

# Labeled Directed Acyclic Graphs: a generalization of context-specific independence in directed graphical models

Johan Pensar<sup>\*,1</sup>, Henrik Nyman<sup>1</sup>, Timo Koski<sup>3</sup>, Jukka Corander<sup>1,2</sup>

<sup>1</sup>Department of Mathematics, Åbo Akademi University, Finland

<sup>2</sup>Department of Mathematics and statistics, University of Helsinki, Finland

<sup>3</sup>Department of Mathematics

KTH Royal Institute of Technology, Stockholm, Sweden

\*Corresponding author, Email: jopensar@abo.fi

## Abstract

We introduce a novel class of labeled directed acyclic graph (LDAG) models for finite sets of discrete variables. LDAGs generalize earlier proposals for allowing local structures in the conditional probability distribution of a node, such that unrestricted label sets determine which edges can be deleted from the underlying directed acyclic graph (DAG) for a given context. Several properties of these models are derived, including a generalization of the concept of Markov equivalence classes. Efficient Bayesian learning of LDAGs is enabled by introducing an LDAG-based factorization of the Dirichlet prior for the model parameters, such that the marginal likelihood can be calculated analytically. In addition, we develop a novel prior distribution for the model structures that can appropriately penalize a model for its labeling complexity. A non-reversible Markov chain Monte Carlo algorithm combined with a greedy hill climbing approach is used for illustrating the useful properties of LDAG models for both real and synthetic data sets.

**Keywords:** Directed acyclic graph; Graphical model; Context-specific independence; Bayesian model learning; Markov chain Monte Carlo

## 1 Introduction

Directed acyclic graphs have gained widespread popularity as representations of complex multivariate systems (Koski and Noble (2009); Koller and Friedman (2009)). Despite their advantageous properties for representing dependencies among variables in a modular fashion, several proposals for making them more flexible and parsimonious have been presented (Boutilier et al (1996); Friedman and Goldszmidt (1996); Chickering et al (1997); Eriksen (1999); Poole and Zhang (2003); Koller and Friedman (2009)). In particular, an important notion is to allow the dependencies to have local structures, such that a node need not explicitly depend on all the combinations of values of its parents. This leads to context-specific independence which can substantially reduce the parametric dimensionality of a network model and lead to a more expressive interpretation of the dependence structure (Boutilier et al (1996); Friedman and Goldszmidt (1996); Poole and Zhang (2003); Koller and Friedman (2009)). Context-specific independencies have also been seemingly separately considered for undirected graphical models by multiple authors (Corander (2003); Højsgaard (2003, 2004)).

Here we generalize the context-specific independence models proposed in Boutilier et al (1996) by allowing the independencies to be represented in terms of labels for the parental configurations of a node in an unrestricted manner. This approach thus goes beyond the trees of conditional probability tables considered in Boutilier et al (1996) as it instead introduces a partition of the parental configurations into classes with invariant conditional probability distributions for the outcomes that are assigned to the same class. It is shown that such a definition leads to a model class with a number of desirable

properties, and we derive several properties of the models, including their identifiability and an LDAG version of the concept of a Markov equivalence class.

We develop an efficient method for Bayesian learning of LDAG models from a set of training data by introducing a prior distribution for the model parameters that factorizes in a comparable manner as the standard Dirichlet distribution used for learning DAG models. Since the prior enables an analytical evaluation of the marginal likelihood of an LDAG, the model space can be searched relatively fast for structures that are associated with high posterior probabilities. To do this in practice, we combine a non-reversible Markov chain Monte Carlo algorithm with a greedy hill climbing approach to obtain a method that is not computationally too expensive.

The structure of this article is as follows. In Section 2 we introduce the LDAG models (2.1) and investigate their properties (2.2). In Section 3 we develop the Bayesian learning method and apply it to both real and synthetic data sets in Section 4 to illustrate the favorable properties of our approach. Some concluding remarks are provided in the final section.

## 2 DAG- and LDAG-based graphical models

### 2.1 Preliminaries and introduction of LDAGs

A DAG is a directed graph that is built up by nodes and directed edges. The acyclic property ensures that no directed path starting from a node leads back to that particular node. The concept of DAG-based graphical models, or Bayesian networks, was formalized by Pearl (1988). In a Bayesian network, the nodes represent variables and the directed edges represent direct dependencies among the variables. Correspondingly, absence of edges represents statements of conditional independence. The constraints imposed by the structure of a DAG alone have been recognized to be unnecessarily stringent under certain circumstances where context-specific or asymmetric independence can play a natural role in the models. In general, two approaches have been considered for this problem. The most common approach is based on different representations of conditional probability distributions (CPD) that are hidden behind the graph topology (Boutilier et al (1996); Chickering et al (1997); Poole and Zhang (2003)). The other approach has focused on the topology of the graph structure itself (Heckerman (1991); Geiger and Heckerman (1996)). Heckerman (1991) introduced similarity networks which are made up of multiple networks. This representation and the related Bayesian multinets are further examined in Geiger and Heckerman (1996). They show how asymmetric independencies can be represented by multiple Bayesian networks and how these independence assertions can speed up inference.

In this paper we will bring the CPD- and graph-based approaches together by introducing a graphical representation scheme in the form of labeled DAGs whose associated CPDs can be stored in compact tables based on rules. To illustrate the concept of LDAGs, we consider the following example from Geiger and Heckerman (1996), p. 52:

A guard of a secured building expects three types of persons ( $h$ ) to approach the building's entrance: workers in the buildings, approved visitors, and spies. As a person approaches the building, the guard can note its gender ( $g$ ) and whether or not the person wears a badge ( $b$ ). Spies are mostly men. Spies always wear badges in an attempt to fool the guard. Visitors don't wear badges because they don't have one. Female workers tend to wear badges more often than do male workers. The task of the guard is to identify the type of person approaching the building.

This scenario can be represented by the DAG on top in Figure 1. The topology of this graph, however, hides the fact that gender and badge wearing are conditionally independent, given that the person is a spy or visitor. The corresponding joint probability distribution is, as a result of this, overparameterized in the sense that it requires a total of 11 free parameters although some of these are identical. Geiger and Heckerman (1996) noticed that this scenario is better represented by the multiple graphs reproduced here in the middle of Figure 1. This representation is made up of two context-specific

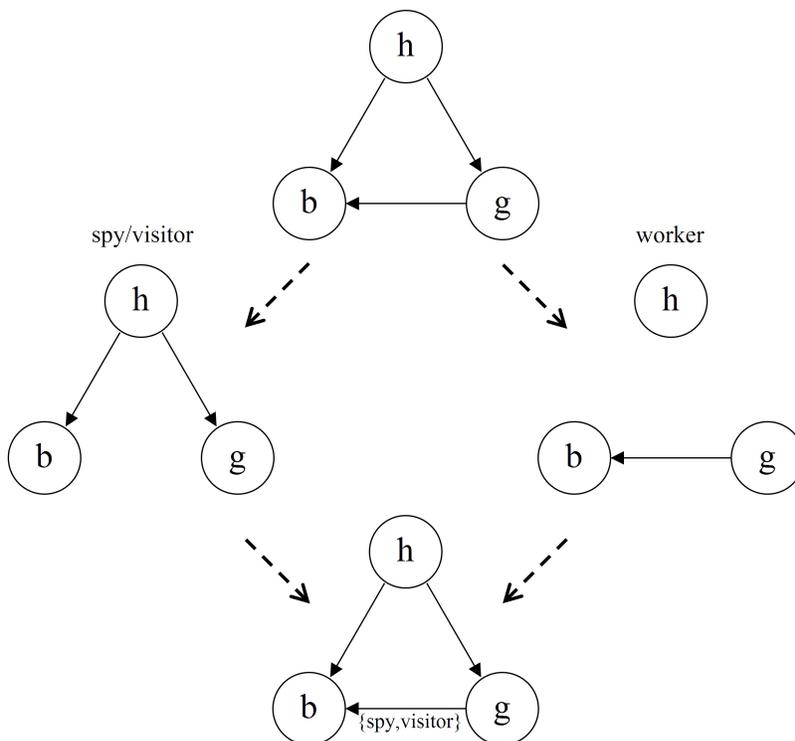


Figure 1: Graph structures describing the spy/visitor/worker-scenario.

graphs that together show that the dependence between gender and badge wearing only holds in the context of the person being a worker. The corresponding joint probability distribution now only requires 9 free parameters. Now consider the labeled DAG on the bottom in Figure 1. We have added the label  $\{spy, visitor\}$  to the edge  $(g, b)$ . This label implies that gender and badge are independent given that the person approaching the building is a spy or a visitor. Although an LDAG is global in its representation, it can still represent independencies that only hold in certain contexts. This allows it to represent the same dependence structure that requires multiple graphs using the multinet approach. As for the multinet approach, this type of representation requires 9 free parameters.

Before stating any formal definitions, we provide some notations. A DAG will be denoted by  $G = (V, E)$  where  $V = \{1, \dots, d\}$  is the set of nodes and  $E \subset V \times V$  is the set of edges such that if  $(i, j) \in E$  then the graph contains a directed edge from node  $i$  to  $j$ . Nodes, from which there is a directed edge to node  $j$ , are called parents of  $j$  and the set of all such nodes is denoted by  $\Pi_j = \{i \in V : (i, j) \in E\}$ . The nodes  $V$  give the indices of a set of stochastic variables  $X = \{X_1, \dots, X_d\}$ . Due to the close relationship between a node and its corresponding variable, the terms node and variable are used interchangeably. We use small letters  $x_j$  to denote a value taken by the corresponding variable. If  $S \subseteq V$ , then  $X_S$  denotes the corresponding set of variables. The outcome space of a variable  $X_j$  is denoted by  $\mathcal{X}_j$  and the joint outcome space of a set of variables by the Cartesian product  $\mathcal{X}_S = \times_{j \in S} \mathcal{X}_j$ . The cardinality of the outcome space of  $X_S$  is denoted by  $|\mathcal{X}_S|$ .

An ordinary DAG encodes independence statements in the form of conditional independencies.

**Definition 1.** *Conditional Independence (CI)*

Let  $X = \{X_1, \dots, X_d\}$  be a set of stochastic variables where  $V = \{1, \dots, d\}$  and let  $A, B, S$  be three disjoint subsets of  $V$ .  $X_A$  is conditionally independent of  $X_B$  given  $X_S$  if

$$p(x_A | x_B, x_S) = p(x_A | x_S)$$

holds for all  $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$  whenever  $p(x_B, x_S) > 0$ . This will be denoted by

$$X_A \perp X_B \mid X_S.$$

If we let  $X_S = \emptyset$ , then  $X_A \perp X_B$  simply denotes marginal independence between the two sets of variables. The most basic statements of conditional independence, reflected by a DAG, follow the directed local Markov property. It implies that each variable  $X_j$  is conditionally independent of its non-descendants given its parental variables  $X_{\Pi_j}$ . This leads to a unique explicit factorization of the joint distribution into lower order distributions,

$$p(X_1, \dots, X_d) = \prod_{j=1}^d p(X_j \mid X_{\Pi_j}), \quad (1)$$

where the factors are CPDs that correspond to local structures. By local structure, we refer to the node itself, its parents and the edges from the parents to the node. The topology of an ordinary DAG restricts it to only encoding for independence relations that hold globally. However, as shown in the example above, it is natural to consider independence relations that only hold in certain contexts. The following notion of context-specific independence was formalized by Boutilier et al (1996).

