Proceedings of Machine Learning Research 343:473–496, 2026 Probabilistic Graphical Models (PGM)

Bayesian estimation of causal effects in Bayesian networks with local structure

Vera Kvisgaard VERAHK@UIO.NO
Department of Mathematics
University of Oslo
Oslo, Norway

Johan Pensar JOHANPEN@UIO.NO
Department of Mathematics
University of Oslo
Oslo, Norway

Editors: Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

Abstract

We consider the problem of estimating causal effects from observational categorical data when the data-generating process can be modeled by an unknown causal Bayesian network with local dependence structure in its conditional distributions. Capturing local structure in a categorical Bayesian network may significantly reduce the number of parameters and facilitate the discovery of weak dependencies and edge orientation, which is crucial for causal inference. We extend the Bayesian IDA (BIDA) framework to Bayesian networks with local structure, focusing on a specific class of local independencies called context-specific independencies (CSI). To this end, we build upon previous work that combines Markov chain Monte Carlo methods over DAGs with deterministic greedy optimisation of CSI-trees to learn the full network structure, and propose a local approach for efficient estimation of pairwise causal effects in a given structure, where we apply the same greedy optimisation procedure to infer the conditional distributions involved in the backdoor formula. We evaluate the proposed method in a simulation study in which we simulate data from synthetic Bayesian networks with local structure, as well as a set of benchmark networks.

Keywords: Bayesian networks, Causal inference, MCMC, Context-specific independence

1. Introduction

Methods for detecting causal relationships from observational data have received much attention in recent decades. In particular, many of these causal discovery methods have been developed under the framework of causal Bayesian networks. A Bayesian network is a probabilistic graphical model based on a directed acyclic graph (DAG) in which the nodes correspond to random variables and the edges correspond to probabilistic dependencies between them (Koller and Friedman, 2009, Ch. 3). When given a causal interpretation the edges are taken to represent not only probabilistic dependencies but also the direct causal relationships that produce these dependencies. The model can then be used to predict the (causal) effect of external interventions on the system (Spirtes et al., 1993; Pearl, 2009).

The DAG describes a set of conditional independencies that form the *global* structure of the model; the joint distribution factorises into a set of conditional probability distributions (CPDs) accordingly, facilitating efficient learning and inference. Several extensions to the

© 2026 V. Kvisgaard & J. Pensar.

standard Bayesian network model have been proposed to exploit additional *local* structure in the CPDs (Koller and Friedman, 2009, Ch. 5), including models of context-specific independence (CSI) and related notions of partial independence that hold only in specific contexts (Pensar et al., 2016). Modelling local independence structure in the network can reduce the number of parameters, facilitate the discovery of weak dependencies, and, in theory, improve edge orientation in the inferred DAG, which is crucial in the causal setting.

However, most of the previous work on learning Bayesian networks with local independence structure have focused on improving the predictive accuracy of the model in the purely probabilistic setting (Boutilier et al., 1996; Chickering et al., 1997; Friedman and Goldszmidt, 1998; Pensar et al., 2015, 2016). Here, we study how local independence structures can be used to improve the estimation of causal effects and extend the Bayesian IDA (BIDA, Pensar et al. (2020)) method to categorical Bayesian networks with context-specific independencies in the local CPDs. More specifically, our aim is to discover strong causal relationships from observational categorical data, by estimating the causal effect between every cause-effect pair of variables, that is, the effect on X_j after intervening on X_i, for every ordered pair (i, j). We assume that the data are generated by an unknown causal Bayesian network with local structure and that there are no unobserved confounders.

Similar to the original IDA approach, first presented in Maathuis et al. (2009), BIDA decouples inference about the causal structure from inference about the causal effects in a given structure. More concretely, it combines Bayesian model averaging over causal structures with independent, local estimation of each pairwise causal effect given a structure, via the backdoor formula (Pearl, 2009). The resulting posterior can be seen as a scalable approximation of that produced by more conventional Bayesian approaches, in which inference about model and parameters is performed jointly (Kvisgaard and Pensar, 2025).

The main challenges posed by the additional local CSI structure is increased model flexibility and a considerably enlarged model space, which already increases super-exponentially in the number of nodes for standard DAG-based models. To reduce the computational complexity, we adapt the approach of Pensar et al. (2015) and combine stochastic sampling of DAGs with deterministic greedy optimisation of the local structure in a given DAG. Pairing each DAG with a single local structure, the size of the resulting model space equals that of standard DAGs. Moreover, we propose two different backdoor estimators for local estimation of causal effects, each of which infers CSI-structured CPDs based on the same greedy procedure.

