: Group DAG G

1: Learn a DAG H on N using procedure FINDVARIABLEDAG.
2: Learn a group DAG G on W using the PC algorithm and d-separation in H as an independence test.
3: **return** G

5. Experiments

5.1 Implementations

We implemented our algorithms using Matlab. The implementation is available at `http://research.cs.aalto.fi/pml/software/GroupBN/`. The implementation of PC algorithm from the BNT toolbox[1] was used as the constraint-based version of procedure FINDVARIABLEDAG. As the score-based version, we used the state-of-the-art integer linear programming algorithm GOBNILP[2].

5.2 Simulations

We generated data from three different Bayesian network structures called structures 1, 2, and 3 having 30, 40, and 50 nodes, respectively, divided into 10 equally sized groups. All structures were groupwise faithful to the group DAG; the network structures are not shown due to space constraints. For each structure we generated 50 binary-valued Bayesian networks by sampling the parameters uniformly at random. Then, we sampled data sets of size 100, 500, 2000, and 10000 from each of the Bayesian networks.

1. `https://code.google.com/p/bnt/`
2. `http://www.cs.york.ac.uk/aig/sw/gobnilp/`

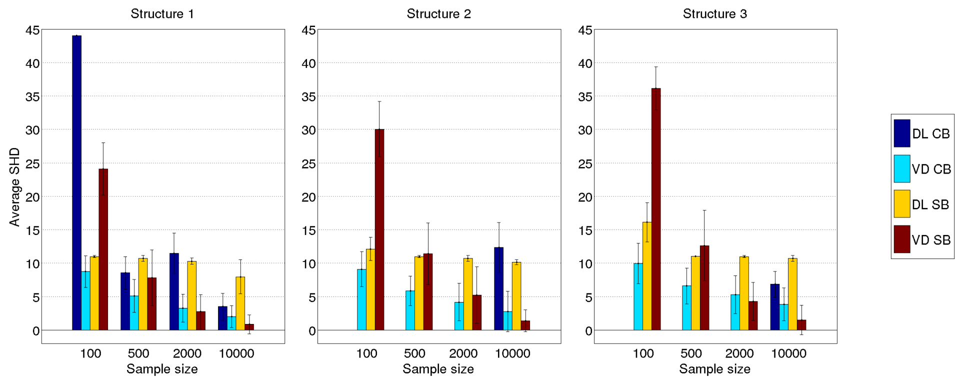


Figure 3: Average SHD (Structural Hamming Distance) between the learned group DAG and the true group DAG when the data were generated from three different structures. DL = direct learning, VD = learning using variable DAGs, CB = constraint-based, SB = score-based. The numbers on the x-axis are sample sizes. Missing bars for constraint-based direct learning are due to the algorithm running out of memory.

We ran both the constraint-based and score-based version of Algorithms 1 and 2. In all tests we used a 2 GB memory limit. The results are shown in Figure 3. It is clear that direct learning is inferior compared to learning via variable DAGs. This is due to the fact that variables in the direct learning approach have lots of states and thus direct learning requires lots of data to draw any conclusions. Based on the results, it seems that the constraint-based approach outperforms the score-based approach when there are few samples, and the roles are reversed once the sample size grows.

6. Discussion

In this paper we introduced the concept of group DAG for modeling conditional independencies and dependencies between groups of random variables, and studied prospects of learning group DAGs. It turned out, perhaps unsurprisingly, that many aspects become more complicated when moving from individual variables to groups of variables.

We have assumed that the variable groups are known beforehand, as prior knowledge, and asked what can be done with the extra prior knowledge. A natural follow-up question is that can the groups be learned from data. Even though this interesting question is superficially related it is, however, a distinct and very different problem that is likely to require a different machinery. Multiple different goals for such a clustering of variables are possible and sensible.

Acknowledgements

The authors thank Antti Hyttinen, Esa Junttila, Jefrey Lijffijt, and Teemu Roos for useful discussions. The work was partially funded by The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN). The experimental results were computed using computer resources within the Aalto University School of Science "Science-IT" project.

References

C. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. Koutsoukos. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, 11: 171–234, 2010.

J. Burge and T. Lane. Improving Bayesian Network Structure Search with Random Variable Aggregation Hierarchies. In *ECML*, pages 66–77. Springer, Berlin, Heidelberg, 2006.

D. Chickering. Learning Bayesian networks is NP-Complete. In *Learning from Data: Artificial Intelligence and Statistics*, pages 121–130. Springer-Verlag, 1996.

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

D. Entner and P. Hoyer. Estimating a Causal Order among Groups of Variables in Linear Models. In *ICANN*, pages 83–90. Spinger, 2012.

E. Gyftodimos and P. Flach. Hierarchical Bayesian networks: a probabilistic reasoning model for structured domains. In *ICML-2002 Workshop on Development of Representations*, 2002.

T. Jaakkola, D. Sontag, A. Globerson, and M. Meila. Learning Bayesian Network Structure using LP Relaxations. In *AISTATS*, pages 358–365, 2010.

D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In *UAI*, pages 302–313. Morgan Kaufmann Publishers Inc., 1997.

C. Meek. Causal Inference and Causal Explanation with Background Knowledge. In *UAI*, pages 403–410. Morgan Kaufmann, 1995.

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

E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman. Learning Module Networks. *Journal of Machine Learning Reseach*, 6:557–588, Oct. 2005.

T. Silander and P. Myllymäki. A simple approach for finding the globally optimal Bayesian network structure. In *UAI*, pages 445–452. AUAI Press, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer Verlag, 2000.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *UAI*, pages 255–270. Elsevier, 1990.