**Definition 2.** *Context-Specific Independence (CSI)*

Let  $X = \{X_1, \dots, X_d\}$  be a set of stochastic variables where  $V = \{1, \dots, d\}$  and let  $A, B, C, S$  be four disjoint subsets of  $V$ .  $X_A$  is contextually independent of  $X_B$  given  $X_S$  and the context  $X_C = x_C$  if

$$p(x_A \mid x_B, x_C, x_S) = p(x_A \mid x_C, x_S),$$

holds for all  $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$  whenever  $p(x_B, x_C, x_S) > 0$ . This will be denoted by

$$X_A \perp X_B \mid x_C, X_S.$$

It has been discovered by numerous authors that certain CSIs can naturally be captured simply by further refining (1). We will refer to these statements as local CSIs as they are confined to the local structures.

**Definition 3.** *Local CSI in a DAG*

A CSI in a DAG is local if it is of the form  $X_j \perp X_B \mid x_C$ , where  $B$  and  $C$  form a partition of the parents of node  $j$ .

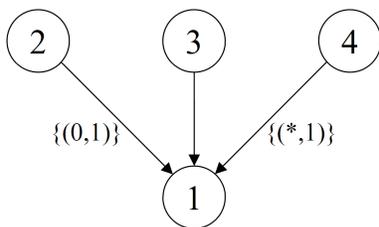
In the CPD-based approaches to including CSI in Bayesian networks, the context-specific local structures cannot be read directly off the graph structure. This is the key to the usefulness of multinets. A multinet offers a natural representation of the dependence structure by explicitly showing the independencies in a graphical form. In an attempt to further pursue this idea, we introduce a graph topology that is able to visualize the local CSIs directly as a part of a single graph structure as done in Figure 1. To achieve this we add labels to the edges in a similar way as Corander (2003). This enables incorporation of local CSIs in a single graph as opposed to multiple networks -approaches where one might need one graph for each distinct context. An LDAG is now formally defined as a DAG with labels representing local CSIs.

**Definition 4.** *Labeled Directed Acyclic Graph (LDAG)*

Let  $G = (V, E)$  be a DAG over the stochastic variables  $\{X_1, \dots, X_d\}$ . For all  $(i, j) \in E$ , let  $L_{(i,j)} = \Pi_j \setminus \{i\}$ . A label on an edge  $(i, j) \in E$  is defined as the set

$$\mathcal{L}_{(i,j)} = \{x_{L_{(i,j)}} \in \mathcal{X}_{L_{(i,j)}} : X_j \perp X_i \mid x_{L_{(i,j)}}\}.$$

An LDAG is a DAG to which the label set  $\mathcal{L}_E = \{\mathcal{L}_{(i,j)} : \mathcal{L}_{(i,j)} \neq \emptyset\}_{(i,j) \in E}$  has been added, it is denoted by  $G_L = (V, E, \mathcal{L}_E)$



$$\mathcal{L}_{(2,1)} = (0, 1) \quad \Rightarrow X_1 \perp X_2 \mid (X_3, X_4) = (0, 1)$$

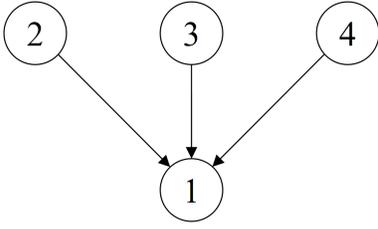
$$\begin{aligned} \mathcal{L}_{(4,1)} = \mathcal{X}_2 \times \{1\} &\quad \Rightarrow X_1 \perp X_4 \mid X_2 \in \mathcal{X}_2, X_3 = 1 \\ &\quad \Leftrightarrow X_1 \perp X_4 \mid X_2, X_3 = 1 \end{aligned}$$

Figure 2: Local CSI-structure and the corresponding local CSIs.

With respect to a fixed ordering of the variables, the labels do not have to contain any variable indices as  $X_{L(i,j)}$  contains all the parental variables in the node's local structure except the one that is part of the edge. A node must naturally have at least two parents for it to be possible for an incoming edge to contain a label. In subsequent examples, we assume that the variables are binary with  $\mathcal{X}_j = \{0, 1\}$ . However, the derived theory applies to non-binary variables as well. For  $L(i,j) = \{k, l\}$ , we will use  $(*, x_l)$  to denote a label set of the form  $\mathcal{X}_k \times \{x_l\}$ . Figure 2 now illustrates how labels may be added to the edges of a local structure and how they should be interpreted. The number of configurations relative to the number of possible configurations in a label can be considered an indication of the strength of the dependence conveyed by the corresponding edge in that particular local structure.

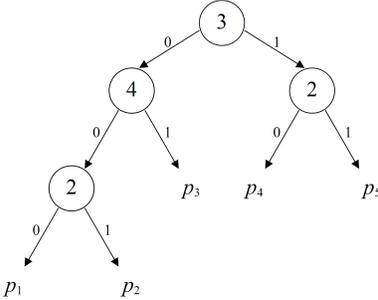
The CPD-based approach for generalizing Bayesian networks utilizes the fact that local CSIs correspond to certain regularities among the CPDs arising in the factorization (1). The focus has therefore been to find different ways of representing the CPDs. Boutilier et al (1996) use decision trees in which certain local CSI-structures can be captured in a very natural way. However, due to the replication problem (Pagallo and Haussler (1990)), trees are somewhat limited in their expression power when it comes to certain types of CSI-structure. Chickering et al (1997) overcome this shortcoming by using a more general type of representation in the form of decision graphs. Unfortunately, decision graphs usually leave the scope of CSI complicating the interpretability of the models from a (in)dependence perspective. In addition, such models lack the ability to exploit CSI in inference. Next we investigate how the tree- and decision graph-based approaches are connected to the LDAG representation which in fact can be considered a compromise between the two representations. LDAGs allow for more expressive models than CPT-trees without leaving the scope of CSI which provides a natural interpretation and proven computational advantages when performing inference.

The textbook way of representing the CPDs is in the form of full conditional probability tables (CPT) of sizes  $(|\mathcal{X}_j| - 1) \cdot |\mathcal{X}_{\Pi_j}|$ . This form of representation requires tables that grow exponentially with the numbers of parents as it fails to capture regularities present in the CPDs. Including local CSIs in the model, however, directly implies that certain CPDs must be similar and need only be defined once. Consider the local structure and associated CPT in Figure 3. We use complete AND-rules to represent the distinct parent configurations. A rule is complete if all parental variables are part of it. By investigating the right column of the CPT, we see that there are only five distinct CPDs. Still, the naive approach requires us to define  $p(X_1 \mid x_{\Pi_1})$  for each distinct parent configuration  $x_{\Pi_1} \in \mathcal{X}_{\Pi_1}$ . It is therefore easy to realize that this representation is far from minimal. Using the approach of Boutilier et al (1996), the regularities in the CPT in Figure 3 can be captured by the CPT-tree in Figure 4. Each path in the tree corresponds to a rule that can be described by the AND-operator. By simply traversing down each distinct path until we reach a terminal node or leaf, we can transform the CPT in Figure 3 into its reduced counterpart on the right in Figure 4. All parent configurations



$X_{\Pi_1}$	$p(X_1   X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	$p_1$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	$p_3$
$X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 0$	$p_4$
$X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 1$	$p_4$
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	$p_2$
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	$p_3$
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 0$	$p_5$
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	$p_5$

Figure 3: Local structure and the associated CPT.



$X_{\Pi_j}$	$p(X_1   X_{\Pi_j})$
$X_3 = 0 \wedge X_4 = 0 \wedge X_2 = 0$	$p_1$
$X_3 = 0 \wedge X_4 = 0 \wedge X_2 = 1$	$p_2$
$X_3 = 0 \wedge X_4 = 1$	$p_3$
$X_3 = 1 \wedge X_2 = 0$	$p_4$
$X_3 = 1 \wedge X_2 = 1$	$p_5$

Figure 4: CPT-tree and the corresponding rule-based reduced CPT.

satisfying a certain rule give rise to the same CPD. This implies that the rules in a reduced CPT must be mutually exclusive for the representation to be minimal. The rules corresponding to a tree are mutually exclusive as two distinct paths cannot lead to the same leaf. If a variable is not part of a path (or the corresponding AND-rule), it implies that the particular variable is contextually independent of the variable associated with the CPT given the context encoded by the path (or rule). Following this method we can read off the following local CSIs:

$$\left. \begin{aligned}
 X_3 = 0 \wedge X_4 = 1 &\Rightarrow X_1 \perp X_2 \mid (X_3, X_4) = (0, 1) \\
 X_3 = 1 \wedge X_2 = 0 &\Rightarrow X_1 \perp X_4 \mid (X_2, X_3) = (0, 1) \\
 X_3 = 1 \wedge X_2 = 1 &\Rightarrow X_1 \perp X_4 \mid (X_2, X_3) = (1, 1)
 \end{aligned} \right\} \Leftrightarrow X_1 \perp X_4 \mid X_2, X_3 = 1$$

If we once more consider Figure 2, we see that the CSIs above coincide with the labels of this specific LDAG. More generally, any CPT-tree can be transformed into a reduced CPT by mutually exclusive AND-rules. Subsequently, incomplete rules can be turned into labels as illustrated in the above example.

Now consider the LDAG on the top in Figure 5 and its associated minimal reduced CPT on the bottom in Figure 6. This CSI-structure cannot be compactly represented by the structure of a CPT-tree. Depending on the order of the variables, we can at best represent either of the CSIs (tree 1 or tree 2 in Figure 5). As a result, duplicate subtrees can be found in both trees. This illustrates the weakness of tree structures. Once we split on a variable, it is rendered essential for the particular context even if it in the end does not affect the CPD. To represent this type of CSI-structures through a graphical representation, we need the more general decision graph used by Chickering et al (1997). As a node in a decision graph may have multiple parents, we can merge leaves with similar CPDs in a tree. Merging such leaves in tree 1 in Figure 5 results in the decision graph at the bottom in which all CSIs are represented. When merging leaves in a CPT-tree, situations may arise where some of the corresponding AND-rules are not mutually exclusive anymore. In order to achieve a minimal representation of the reduced CPT, any mutually non-exclusive AND-rules must be combined with the OR-operator.

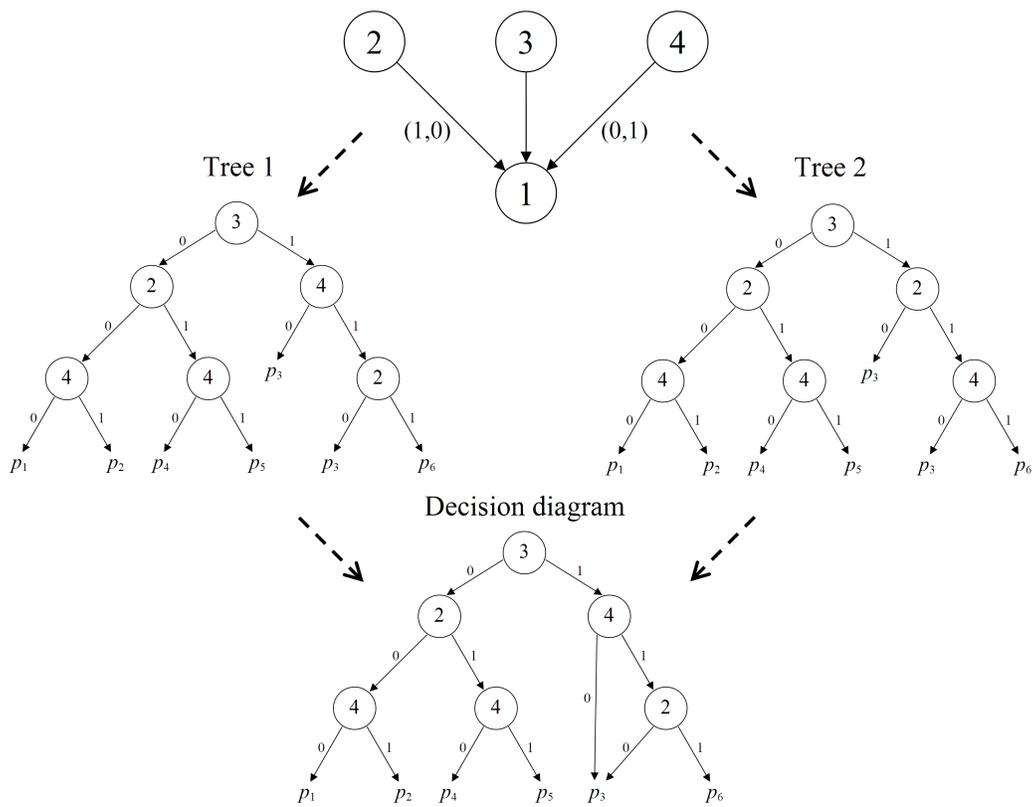


Figure 5: Tree- and graph-based representations of a local CSI-structure.