We compare the proposed BIDA methods to the original BIDA in a simulation study and find that, given sufficient data, they are more accurate in discovering the strong effects in synthetic networks with local CSI structure. However, the relative performance is overall more mixed on a set of benchmark networks without explicit CSI structure, highlighting a tradeoff between model flexibility and learnability that depends on the degree to which the true model exhibits (approximate) CSI structure.

The remainder of the paper is organised as follows. Section 2 reviews Bayesian networks, CSI and the original BIDA approach. Section 3 describes our procedure for joint DAG sampling and CSI optimisation, as well as the proposed local estimators for causal effects. Section 4 presents the results from our simulation experiments. Finally, Section 5 concludes with a discussion and directions for future work.

2. Background

In this section, we first review the causal Bayesian network model and the BIDA approach for inferring pairwise causal effects in this model, before extending the causal model with local independence structure in the form of context-specific independencies (CSI). Throughout the paper, we consider a set of random categorical variables $X_V = \{X_1, \ldots, X_n\}$, with index set $V = \{1, \ldots, n\}$. For any subset $A \subseteq V$, we use x_A to denote a particular instance and $\mathcal{X}_A$ to denote the joint outcome space of X_A.

2.1. Causal Bayesian networks

A Bayesian network for X_V is a pair (G, θ), where G is a directed acyclic graph (DAG) that represents the model structure and θ is a set of parameters that specify the joint distribution over X_V (Koller and Friedman, 2009, Ch. 3). The nodes in the DAG correspond to the variables, and, under a causal interpretation, the edges represent direct causal relationships between them (Spirtes et al., 1993; Pearl, 2009). These causal relationships produce a set of conditional independencies in the distribution over X_V, so that the joint distribution factorises according to the Markov condition:[1]

$$p(x_V \mid G, \theta) = \prod_{i=1}^{n} p(x_i \mid G, \theta_{X_i|x_{G_i}}, x_{G_i}), \tag{1}$$

where G_i are the indices of the parents of node i and $\theta_{X_i|x_{G_i}}$ are the parameters of the conditional probability distribution (CPD) of X_i given its parents.

In the categorical setting, $\theta_{X_i|x_{G_i}}$ is the set of event probabilities $\{\theta_{x_i|x_{G_i}}\}$ for each value of x_i. We let $\theta_{X_i|X_{G_i}}$ denote the full conditional probability table (CPT). The model parameters are then the collection of CPTs for each node, $\theta = \{\theta_{X_i|X_{G_i}}\}_{i=1}^{n}$. For any disjoint subsets $A, B \subseteq V$, we write $\theta_{x_A|x_B}$ for the probability $p(x_A \mid x_B, G, \theta)$ implied by (1), even if the conditional distribution is not explicitly represented by the model parameters.

The causal Bayesian network model in (1) can be used to predict the effect of external interventions on the considered system. We use the do-operator $do(X_i = x_i')$, or simply $do(x_i')$, to denote that X_i is set to the value x_i' by an intervention that, importantly, leaves all other variables in the model undisturbed. The intervention induces a new distribution over the variables, whose marginals can be expressed by the backdoor formula (Pearl, 2009):

$$p(x_j \mid G, \theta, do(x_i)) = \begin{cases} \theta_{x_j}, & \text{if } X_j \in X_{G_{ij}}, \\ \sum_{x_{G_{ij}}} \theta_{x_j|x_i, x_{G_{ij}}} \theta_{x_{G_{ij}}}, & \text{otherwise.} \end{cases} \tag{2}$$

Here G_{ij} is a backdoor adjustment set with respect to cause-effect pair (i, j) and DAG G.[2] The right-hand side of the formula involves only parameters of the pre-intervention distribution, which can be consistently estimated from observational data, assuming causal sufficiency (no unobserved confounders) and that the distribution is strictly positive.

1. Under the operating assumption that there are no latent confounders.
2. Following Pearl's original backdoor criterion (2009, Definition 3.3.1), G_{ij} is a valid adjustment set if: (i) it contains no descendants of node i; and (ii) it d-separates node i and j in the backdoor graph (the graph obtained from G by removing all edges that go out of node i).

To summarise the causal effect of X_i on X_j in a single quantity τ_{ij} in the categorical setting, we use the Jensen-Shannon divergence (JSD, Lin (1991); Griffiths et al. (2015)) between the relevant intervention distributions:

$$\tau_{ij} = \frac{1}{r_i} \sum_{x_i} \sum_{x_j} \pi_{x_j|x_i} \log \frac{\pi_{x_j|x_i}}{\frac{1}{r_i} \sum_{x_i} \pi_{x_j|x_i}}, \tag{3}$$

where $\pi_{x_j|x_i} = p(x_j \mid G, \theta, do(x_i))$ are parameters denoting the interventional distribution in (2), which, notably, are functions of the model parameters, θ.