$X_{\Pi_1}$	$p(X_1   X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	$p_1$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	$p_2$
$X_2 = 0 \wedge X_3 = 1$	$p_3$
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	$p_4$
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	$p_5$
$X_3 = 1 \wedge X_4 = 0$	$p_3$
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	$p_6$



$X_{\Pi_1}$	$p(X_1   X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	$p_1$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	$p_2$
$(X_2 = 0 \wedge X_3 = 1) \vee (X_3 = 1 \wedge X_4 = 0)$	$p_3$
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	$p_4$
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	$p_5$
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	$p_6$

Figure 6: Constructing a minimal reduced CPT through a multi-step process.

To show how the minimal reduced CPT is recovered from the labels we proceed stepwise as shown in Figure 6. Each of the labels corresponds to a single reduced AND-rule resulting in the upper table. The rules on row 3 and 6 are not mutually exclusive at this point as they both are satisfied by the common parent configuration  $(X_2, X_3, X_4) = (0, 1, 0)$ . This implies that any parent configuration satisfying any of these rules must give rise to the same CPD. The AND-rules are therefore combined with the OR-operator resulting in the minimal reduced CPT on the bottom of the figure. More generally, each configuration in the labels of a local structure corresponds to an AND-rule. If any two rules overlap, they are combined with the OR-operator. The rules of a minimal reduced CPT created by this method will thus be mutually exclusive and exhaustive with respect to the outcome space of the parental variables.

A CPT-representation may in fact be viewed as a function that given a parent configuration returns a CPD. The common factor among the different CPD-based representations is that they all induce partitions of the outcome space of the parental variables. If the representation is based on the notion of CSI, the corresponding partition will be referred to as CSI-consistent. Decision graphs in general go beyond the scope of CSI as they are able to represent any arbitrary partition. Subsequently, a decision graph must fulfill certain criteria structure-wise for it to be consistent with CSI. Still, even if a decision graph indeed is consistent with respect to CSI, the interpretation of the local CSIs is not trivial. In an LDAG, however, all local CSIs can readily be recovered from the labels. By introducing the class of LDAGs we aim at balancing the expressive power of the models against their interpretability. CSI has a sound interpretation and arises naturally in various situations. From a computational perspective, CSI has also proven to be particularly useful in probabilistic inference.

Considerable research efforts have been devoted to outlining how CSI can be exploited in probabilistic inference. Probabilistic inference refers to the process of computing the posterior probability of a list of query variables given some observed variables. The key to efficient inference lies in the concept of factorization of the joint distribution. Incorporating local CSIs into the models, allows a further decomposition of (1) into a finer-grained factorization which in turn can speed up the inference. Boutilier et al (1996) investigate how CPT-trees can be used to speed up various inference algorithms. As a consequence of the replication problem, Poole (1997) concludes that rule-based versions may be more efficient than tree-based. Zhang and Poole (1999) give a more general analysis of the computational leverages that CSI has to offer without referring to any particular inference algorithms. Poole and Zhang (2003) further improve the efficiency of the approach of Poole (1997) by using the concept

of confactors which is a combination of contexts and tables. They introduce contextual belief networks which are similar to traditional Bayesian networks except that the CPDs are associated with parent contexts rather than explicit parent configurations. The labels in an LDAG correspond directly to the parent contexts of a contextual belief network. However, in the process of making the contexts mutually exclusive they proceed in a different manner than in Figure 6. Their approach is more beneficial inference-wise but less compact, which interferes with the learning procedure we discuss in the next section. It is worth noting, however, that one may simply choose the approach more suitable for the problem at hand.

Inference that exploits CSI has been quite thoroughly investigated by numerous authors. In this paper we thus focus on model identifiability and learning. From a practical point of view, one might argue that the sole existence of expressive and efficient models is not enough if these models cannot be accurately learned from a set of data. The more expressive the models, the harder the learning tends to get due to added flexibility. In section 3 we present a Bayesian learning scheme for LDAGs but first we attend the aspect of model identifiability and interpretability.

## 2.2 Properties of LDAGs

To facilitate the interpretation of the CSI-structure of LDAGs, we utilize two conditions introduced for labeled undirected graphical models (LGMs) in Corander (2003); maximality and regularity.

Given a local structure, different label combinations may induce the same local CSI-structure. This phenomenon may arise when the contexts of two or more labels overlap. If two distinct label combinations induce equivalent CSI-structures, it implies that the dissimilar label configurations can be added to each of the label combinations without inducing any new restrictions. To avoid this type of ambiguities we introduce the maximality condition for LDAGs.

**Definition 5.** *Maximal LDAG*

*An LDAG  $G_L = (V, E, \mathcal{L}_E)$  is called maximal if there exists no configuration  $x_{L(i,j)}$  that can be added to the label  $\mathcal{L}_{(i,j)}$  without inducing an additional local CSI.*

If we add a configuration to a label in a maximal LDAG, it must result in an additional restriction in form of a local CSI. This will in turn result in a reduction of the associated minimal reduced CPT.

**Theorem 1.** *Let  $G_L = (V, E, \mathcal{L}_E)$  and  $G_L^* = (V, E, \mathcal{L}_E^*)$  be two maximal LDAGs with the same underlying DAG  $G = (V, E)$ . Then  $G_L$  and  $G_L^*$  represent equivalent dependence structures if and only if  $\mathcal{L}_E = \mathcal{L}_E^*$ , i.e.  $G_L = G_L^*$ .*

*Proof.* Assume that  $G_L = (V, E, \mathcal{L}_E)$  and  $G_L^* = (V, E, \mathcal{L}_E^*)$  are two maximal LDAGs representing the same dependence structure and having the same underlying DAG  $G = (V, E)$ . Assume further that they have different labelings, i.e.  $\mathcal{L}_E \neq \mathcal{L}_E^*$ . There must thereby exist a configuration, say,  $x_{L(i,j)}^* \in \mathcal{L}_{(i,j)}^*$  such that  $x_{L(i,j)}^* \notin \mathcal{L}_{(i,j)}$ . This corresponds to the local CSI  $X_j \perp X_i \mid x_{L(i,j)}^*$  being explicitly represented in  $G_L^*$  but not in  $G_L$ . For the LDAGs to represent the same dependence structure, the local CSI must, however, hold for  $G_L$  as well. If the local CSI is implicitly represented by other labels in  $G_L$ , it implies that  $x_{L(i,j)}^*$  can be added to  $\mathcal{L}_{(i,j)}$  with inducing an additional local CSI. This contradicts the maximality condition. If the CSI is not implicitly represented by other labels in  $G_L$ , it implies that adding  $x_{L(i,j)}^*$  to  $\mathcal{L}_E$  induces an additional local CSI. This contradicts the assumption of the LDAGs representing the same dependence structure. This leads us to the conclusion that  $G_L$  and  $G_L^*$  represent the same dependence structures if and only if  $\mathcal{L}_E = \mathcal{L}_E^*$ .  $\square$

Each configuration in a label corresponds to a local CSI according to Definition 4. If an LDAG is not maximal, some CSIs cannot be obtained directly from the graph following the definition. These CSIs are, however, implicitly reflected by the graph as they arise from a combination of other CSIs explicitly represented by labels in the graph. In maximal LDAGs, all local CSIs can be obtained directly from the graph.

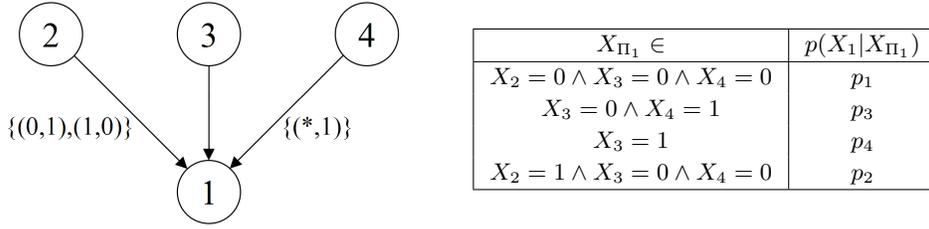


Figure 7: Local CSI-structure and the associated minimal reduced CPT.

To illustrate the maximality property, consider the local structure in Figure 7. The local structure is similar to the one in Figure 2 except that configuration  $x_{L(2,1)} = (1, 0)$  has been added to its label. The local structure in Figure 7 is now not maximal since  $(1, 1)$  can be added to  $\mathcal{L}_{(2,1)}$  without resulting in an additional CSI. An intuitive way of reaching this conclusion is to consider the rules in the associated minimal reduced CPT next to the local structure. The rule on the third row arises from a combination of mutually non-exclusive rules:

$$\left. \begin{array}{l} x_{L(2,1)} = (1, 0) \Rightarrow X_3 = 1 \wedge X_4 = 0 \\ x_{L(4,1)} = (0, 1) \Rightarrow X_2 = 0 \wedge X_3 = 1 \\ x_{L(4,1)} = (1, 1) \Rightarrow X_2 = 1 \wedge X_3 = 1 \end{array} \right\} \Rightarrow X_3 = 1$$

As  $(0, 1, 1)$  and  $(1, 1, 1)$  satisfy this rule, no further merging of rules is done when  $x_{L(2,1)} = (1, 1)$  is added to its label. This corresponds to the local CSI

$$X_1 \perp X_2 \mid X_3 = 1, X_4 = 1$$

implicitly being encoded by the other labels. This type of situation may arise when different label induced rules overlap and are combined with the OR-operator in order to achieve a minimal number of mutually exclusive rules.

The maximality condition is proven to be an essential condition for LDAGs. Without it we may fail in the interpretation of both local and, consequently, non-local CSIs which we consider later in this section. Failing in the interpretation of local CSIs hampers the efficiency of inference algorithms as useful CSIs may be neglected. A naive approach for testing whether an LDAG is maximal or not is simply to test each configuration that is not part of a label by adding it and checking if it results in a combining of rules or not. If there exists a configuration  $x_{L(i,j)} \notin \mathcal{L}_{(i,j)}$  for which all parent contexts  $\{x_{L(i,j)}\} \times \mathcal{X}_i$  satisfy the same rule in the associated minimal reduced CPT, the LDAG is not maximal.

To ensure that the effect of an edge cannot completely vanish due to labels, we introduce the regularity condition for maximal LDAGs.

**Definition 6.** *Regular maximal LDAG*

A maximal LDAG  $G_L = (V, E, \mathcal{L}_E)$  is regular if  $\mathcal{L}_{(i,j)}$  is a strict subset of  $\mathcal{X}_{L(i,j)}$  for every label in  $G_L$ .

The regularity condition is illustrated in Figure 8. The LDAG in Figure 8a satisfies the regularity condition, however, this LDAG is not maximal. To make the LDAG maximal, the configuration  $(1, 1)$  must be added to  $\mathcal{L}_{(4,1)}$  (Figure 8b) which now contains all possible configurations and thereby renders the maximal LDAG non-regular.

**Theorem 2.** *In a regular maximal LDAG, a label  $\mathcal{L}_{(i,j)}$  cannot induce an independence assertion of the form  $X_j \perp X_i \mid x_{L(i,j)}$  for all  $x_{L(i,j)} \in \mathcal{X}_{L(i,j)}$ , i.e.  $X_j \perp X_i \mid X_{L(i,j)}$ .*

*Proof.* Assume that  $\mathcal{L}_{(i,j)}$  is a label in a regular maximal LDAG  $G_L$ . The maximality condition ensures that we cannot add configurations to  $\mathcal{L}_{(i,j)}$  without altering the dependence structure. This means that  $X_j \perp X_i \mid x_{L(i,j)}$  must hold for  $x_{L(i,j)} \in \mathcal{L}_{(i,j)}$  but not for  $x_{L(i,j)} \in \mathcal{X}_{L(i,j)} \setminus \mathcal{L}_{(i,j)}$ . Due to the regularity condition, we have that  $\mathcal{L}_{(i,j)} \subset \mathcal{X}_{L(i,j)}$  and, consequently,  $\mathcal{X}_{L(i,j)} \setminus \mathcal{L}_{(i,j)} \neq \emptyset$ . It must thereby exist an outcome  $x_{L(i,j)}^*$  for which  $X_j \not\perp X_i \mid x_{L(i,j)}^*$ .  $\square$

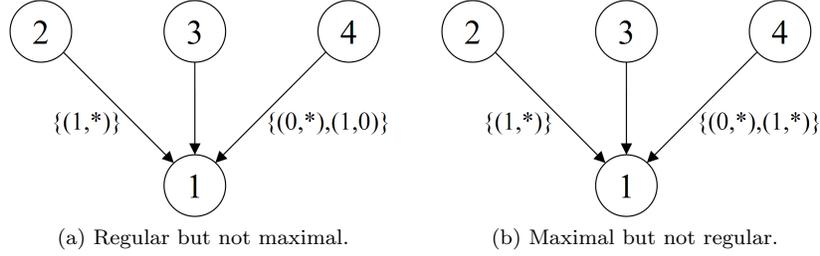


Figure 8: Regularity condition for a maximal LDAG.

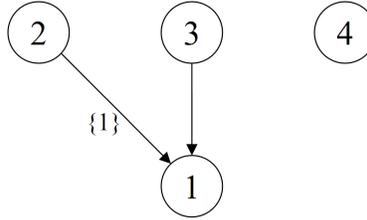


Figure 9: Simplified version of the LDAG from Figure 8b.

If we consider the non-regular LDAG in Figure 8b, it can be simplified as the maximal regular LDAG in Figure 9 without altering the dependence structure. We can restrict our model space to the class of maximal regular LDAGs which is considerably smaller than the class of all LDAGs, without suffering any loss of generality as the dependence structure of an arbitrary LDAG can be represented by a maximal regular LDAG.

The independence assertions that can be recovered directly or indirectly from the structure of an LDAG  $G_L$  can be divided into local and non-local. The local CIs follow from the directed local Markov property while the local CSIs can be attained from the labels. The set of all local independencies associated with  $G_L$  will be denoted by  $\mathcal{I}_{loc}(G_L)$ . In addition to the local independencies, there are additional non-local independencies which can be derived from  $\mathcal{I}_{loc}(G_L)$ . The set of all local and non-local independencies, denoted by  $\mathcal{I}(G_L)$ , fully describes the dependence structure of  $G_L$ . However, the dependence structure of an LDAG or  $\mathcal{I}(G_L)$  is still ultimately determined by the local independencies since all non-local independencies are implicitly represented by  $\mathcal{I}_{loc}(G_L)$ .

Let  $P$  denote a distribution over the same set of variables as an LDAG  $G_L$  and let  $\mathcal{I}(P)$  denote the set of CSIs satisfied by  $P$ . If  $P$  factorizes according to  $G_L$ , it must hold that  $\mathcal{I}_{loc}(G_L) \subseteq \mathcal{I}(G_L) \subseteq \mathcal{I}(P)$  and  $G_L$  is called a CSI-map of  $P$ . There may, however, exist distribution-specific independencies that hold in  $P$  even when they are not represented by the structure of  $G_L$ . A distribution  $P$  is said to be faithful to  $G_L$  if equality  $\mathcal{I}(G_L) = \mathcal{I}(P)$  holds. The LDAG is then called a perfect CSI-map of  $P$  and can be considered a true representation in the sense that no artificial dependencies are introduced.

The derivation of non-local CIs in ordinary DAGs can be very cumbersome. Instead, non-local CIs can be verified utilizing the concept of  $d$ -separation. Boutilier et al (1996) introduce a sound counterpart for context-specific independence; CSI-separation. They reduce the problem of checking for CSI-separation by checking for ordinary variable independence in a simpler context-specific graph. To formulate the concept of CSI-separation for LDAGs, the following notions are introduced.

**Definition 7.** *Satisfied label*

Let  $G_L = (V, E, \mathcal{L}_E)$  be an LDAG and  $X_C = x_C$  a context where  $C \subseteq V$ . In the context  $X_C = x_C$ , a label  $\mathcal{L}_{(i,j)} \in \mathcal{L}_E$  is satisfied if  $L_{(i,j)} \cap C \neq \emptyset$  and  $\{x_{L_{(i,j)} \cap C}\} \times \mathcal{X}_{L_{(i,j)} \setminus C} \subseteq \mathcal{L}_{(i,j)}$ .

**Definition 8.** *Context-specific LDAG*

Let  $G_L = (V, E, \mathcal{L}_E)$  be an LDAG. For the context  $X_C = x_C$ , where  $C \subseteq V$ , the context-specific

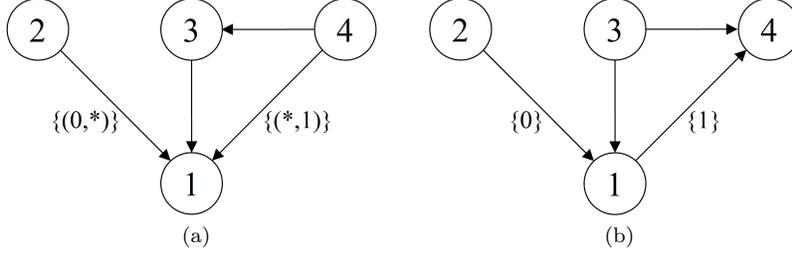


Figure 10: LDAGs with CI inducing CSI-structures.

LDAG is denoted by  $G_L(x_C) = (V, E \setminus E', \mathcal{L}_{E \setminus E'})$  where  $E' = \{(i, j) \in E : \mathcal{L}_{(i,j)} \text{ is satisfied}\}$ . The underlying DAG of the context-specific LDAG is denoted by  $G(x_C) = (V, E \setminus E')$ .

A context-specific LDAG is a reduced version of an LDAG where all satisfied edges are removed. CSI-separation can now be defined in a similar manner as in Boutilier et al (1996).

**Definition 9.** *CSI-separation in LDAGs*

Let  $G_L = (V, E, \mathcal{L}_E)$  be an LDAG and let  $A, B, S, C$  be four disjoint subsets of  $V$ .  $X_A$  is CSI-separated from  $X_B$  by  $X_S$  in the context  $X_C = x_C$  in  $G_L$ , denoted by

$$X_A \perp X_B \parallel_{G_L} x_C, X_S,$$

if  $X_A$  is  $d$ -separated from  $X_B$  by  $X_{S \cup C}$  in  $G(x_C)$ .

If  $C = \emptyset$  in the above definition, the method describes the procedure of  $d$ -separation with respect to the underlying DAG. CSI-separation is proven to be a sound method for verifying CSIs, i.e.

$$X_A \perp X_B \parallel_{G_L} x_C, X_S \Rightarrow X_A \perp X_B \mid x_C, X_S.$$

Unfortunately, it is not complete in the sense that there may arise situations where certain structure induced independencies cannot be discovered directly by the CSI-separation algorithm. Koller and Friedman (2009) noticed that it may be necessary to perform reasoning by cases to recover all independencies reflected by CSI-based structures. If CSI-separation holds for every  $x_C \in \mathcal{X}_C$  in the above definition it will imply conditional independence. Consider the LDAG in Figure 10a. When only considering the underlying DAG, it appears that

$$X_2 \not\perp X_4 \parallel_G X_1, X_3 \Rightarrow X_2 \not\perp X_4 \mid X_1, X_3$$

However, through CSI-separation and reasoning by cases we recover

$$\begin{aligned} X_2 \perp X_4 \parallel_{G_L} X_1, X_3 = 0 &\Rightarrow X_2 \perp X_4 \mid X_1, X_3 = 0 \\ X_2 \perp X_4 \parallel_{G_L} X_1, X_3 = 1 &\Rightarrow X_2 \perp X_4 \mid X_1, X_3 = 1 \end{aligned}$$

which eventually leads us to the conclusion that

$$X_2 \perp X_4 \mid X_1, x_3 \quad \forall x_3 \in \mathcal{X}_3 \Leftrightarrow X_2 \perp X_4 \mid X_1, X_3.$$

$d$ -separation is based on the notion of active trails, i.e. trails along which information can flow from one variable to another, and a lack of such trails will imply  $d$ -separation. Labels in an LDAG have the ability to cut off an active trail for a certain context by removing an edge in it and render the trail non-active or blocked in that context. The regularity condition prohibits this from occurring throughout the outcome space for a single edge but certain combinations of labels can still deactivate a trail that appears active when only considering the underlying DAG.

Now consider the LDAG in Figure 10b. When considering the underlying DAG alone, it appears that

$$X_2 \not\perp X_4 \parallel_G \emptyset \Rightarrow X_2 \not\perp X_4.$$

However, we can recover the following CSIs through CSI-separation:

$$\left. \begin{array}{l} X_2 \perp X_4 \parallel_{G_L} X_3 = 0 \\ X_2 \perp X_4 \parallel_{G_L} X_3 = 1 \end{array} \right\} \Rightarrow X_2 \perp X_4 \mid X_3$$

$$X_2 \perp X_3 \parallel_G \emptyset \Rightarrow X_2 \perp X_3$$

The first of the CIs must be discovered through reasoning by cases in the same way as in the previous example while the second is easily discovered from the underlying DAG. Combining the CIs leads us indirectly to the conclusion that

$$X_2 \perp X_4$$

indeed holds due to the structural properties of the LDAG. Several CSI-separation statements work together in order to achieve a non-local independence that is not easily discovered. However, both these situations are special cases that can only arise when the complete outcome space of a subset of variables is split up over several labels.

Earlier we introduced the class of regular maximal LDAGs and concluded that we can restrict the model space to this substantially smaller subclass without losing any generality. However, in this subclass there still exist large classes of distinct LDAGs that encode equivalent dependence structures. Heckerman et al (1995) highlighted a fundamental aspect in that classes of distinct DAGs may determine the same statistical model. Every DAG within such a class will determine the same set of CI restrictions among the variables in the model. Andersson et al (1997) characterized these so called Markov equivalence (also known as  $\mathcal{I}$ -equivalence) classes by concluding that each class corresponds to an essential graph in the form of a chain graph. As for DAGs, the difference between two equivalent LDAGs can occur from reversing non-essential edges. It is worth noting that the criteria for an edge being essential will differ from DAGs. This observation is based on the fact that the direction of the edges determines which local CSIs may be included in an LDAG.

All DAGs within the same Markov equivalence class share the same dependence structure or  $\mathcal{I}(G)$ . Correspondingly, we now define CSI-equivalence for LDAGs.

**Definition 10.** *CSI-equivalence for LDAGs*

Let  $G_L = (V, E, \mathcal{L}_E)$  and  $G_L^* = (V, E^*, \mathcal{L}_E^*)$  be two distinct regular maximal LDAGs. The LDAGs are said to be CSI-equivalent if  $\mathcal{I}(G_L) = \mathcal{I}(G_L^*)$ . A set containing all CSI-equivalent LDAGs forms a CSI-equivalence class.

In the remainder of this section we will discuss some structural properties that two distinct LDAGs must fulfill to belong to the same CSI-equivalence class. We begin by considering the underlying DAG.

**Theorem 3.** *Let  $G_L = (V, E, \mathcal{L}_E)$  and  $G_L^* = (V, E^*, \mathcal{L}_E^*)$  be two regular maximal LDAGs belonging to the same CSI-equivalence class. Their underlying DAGs  $G = (V, E)$  and  $G^* = (V, E^*)$  must then have the same skeleton.*

*Proof.* This theorem is a direct consequence of Theorem 2.  $\square$

Next we introduce a criterion that ties together the concept of CSI-equivalence among LDAGs and the concept of Markov equivalence among DAGs.

**Theorem 4.** *Let  $G_L = (V, E, \mathcal{L}_E)$  and  $G_L^* = (V, E^*, \mathcal{L}_E^*)$  be two maximal regular LDAGs for which there exists distributions  $P$  and  $P^*$  such that  $\mathcal{I}(G_L) = \mathcal{I}(P)$  and  $\mathcal{I}(G_L^*) = \mathcal{I}(P^*)$ .  $G_L$  and  $G_L^*$  are CSI-equivalent if and only if their corresponding context-specific LDAGs  $G_L(x_V) = G(x_V)$  and  $G_L^*(x_V) = G^*(x_V)$  are Markov equivalent for all  $x_V \in \mathcal{X}_V$ .*

*Proof.* Let  $G_L = (V, E, \mathcal{L}_E)$  and  $G_L^* = (V, E^*, \mathcal{L}_E^*)$  be two maximal regular LDAGs for which there exists distributions  $P$  and  $P^*$  such that  $\mathcal{I}(G_L) = \mathcal{I}(P)$  and  $\mathcal{I}(G_L^*) = \mathcal{I}(P^*)$ .

( $\Rightarrow$ ) Assume that  $G_L$  and  $G_L^*$  are CSI-equivalent. Assume further that there exists a joint outcome  $x_V \in \mathcal{X}_V$  for which  $G(x_V)$  and  $G^*(x_V)$  are not Markov equivalent, i.e. they have different (1) skeletons or (2) immoralities. (1) If they have different skeletons, there exists an edge  $\{i, j\}$  in, say, the skeleton of  $G(x_V)$  that does not exist in the skeleton of  $G^*(x_V)$ . Due to Theorem 3, the underlying DAGs must have the same skeleton. The lack of the edge in  $G^*(x_V)$  implies that a local CSI of the form  $X_j \perp X_i \mid x_{L(i,j)}$  holds in  $G_L^*$  but not in  $G_L$ . (2) Assume that there exists an immorality  $i \rightarrow j \leftarrow k$  in, say,  $G(x_V)$  that does not exist in  $G^*(x_V)$ . If there does not exist an edge between  $i$  and  $k$  in  $G_L$  (and  $G_L^*$ ), there must exist some  $S \subseteq V \setminus \{i, j, k\}$  for which  $X_i \perp X_k \parallel_G X_S$  and consequently  $\{X_i \perp X_k \mid X_S\} \in \mathcal{I}(G_L)$  while  $\{X_i \perp X_k \mid X_S\} \notin \mathcal{I}(G_L^*)$  since there exists at least one active trail between  $X_i$  and  $X_k$  via  $X_j$ . If there exists an edge between  $i$  and  $k$  in  $G_L$  (and  $G_L^*$ ), there must exist a local CSI of the form  $X_i \perp X_k \mid x_{L(k,i)}$  (or  $X_k \perp X_i \mid x_{L(i,k)}$ ) that holds in  $G_L^*$  but not in  $G_L$  since  $j \in L(k,i)$  (or  $j \in L(i,k)$ ) in  $G_L^*$  while  $j \notin L(k,i)$  (and  $j \notin L(i,k)$ ) in  $G_L$ . (1) and (2) allow us to conclude that  $\mathcal{I}(G_L) \neq \mathcal{I}(G_L^*)$  which contradicts the assumption of  $G_L$  and  $G_L^*$  being CSI-equivalent.  $G_L(x_V)$  and  $G_L^*(x_V)$  must be Markov equivalent for all  $x_V \in \mathcal{X}_V$ .

( $\Leftarrow$ ) Assume that  $G_L(x_V)$  and  $G_L^*(x_V)$  are Markov equivalent for all  $x_V \in \mathcal{X}_V$ . Let  $P$  be a distribution for which  $G_L$  is a perfect CSI-map. Each joint probability  $p(X_V = x_V)$  factorizes according to  $G_L(x_V)$ . Since  $G_L(x_V)$  and  $G_L^*(x_V)$  are Markov equivalent, we can refactorize each joint probability  $p(X_V = x_V)$  according to  $G_L^*(x_V)$  without altering the joint distribution or inducing any additional dependencies. This means that  $G_L^*$  is also a perfect CSI-map of  $P$ . Since  $\mathcal{I}(G_L) = \mathcal{I}(P) = \mathcal{I}(G_L^*)$ , we can conclude that  $G_L$  and  $G_L^*$  are CSI-equivalent.  $\square$

From this theorem it is clear that the LDAGs in Figure 10 are indeed CSI-equivalent even if some of the independencies are not so obvious. In order to check CSI-equivalence between two LDAGs, it suffices to compare context-specific graphs for only a subset of variables since not all will affect the structure of the graphs. Furthermore, all outcomes for which no labels in either graph are satisfied, need only to be checked once as the context-specific graphs in all these cases are equal to the underlying DAG. This last observation leads to a more strict version of Theorem 3 given that a specific condition is satisfied.

**Theorem 5.** *Let  $G_L = (V, E, \mathcal{L}_E)$  and  $G_L^* = (V, E^*, \mathcal{L}_E^*)$  be two regular maximal LDAGs that are CSI-equivalent and let their labelings be such that there exists at least one joint outcome,  $x_V \in \mathcal{X}_V$ , for which no label is satisfied. Their underlying DAGs  $G = (V, E)$  and  $G^* = (V, E^*)$  must then be Markov equivalent.*

*Proof.* This theorem is a direct consequence of Theorem 4.  $\square$

### 3 Bayesian learning of LDAGs by non-reversible MCMC

This section will attend the intricate problem of learning the LDAG structure from a set of data. This poses some obvious problems due to the extremely vast model space as well as some additional not so obvious problems due to the flexibility of the models. We introduce a structural learning method that utilizes a non-reversible Markov Chain Monte Carlo (MCMC) method combined with greedy hill climbing. Such a combination of a stochastic and a deterministic algorithm provides solid performance with a reasonable time complexity. A Bayesian score is used to evaluate the appropriateness of an LDAG given a set of observed data. In order to prevent overfitting, we impose a prior distribution that allows us to balance the ability of an LDAG to match the available learning data with its complexity. We begin with some additional notations.

Let  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$  denote a set of training data consisting of  $n$  observations  $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$  of the variables  $\{X_1, \dots, X_d\}$  such that  $\mathbf{x}_i \in \mathcal{X}$ . We assume that  $\mathbf{X}$  is complete in the sense that it contains no missing values. We denote an LDAG by  $G_L$  and  $\mathcal{G}_L$  denotes the set of all regular maximal LDAGs.

We let  $\Theta_{G_L}$  denote the parameter space induced by an LDAG and  $\dim(\Theta_{G_L})$  denotes the number of free parameters spanning the parameter space. An instance  $\theta \in \Theta_{G_L}$  corresponds to a specific joint distribution that factorizes according to the LDAG  $G_L$ . The CSI-consistent partition of the outcome space  $\mathcal{X}_{\Pi_j}$  is denoted by  $\mathcal{S}_{\Pi_j} = \{S_{j1}, \dots, S_{jk_j}\}$  where  $k_j = |\mathcal{S}_{\Pi_j}|$  is the number of outcome classes. We let  $r_j = |\mathcal{X}_j|$  and  $q_j = |\mathcal{X}_{\Pi_j}|$  denote the cardinality of the outcome space of variable  $X_j$  and its parents  $X_{\Pi_j}$ , respectively. Finally, we use  $n(x_{ij} \times S_{jl})$  to denote the total count of the configurations  $\{x_{ij}\} \times S_{jl}$  in  $\mathbf{X}$ .

In the Bayesian approach to model learning, one considers the posterior distribution of the models given some data,

$$p(G_L | \mathbf{X}) = \frac{p(\mathbf{X} | G_L) \cdot p(G_L)}{\sum_{G_L \in \mathcal{G}_L} p(\mathbf{X} | G_L) \cdot p(G_L)}. \quad (2)$$

Here  $p(\mathbf{X} | G_L)$  is the marginal probability of observing the data  $\mathbf{X}$  (evidence) given a specific LDAG  $G_L$  and  $p(G_L)$  denotes the prior probability of the LDAG. The denominator is a normalizing constant that does not depend on  $G_L$  and it can be ignored for the purpose of comparing particular graphs. Our main interest is to find the maximum a posteriori model, i.e. the solution to

$$\arg \max_{G_L \in \mathcal{G}_L} p(\mathbf{X} | G_L) \cdot p(G_L). \quad (3)$$

To evaluate  $p(\mathbf{X} | G_L)$ , we need to consider all possible instances of the parameter vector satisfying the independencies encoded by the LDAG and weight them with respect to a prior according to

$$p(\mathbf{X} | G_L) = \int_{\theta \in \Theta_{G_L}} p(\mathbf{X} | G_L, \theta) \cdot f(\theta | G_L) d\theta, \quad (4)$$

where  $p(\mathbf{X} | G_L, \theta)$  and  $f(\theta | G_L)$  are the respective likelihood function and prior distribution over the parameters, given the graph  $G_L$ .

Under certain assumptions, (4) can be solved analytically for DAGs, see Cooper and Herskovitz (1992) and Buntine (1991). Heckerman et al (1995) identify and discuss the assumptions in detail. Friedman and Goldszmidt (1996) and Chickering et al (1997) derive a corresponding closed-form expression for structures based on CPT-trees and decision graphs, respectively. Since an LDAG induces partitionings of the parental outcome spaces in a similar manner as these previous works, the marginal likelihood of an LDAG can be expressed as

$$p(\mathbf{X} | G_L) = \prod_{j=1}^d \prod_{l=1}^{k_j} \frac{\Gamma(\sum_{i=1}^{r_j} \alpha_{ijl})}{\Gamma(n(S_{jl}) + \sum_{i=1}^{r_j} \alpha_{ijl})} \prod_{i=1}^{r_j} \frac{\Gamma(n(x_{ji} \times S_{jl}) + \alpha_{ijl})}{\Gamma(\alpha_{ijl})}, \quad (5)$$

where  $n(\cdot)$  is the count defined earlier and the  $\alpha_{ijl}$ 's are hyperparameters (also known as pseudocounts) defining a collection of local Dirichlet distributions. The hyperparameters characterize our prior belief about the CPDs and must be established to evaluate (5). Buntine (1991) defines a non-informative prior for ordinary Bayesian networks. As each joint outcome is equally likely for this prior, it ensures that equivalent networks are evaluated equally by the marginal likelihood. Under some additional assumptions, Heckerman et al (1995) showed that likelihood equivalence can also be achieved by deriving each  $\alpha_{ijl}$  from a prior network. The priors discussed in Friedman and Goldszmidt (1996) and Chickering et al (1997) extend this idea to structures based on compact representations of the CPDs. We define our prior by setting

$$\alpha_{ijl} = \frac{N}{r_j \cdot q_j} \cdot |S_{jl}|, \quad (6)$$

where  $q_j$  is with respect to the underlying DAG and  $|S_{jl}|$  denotes the number of configurations in that specific part. This non-informative prior can thus be considered an extension of the one used in Buntine (1991) which in turn is a special case of Heckerman et al (1995). The parameter  $N$ , known

as the equivalent sample size, reflects the strength of our prior belief on the parameter distributions. Its effect on the choice of Bayesian network structures has been investigated by Silander et al (2007).

The only remaining issue at this point is to define the prior distribution over the set of LDAGs. This part of (2) is generally not given too much attention in Bayesian model learning but for LDAGs it plays a vital role. A common approach is to assume a uniform prior and simply base the scoring function on the marginal likelihood alone. A uniform prior has been shown to work quite well for ordinary DAGs. Chickering et al (1997) use this prior as their main focus is to maximize the marginal likelihood rather than looking at criteria such as predictive performance or structural differences from a generative model. However, they also propose another prior that penalizes complexity of a model in terms of the number of free parameters,

$$p(G_L) \propto \kappa^{\dim(\Theta_{G_L})} = \prod_{j=1}^d \kappa^{\dim(\Theta_{G_L}(j))} \quad (7)$$

where  $\kappa \in (0, 1]$ . This approach is more in line with the one used by Friedman and Goldszmidt (1996) who use Kullback-Leibler divergence to further analyze the models chosen according to the scoring function.

The choice of model prior turns out to be an essential part of the Bayesian scoring function for LDAGs. We show in the result section that the marginal likelihood alone has a tendency to overfit the dependence structure for limited sample sizes by favoring dense graphs with complex labelings. The number of free parameters associated with such a LDAG is low compared to the number of free parameters associated with its underlying DAG and the LDAG is said to have a high CSI-complexity. The overfitting effect is thus reflected through a high CSI-complexity rather than an excessive number of free parameters. Although high CSI-complexity models may lead to high marginal likelihoods, they are more prone to contain false dependencies and thereby fail to capture the true global dependence structure. This has a direct negative effect on their out-of-sample predictive performance. Another drawback is that their high density will yield bulky CPDs in (1). This basically counteracts the fundamental idea of modularity on which the concept of graphical models is based.

The overfitting phenomenon vanishes asymptotically when  $n \rightarrow \infty$ , since maximization of the marginal likelihood leads to a consistent estimator of the model structure. Consequently, we construct our prior such that it acts as a regularizer for smaller sample sizes and its effect will gradually vanish as the sample size is increased,

$$p(G_L) \propto \kappa^{\dim(\Theta_G) - \dim(\Theta_{G_L})} = \prod_{j=1}^d \kappa^{\dim(\Theta_{G(j)}) - \dim(\Theta_{G_L(j)})}, \quad (8)$$

where  $\dim(\Theta_{G_L})$  and  $\dim(\Theta_G)$  are the number of free parameters associated with the LDAG and its underlying DAG, respectively. The parameter  $\kappa \in (0, 1]$  can be considered a measure of how strongly a CSI inducing label configuration must be supported by the data in order for it to be included in the model. For small values on  $\kappa$ , addition of a label configuration increases the score only if its associated CSI is firmly supported by the data while  $\kappa = 1$  corresponds to a uniform prior. This prior is similar to (7) but with the important distinction that the penalty degree is now determined by CSI-complexity rather than complexity in terms of number of free parameters which is implicitly restrained by the marginal likelihood. Regardless of the structure of the underlying DAG, all LDAGs with the same amount of label induced CSIs will thus have the same prior probability. This is motivated by the fact that we do not know the true global dependence structure, i.e. the underlying DAG. Instead we adjust our prior belief in how high degree of CSI-complexity the data is able to faithfully express without imposing false dependencies.

Our prior shares some desirable properties with the marginal likelihood (5). Given Markov equivalent underlying DAGs, all CSI-equivalent LDAGs are evaluated equally. When considering two equivalent LDAGs with non-equivalent underlying DAGs, the prior will favor the one with lower CSI-complexity. This is, however, the one to be preferred as it has a simpler interpretation. Another

---

**Algorithm 1** Procedure for optimizing the local CSI structure for  $X_j$

---

```

Procedure Optimize-Local-Structure(
     $X_j$ , // Variable whose local structure is optimized
     $X_{\Pi_j}$ , // Parental variables
     $\mathbf{X}$ , // A set of complete data over  $X_{\Pi_j \cup \{j\}}$ 
)
1:  $\mathcal{L}_j = \{\mathcal{L}_{(i,j)}\}_{i \in \Pi_j} \leftarrow \emptyset$ 
2:  $keepClimb \leftarrow True$ 
3: while  $keepClimb$ 
4:    $\mathcal{L}_j^{top} \leftarrow \mathcal{L}_j$ 
5:   for  $x_{L(i,j)} \notin \mathcal{L}_j : \{x_{L(i,j)} \cup \mathcal{L}_{(i,j)}\} \subset \mathcal{X}_{L(i,j)}$ 
6:      $\mathcal{L}_j^{cand} \leftarrow \mathcal{L}_j \cup x_{L(i,j)}$ 
7:     if  $p(\mathbf{X}_j | \mathbf{X}_{\Pi_j}, \mathcal{L}_j^{cand}) > p(\mathbf{X}_j | \mathbf{X}_{\Pi_j}, \mathcal{L}_j^{top})$ 
8:        $\mathcal{L}_j^{top} \leftarrow \mathcal{L}_j^{cand}$ 
9:     end
10:  end
11:  if  $\mathcal{L}_j < \mathcal{L}_j^{top}$ 
12:     $\mathcal{L}_j \leftarrow \mathcal{L}_j^{top}$ 
13:     $\mathcal{L}_j \leftarrow makeMaximal(\mathcal{L}_j)$ 
14:  else
15:     $keepClimb \leftarrow False$ 
16:  end
17: end
18: Return  $\mathcal{L}_j$ 

```

---

important property of (8) is that it decomposes variable-wise. From a computational perspective, this greatly enhances the efficiency of the search algorithm introduced later. On the downside, an unavoidable issue with an adjustable prior (or regularizer) is the task of determining the optimal value of some tuning parameter (in our case  $\kappa$ ). In the end of this section we propose a cross-validation-based method which allows us to choose among several values on  $\kappa$  before the actual model learning.

Given a scoring function, the task of learning an LDAG structure reduces to finding the model that maximizes the score given the data. This is, however, a very challenging problem since the model space is enormous. The number of DAGs for  $d$  variables grows super-exponentially with  $d$  (Robinson (1977)). In practice it is hence infeasible to calculate the posterior distribution (2) even for a small number of variables. Furthermore, this only covers ordinary DAGs and an expansion of the model space to include LDAGs will further increase the intractability of an exhaustive evaluation. Consequently, we need to apply some form of a search method. For this purpose we introduce a search algorithm which utilizes a non-reversible MCMC method, introduced and discussed by Corander et al (2006, 2008), combined with a direct form of optimization. The general idea is that the stochastic part of the algorithm jumps between neighbouring underlying DAGs, whose CSI structures are optimized by adding labels in a *greedy hill climbing*-manner. As our score decomposes variable-wise, instead of considering the whole DAG, we can optimize the local structure of one variable at a time. The procedure is described in Algorithm 1. For the score derived in the previous section, the termination of the algorithm occurs when the improvement of the marginal likelihood falls below the predetermined value of  $\kappa$ . Any deterministic optimization strategy similar to Algorithm 1 basically maps the set of DAGs onto a subset of regular maximal LDAGs  $\mathcal{G}_L^{opt} \subseteq \mathcal{G}_L$ . This will bring down the size of the model space explored by our MCMC method to the number of DAGs.

Various forms of MCMC are generally proposed for the Bayesian approach for learning the structural layer of probabilistic models. We utilize a non-reversible version which has been shown to possess several advantageous properties (Corander et al (2006, 2008)). Let  $q(\cdot|G_L)$  denote a generic proposal

distribution over the model space  $\mathcal{G}_L^{opt}$ , given  $G_L$  for all  $G_L \in \mathcal{G}_L^{opt}$ . We let  $G_L(t)$  denote the state of the chain at iteration  $t$ . At iteration  $t = 1, 2, \dots$  of the non-reversible chain,  $q(\cdot|G_L(t))$  is used to generate the next candidate state  $G_L^*$  which is then accepted with probability

$$\min \left( 1, \frac{p(G_L^*)p(\mathbf{X}|G_L^*)}{p(G_L(t))p(\mathbf{X}|G_L(t))} \right).$$

If  $G_L^*$  is accepted, we set  $G_L(t+1) = G_L^*$  and otherwise  $G_L(t+1) = G_L(t)$ . The proposal probabilities need not to be explicitly calculated or even known as long as they remain unchanged over the iterations and the resulting chain is irreducible. The stationary distribution of such a chain does no longer follow the posterior distribution (2). However, our main objective is, as previously stated, to identify only the maximum a posteriori model (3). The approximate solution proposed by a search chain at iteration  $t$  is simply the one with the highest score visited thus far. Satisfying the conditions mentioned, the proposal distributions are defined as uniform distributions over the globally adjacent LDAGs that can be reached by adding, reversing or removing a single edge under the restriction that the resulting LDAG is acyclic.

As the difference between two successive graphs may only differ for a single edge, at most two local structures are modified at each step of the chain. Since our score  $p(\mathbf{X}, G_L)$  decomposes variable-wise, only the modified local structures must be re-evaluated as the score for the rest of the variables remains unchanged. This idea can be further exploited when optimizing the local CSI-structures. At each step of the optimization procedure, we need only to re-evaluate the score with respect to the parts of the partition that are modified. For our algorithm in particular, only a single new part is created for each added label configuration.

Adding of labels yields a flexibility that facilitates the identification of "weaker" edges that might be deemed non-existing in the model space of DAGs. However, optimization of the CSI-structure cannot make up for unrealistic global independence assumptions made by an inferior underlying DAG structure. Hence, a prerequisite for learning a good LDAG structure is that it is based on a sensible underlying DAG. Getting stuck at regions with inferior underlying DAGs, will have a more severe negative effect on the learned LDAGs than not finding the optimal CSI-structure. This motivates the fact that the stochastic part of our method performs global changes whereas the optimization of the CSI-structures is done in a deterministic manner.

To finally attend the problem of choosing an appropriate value of  $\kappa$ , we propose a cross-validation scheme that allows us to assess a set of candidate values. First we partition the data  $\mathbf{X}$  into a training set  $\mathbf{Y}$  and a test set  $\mathbf{Z}$ . We then apply our search method on the training data under some prior (or  $\kappa$ ) and identify the optimal model  $G_L^\kappa$ . We then evaluate the learned model's ability to predict the test data by calculating the posterior predictive probability of the test data given the training data,

$$p(\mathbf{Z} | \mathbf{Y}, G_L^\kappa) = \int_{\theta \in \Theta_{G_L^\kappa}} p(\mathbf{Z} | G_L^\kappa, \theta) \cdot f(\theta | \mathbf{Y}, G_L^\kappa) d\theta. \quad (9)$$

This integral is similar to (4) but the parameter vectors are now weighted with respect to the posterior distributions updated according to the training data. Under similar assumptions made earlier, (9) can be calculated analytically by

$$p(\mathbf{Z} | \mathbf{Y}, G_L^\kappa) = \prod_{j=1}^d \prod_{l=1}^{k_j} \frac{\Gamma(\sum_{i=1}^{r_j} (\alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl})))}{\Gamma(n_{\mathbf{Z}}(S_{jl}) + \sum_{i=1}^{r_j} (\alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl})))} \prod_{i=1}^{r_j} \frac{\Gamma(n_{\mathbf{Z}}(x_{ji} \times S_{jl}) + \alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl}))}{\Gamma(\alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl}))}, \quad (10)$$

where the bold case index indicates to which data set the outcome count refers. To reduce the variability of the method, multiple partitions of  $\mathbf{X}$  are created,  $\{(\mathbf{Y}_1, \mathbf{Z}_1), (\mathbf{Y}_2, \mathbf{Z}_2), \dots, (\mathbf{Y}_M, \mathbf{Z}_M)\}$ ,

$\kappa$	$\log p(\mathbf{X}, G_L)$	$ E $	$\dim(\Theta_G)$	$\dim(\Theta_{G_L})$	$\rho_{pred}$
0.001	-6731.82	5	12	12	-671.30
0.1	-6729.69	6	14	12	-670.88
<b>0.3</b>	<b>-6727.50</b>	<b>6</b>	<b>14</b>	<b>12</b>	<b>-670.51</b>
0.5	-6724.68	7	18	11	-670.89

Table 1: Properties of identified LDAGs for coronary heart disease data.

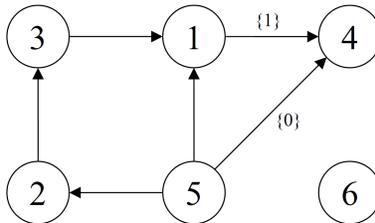


Figure 11: Optimal LDAG for coronary heart disease data.

and the validation results are averaged according to

$$\rho_{pred}(\kappa) = \frac{1}{M} \sum_{m=1}^M \log p(\mathbf{Z}_m | \mathbf{Y}_m, G_L^{\kappa, m}). \quad (11)$$

The value on  $\kappa$  is finally chosen among the candidates as the one that maximizes (11).

## 4 Experimental results with real and simulated data sets

To illustrate the properties of LDAGs, we apply our search algorithm on both a real and two simulated data sets. First we consider a real data set that has been thoroughly investigated in earlier graphical modelling literature. After that we consider synthetic DAG- and LDAG-based models. Throughout this section we set the equivalent sample size  $N = 1$ . By keeping  $N$  constant, we instead focus on investigating how different values on  $\kappa$  will affect the ability of the learned graph to predict the data structure. The learning algorithm was executed for  $\kappa \in \{0.001, 0.1, 0.3, 0.5\}$ . The optimal value on  $\kappa$  is chosen according to the cross-validation scheme described earlier. To create multiple partitions, the data is split up in to ten parts of the same size and each part is successively chosen as test set. For each search, we initiated 50 parallel independent search chains. The empty graph was set as the initial state of each chain and number of iterations was set to 500. The optimal graph was then simply identified as the one with the highest score visited by any of the chains.

Our real data set contains 1841 cases composed of six binary risk factors for coronary heart disease (Whittaker (1990)). The meanings of the variables are explained in Table 2 (Appendix). In Table 1 the structural properties of the graphs, identified for different values on  $\kappa$ , are listed along with their scores. Here we get an indication of how the CSI-complexity increases with higher values on  $\kappa$ . The bold font indicate which  $\kappa$  was chosen as optimal by the cross-validation procedure. The corresponding LDAG is illustrated in Figure 11. The LDAG identified for  $\kappa = 0.001$  contains no labels and is thereby equal to its underlying DAG. The improvement, that an added label configuration induces to the marginal likelihood, is overshadowed by the simultaneous lowering of the prior probability mass. Consequently, when  $\kappa \rightarrow 0$  the direct optimization will map the set of DAGs onto itself and the learning procedure is reduced to a search among ordinary DAGs.

We now consider synthetic models from which data are generated to systematically compare models identified for different prior distributions and sample sizes. Since we know the generating model, we investigate how well the identified models approximate the true distribution. The CPDs of the

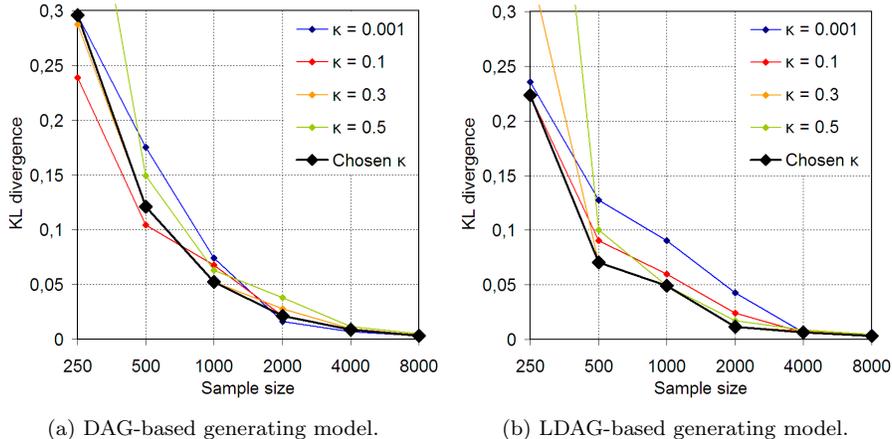


Figure 12: KL divergence for different sample sizes under different priors.

models are estimated by the consistent mean a posteriori estimator as the expected value of the local posterior Dirichlet distributions. To compare the distributions, we utilize the concept of Kullback-Leibler (KL) divergence. Let  $p$  denote the real distribution over  $X = (X_1, \dots, X_d)$  and let  $p^*$  denote an approximation of  $p$ . The KL divergence between the distributions is defined by

$$D_{KL}(p \parallel p^*) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{p^*(x)}.$$

The KL divergence is a non-negative non-symmetric measure of the distance that is equal to zero only if  $p = p^*$ . Under the assumption that no incorrect independence assumptions are made by the model, then  $D_{KL}(p \parallel p^*) \rightarrow 0$  as the sample size  $n \rightarrow \infty$ . In addition to the KL divergence, we also investigate how well the identified LDAGs capture the global dependence structure, i.e. the underlying DAG.

We generated data according to a DAG as well as an LDAG. The LDAG was created by adding labels to the initial DAG. The DAG and labels are illustrated in Figure 2 (Appendix). To generate data according to the DAG, each CPD were randomly drawn from a uniform distribution. Similar CPDs may have arisen by chance but no local CSIs were explicitly included. To generate data according to the LDAG, some of the CPDs were set identical in order to satisfy the labels. We let the sample size  $n$  range from 250 to 8000. For each data set, we executed the learning procedure described earlier. Our results are summarized in Table 3 (Appendix) for the DAG model and in Table 4 (Appendix) for the LDAG model.

As expected, the model distributions approach the true distribution when the sample size increases. This results in a steady improvement of the KL divergence as illustrated in Figure 12. The decrease is evident for all values on  $\kappa$  but our results indicate that different prior distributions are to be preferred depending on the sample size. It also clear how the quality of most of the models begin to suffer under  $\kappa = 0.5$  as a result of overfitting. The reduced out-sample-performance of the models prevents the prior from being picked even once. During the simulations it became evident that the overfitting effect further escalated under even less restrictive priors. On the whole, our procedure for picking the optimal  $\kappa$  performs well. The thick black curve in Figure 12 represents the models chosen by the cross-validation. Ideally, this curve should stay below the other curves. For the DAG-based model (Figure 12a), the curve never diverges too far from the optimal choice. For the LDAG-based model (Figure 12b), our procedure performs very well by always picking the optimal prior from the candidate set.

As we can see from Table 3 and 4, all models identified under  $\kappa = 0.001$  are without labels since  $\dim(\Theta_G) - \dim(\Theta_{G_L}) = 0$ . We can thus use this prior as a reference point for investigating how well

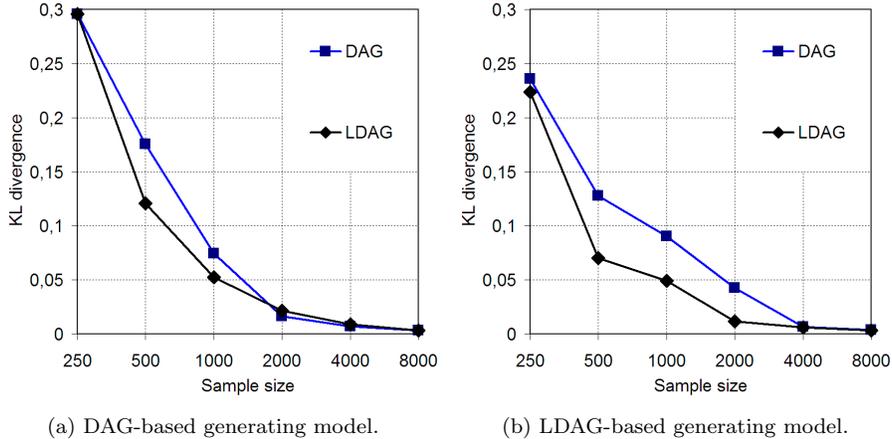


Figure 13: Comparison of DAGs and LDAGs for different sample sizes.

LDAGs perform compared to traditional DAGs. Figure 13 illustrates the difference in KL divergence between the true distribution and the approximate distributions induced by the models. The DAG curve in the figure corresponds to the 0.001-curve from Figure 12 and the LDAG curve corresponds to the thick black curve where the models were chosen by the initial cross-validation method. Note that the method in some cases picks the 0.001-prior which results in a converging of the curves. We see that the LDAGs mostly outperform traditional DAGs by inducing distributions that better approximate the true distribution. This is especially clear for the medium sized samples, even when the generating model does not contain any explicit CSIs. Consequently, this seems to be the range where the models have the most to gain from adding labels. The samples are too small for discovering the true DAG structure without labels yet large enough for the structure learning to be stable even under less restrictive priors. For large enough sample sizes the two curves will eventually converge. This is a natural and inevitable phenomenon that is illuminated when investigating the structure of the underlying DAG.

If we consider the result tables (Table 3 and 4), we see that the point of convergence between the curves coincides with the sample size at which the correct underlying DAG is identified under  $\kappa = 0.001$ , i.e. without labels. When the generating model is based on a DAG, adding labels to the correct underlying DAG will induce restrictions on the corresponding approximate distribution that is not satisfied by the true distribution. When the estimation of the parameters become stable enough, the gain from having to estimate fewer parameters cannot longer outweigh the inaccuracies of the additional restrictions. When the generating model contains explicit CSIs, the DAG curve does not overtake the LDAG curve for any of the considered sample sizes. The DAG curve will eventually catch up with the LDAG curve when  $n \rightarrow \infty$  but the DAG model will, however, require some redundant parameters. Finally, the result tables also illustrate how adding labels facilitates the discovery of the true global dependence structure in the sense that LDAGs require less data to reach the correct underlying DAG compared to traditional DAGs. The flexibility of LDAGs provides an advantage over traditional DAGs in terms of structure learning, since it allows representation of more complex models with fewer parameters. However, the same flexibility may also cause overfitting if not properly regulated in the learning process.

## 5 Discussion

We have further developed the idea of incorporating context-specific independence in directed graphical models by introducing a graphical representation in form of a labeled directed acyclic graph. We have

shown that an LDAG is general in its representation of local CSIs as well as it is able to visualize complex dependence structures as a single entity. We also investigated properties of LDAGs in terms of model interpretability and identifiability by introducing the class of maximal regular LDAGs and the notion of CSI-equivalence.

In terms of structure learning, we have derived an LDAG-based Bayesian score and an MCMC-based search method that combines stochastic global changes with deterministic local changes. Our experimental results agree with previous research in the sense that incorporation of CSI in the learning phase improves model quality. However, we also noted that an appropriate prior must be used for optimal performance.

An interesting extension to LDAGs could be to allow the local dependence structures to go beyond CSI in a way that can still be expressed through some form of labels. It would also be interesting to carry out a more extensive simulation study in which one could compare alternative search methods as well as compare LDAGs to other existing models.

## References

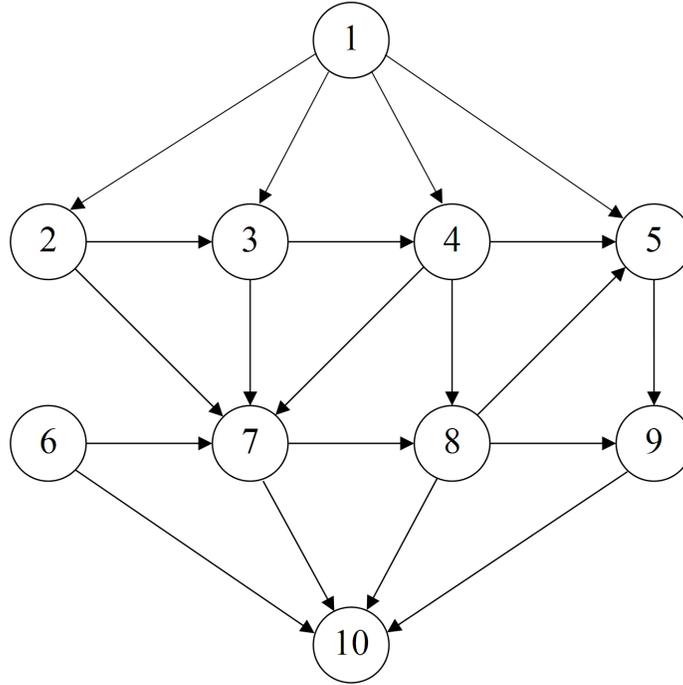
- Andersson S, Madigan D, Perlman M (1997) A characterization of Markov equivalence classes for acyclic digraphs. *Annals of statistics* 25:505–541
- Boutilier C, Friedman N, Goldszmidt M, Koller D (1996) Context-specific independence in Bayesian networks. In: Horvitz E, Jensen F (eds) *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp 115–123
- Buntine W (1991) Theory refinement on Bayesian networks. In: D’Ambrosio B, Smets P, Bonissone P (eds) *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp 52–60
- Chickering DM, Heckerman D, Meek C (1997) A Bayesian approach to learning Bayesian networks with local structure. In: *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann
- Cooper G, Herskovitz E (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9:309–347
- Corander J (2003) Labelled graphical models. *Scandinavian Journal of Statistics* 30:493–508
- Corander J, Gyllenberg M, Koski T (2006) Bayesian model learning based on a parallel MCMC strategy. *Statistics and Computing* 16:355–362
- Corander J, Ekdahl M, Koski T (2008) Parallel interacting MCMC for learning of topologies of graphical models. *Data Mining and Knowledge Discovery* 17:431–456
- Eriksen PS (1999) Context specific interaction models. Tech. rep., Aalborg University
- Friedman N, Goldszmidt M (1996) Learning Bayesian networks with local structure. In: Horvitz E, Jensen F (eds) *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp 252–262
- Geiger D, Heckerman D (1996) Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence* 82:45–74
- Heckerman D (1991) *Probabilistic Similarity networks*. MIT Press, Cambridge, MA
- Heckerman D, Geiger D, Chickering D (1995) Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20:197–243

- Højsgaard S (2003) Split models for contingency tables. *Computational Statistics & Data Analysis* 42:621–645
- Højsgaard S (2004) Statistical inference in context specific interaction models for contingency tables. *Scand J Stat* 31:143–158
- Koller D, Friedman N (2009) *Probabilistic Graphical Models: Principles and Techniques*. MIT Press
- Koski T, Noble J (2009) *Bayesian Networks, An Introduction*. Wiley, Chippenham, Wiltshire
- Pagallo G, Haussler D (1990) Boolean feature discovery in empirical learning. *Machine Learning* 5:71–99
- Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco
- Poole D (1997) Probabilistic partial evaluation: Exploiting rule structure in probabilistic inference. In: *Proc. 15th International Joint Conference on Artificial Intelligence*, pp 1284–1291
- Poole D, Zhang N (2003) Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence Research* 18:263–313
- Robinson R (1977) Counting unlabelled acyclic digraphs. *Springer Lecture Notes in Mathematics: Combinatorial Mathematics V* 622:28–43
- Silander T, Kontkanen P, Myllymäki P (2007) On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In: Parr R, van der Gaag L (eds) *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp 360–367
- Whittaker J (1990) *Graphical Models in Applied Multivariate Statistics*. Wiley, New York
- Zhang N, Poole D (1999) On the role of context-specific independence in probabilistic inference. In: *Proc. 16th International Joint Conference on Artificial Intelligence*, pp 1288–1293

## Appendix

Variable	Definition	Outcomes
$X_1$	Smoking	No := 0, Yes := 1
$X_2$	Strenuous mental work	No := 0, Yes := 1
$X_3$	Strenuous physical work	No := 0, Yes := 1
$X_4$	Systolic blood pressure	< 140 := 0, > 140 := 1
$X_5$	Ratio of $\beta$ and $\alpha$ lipoproteins	< 3 := 0, > 3 := 1
$X_6$	Family anamnesis of coronary heart disease	No := 0, Yes := 1

Table 2: Description of the variables in coronary heart disease data.



$$\begin{array}{ll}
 \mathcal{L}_3 : & \mathcal{L}_{(2,3)} = \{0\} \\
 \mathcal{L}_4 : & \mathcal{L}_{(1,4)} = \{1\} \\
 \mathcal{L}_5 : & \mathcal{L}_{(4,5)} = \{(0, *)\} \\
 & \mathcal{L}_{(8,5)} = \{(0, *)\} \\
 \mathcal{L}_7 : & \mathcal{L}_{(2,7)} = \{(1, 1, 0)\} \\
 & \mathcal{L}_{(3,7)} = \{(0, 1, 1), (1, *, 1)\} \\
 & \mathcal{L}_{(4,7)} = \{(1, 1, *)\} \\
 & \mathcal{L}_{(6,7)} = \{(1, 1, *)\} \\
 \mathcal{L}_9 : & \mathcal{L}_{(5,9)} = \{1\} \\
 \mathcal{L}_{10} : & \mathcal{L}_{(7,10)} = \{(1, *, *)\} \\
 & \mathcal{L}_{(8,10)} = \{(1, *, *)\} \\
 & \mathcal{L}_{(9,10)} = \{(1, *, *)\}
 \end{array}$$

Figure 14: DAG and labels according to which the synthetic data sets were generated.

$n$	$\kappa$	$\log p(\mathbf{X}, G_L)$	$ E $	$\dim(\Theta_G) - \dim(\Theta_{G_L})$	$D_{KL}$	$\rho_{pred}(\kappa)$	
250	<b>0.001</b>	<b>-1478.76</b>	<b>13</b>	<b>0</b>	<b>0.2956</b>	<b>-146.33</b>	
	0.1	-1477.13	14	4	0.2387	-146.96	
	0.3	-1467.68	15	18	0.2873	-150.14	
	0.5	-1451.83	20	35	0.5009	-157.86	
500	0.001	-2855.23	17	0	0.1755	-277.52	
	0.1	-2832.88	19	15	0.1045	-276.69	
	<b>*</b>	<b>0.3</b>	<b>-2813.80</b>	<b>20</b>	<b>24</b>	<b>0.1209</b>	<b>-275.13</b>
	0.5	-2805.54	21	32	0.1492	-280.19	
1000	0.001	-5587.31	18	0	0.0742	-546.52	
	0.1	-5569.53	19	13	0.0676	-544.18	
	<b>*</b>	<b>0.3</b>	<b>-5553.73</b>	<b>20</b>	<b>24</b>	<b>0.0523</b>	<b>-544.13</b>
	0.5	-5541.17	23	33	0.0632	-550.40	
2000 *	0.001	-11132.33	20	0	0.0162	-1095.74	
	<b>*</b>	<b>0.1</b>	<b>-11097.82</b>	<b>20</b>	<b>18</b>	<b>0.0215</b>	<b>-1095.68</b>
	<b>*</b>	0.3	-11073.52	20	23	0.0279	-1097.86
		0.5	-11062.23	24	38	0.0379	-1098.94
4000 *	0.001	-21988.57	20	0	0.0068	-2177.83	
	<b>*</b>	<b>0.1</b>	<b>-21957.06</b>	<b>20</b>	<b>14</b>	<b>0.0088</b>	<b>-2177.49</b>
	<b>*</b>	0.3	-21940.76	20	15	0.0096	-2177.77
	<b>*</b>	0.5	-21929.60	20	17	0.0114	-2178.32
8000 *	<b>0.001</b>	<b>-43822.03</b>	<b>20</b>	<b>0</b>	<b>0.0034</b>	<b>-4357.81</b>	
	<b>*</b>	0.1	-43796.27	20	11	0.0044	-4358.73
	<b>*</b>	0.3	-43784.18	20	11	0.0044	-4361.13
		0.5	-43777.35	21	15	0.0051	-4361.58

Table 3: Properties of identified LDAGs for DAG data.

$n$	$\kappa$	$\log p(\mathbf{X}, G_L)$	$ E $	$\dim(\Theta_G) - \dim(\Theta_{G_L})$	$D_{KL}$	$\rho_{pred}(\kappa)$	
250	0.001	-1468.04	13	0	0.2357	-142.08	
	<b>0.1</b>	<b>-1463.54</b>	<b>13</b>	<b>4</b>	<b>0.2237</b>	<b>-141.93</b>	
	0.3	-1451.55	16	15	0.3406	-146.73	
	0.5	-1436.93	23	47	0.7015	-149.02	
500	0.001	-2943.88	13	0	0.1277	-286.72	
	0.1	-2934.72	14	12	0.0903	-286.34	
	<b>0.3</b>	<b>-2917.84</b>	<b>17</b>	<b>18</b>	<b>0.0703</b>	<b>-284.12</b>	
	0.5	-2906.13	21	41	0.1001	-284.59	
1000	0.001	-5739.70	16	0	0.0905	-565.60	
	0.1	-5725.86	16	15	0.0600	-565.79	
	<b>0.3</b>	<b>-5705.51</b>	<b>19</b>	<b>25</b>	<b>0.0490</b>	<b>-564.70</b>	
	0.5	-5692.74	19	25	0.0490	-567.23	
2000	0.001	-11377.83	17	0	0.0428	-1122.61	
	0.1	-11328.02	18	18	0.0243	-1121.34	
	*	<b>0.3</b>	<b>-11305.91</b>	<b>20</b>	<b>27</b>	<b>0.0116</b>	<b>-1118.74</b>
	0.5	-11294.84	23	45	0.0171	-1119.03	
4000 *	0.001	-22659.45	20	0	0.0065	-2245.44	
	*	<b>0.1</b>	<b>-22600.54</b>	<b>20</b>	<b>23</b>	<b>0.0063</b>	<b>-2243.25</b>
	*	0.3	-22573.11	20	25	0.0078	-2243.89
	*	0.5	-22560.28	20	26	0.0089	-2244.78
8000 *	0.001	-44862.27	20	0	0.0036	-4462.29	
	*	<b>0.1</b>	<b>-44803.87</b>	<b>20</b>	<b>22</b>	<b>0.0031</b>	<b>-4462.08</b>
	*	0.3	-44778.38	20	24	0.0046	-4462.43
	*	0.5	-44766.12	20	24	0.0046	-4463.07

Table 4: Properties of identified LDAGs for LDAG data.