2.2. The Bayesian IDA method

Suppose we are given a set of data, D, consisting of samples $x_V^{(1)}, \ldots, x_V^{(N)}$ drawn from a causal Bayesian network model $p(x \mid G, \theta)$, where both the DAG, G, and the model parameters, θ, are unknown. The Bayesian IDA (BIDA) builds a posterior distribution over the causal effect τ_{ij} for every cause-effect pair $(i, j) \in V \times V$, under the assumptions of causal sufficiency and positivity, by combining local backdoor estimation of causal effects given a causal DAG with Bayesian model averaging over DAGs.

Inspired by the original IDA method, first presented in Maathuis et al. (2009), each pairwise causal effect is estimated locally, considering only the conditional distributions involved in the backdoor formula. More specifically, given a DAG G in which G_{ij} is the backdoor adjustment set of some pre-specified class, it places a prior directly over the set of parameters θ_{ij} on the right-hand side of the formula (2):

$$p(\theta_{ij} \mid G_{ij}) = \begin{cases} p(\theta_{X_j}), & \text{if } j \in G_{ij}, \\ p(\theta_{X_{G_{ij}}}) \prod_{x_i} \prod_{x_{G_i}} p(\theta_{X_j|x_i, x_{G_{ij}}}), & \text{otherwise.} \end{cases} \tag{4}$$

Assuming a conjugate Dirichlet prior for each parameter vector on the right-hand side of the above equation, the observed sample counts over $(X_i, X_j, X_{G_{ij}})$ are sufficient statistics for computing posterior updates. The conditional target posterior, $p(\tau_{ij} \mid G_{ij}, D)$, follows from the backdoor formula (2) and the JSD (3), and can be approximated by Monte Carlo sampling.

The local parameter estimation approach avoids computationally demanding inference in the full causal model (1). Furthermore, model averaging over DAGs can be replaced with averaging over adjustment sets, since the posterior $p(\tau_{ij} \mid G_{ij}, D)$ is identical given any DAG in which G_{ij} is the adjustment set. Consequently, the resulting BIDA posterior can be expressed as a mixture distribution:

$$p(\tau_{ij} \mid D) \;=\; \sum_{G_{ij}} p(\tau_{ij} \mid G_{ij}, D) p(G_{ij} \mid D). \tag{5}$$

The mixture weight $p(G_{ij} \mid D) = \sum_{G:\, G_{ij}} p(G \mid D)$ is the posterior probability of G_{ij} being the target adjustment set (with respect to the pre-specified class), equal to the aggregated posterior probability of all causal structures G in which G_{ij} is the adjustment set. Under standard modularity constraints on the likelihood for G, the weights can be computed exactly in smaller domains (Pensar et al., 2020, using parents for adjustment) or approximated with existing MCMC schemes in the space of DAGs (Kvisgaard and Pensar, 2025).

While the local estimation approach in effect decouples inference about the causal effects from inference about the causal structure, the BIDA posterior (5) can be seen as a scalable approximation to more conventional Bayesian approaches where inference about the model and its parameters is performed jointly (see Kvisgaard and Pensar (2025) for a more detailed discussion).

2.3. Local independence structure: Context-specific independence

The conditional independencies encoded in the DAG G are global in the sense that they hold for every configuration of the variables involved. There may be additional independencies of a more local nature in the distribution, which can be exploited to reduce the number of parameters needed to specify the model, and thus to improve the efficiency of inference and learning in the model (Koller and Friedman, 2009; Boutilier et al., 1996; Chickering et al., 1997; Friedman and Goldszmidt, 1998; Pensar et al., 2015, 2016).

In this work, we focus on a specific class of local independencies called context-specific independence (CSI), formally defined below.

Definition 1 (Context-specific independence, Boutilier et al., 1996, Def. 2.2) *Let X_A, X_B, and X_D be disjoint sets of random variables and let X_C be a set of variables with a particular assignment x_C. We say that X_A is context-specifically independent of X_B given X_D in context $X_C = x_C$, denoted by*

$$X_A \perp X_B \mid X_D, x_C,$$

if for all values $x_A \in \mathcal{X}_A$, $x_B \in \mathcal{X}_B$, and $x_D \in \mathcal{X}_D$ such that $p(x_B, x_D, x_C) > 0$,

$$p(x_A \mid x_B, x_D, x_C) \;=\; p(x_A \mid x_D, x_C).$$

A particularly convenient way to represent CPDs with CSI structure is through decision trees, where each node has at most one parent and each internal node has one outgoing edge for every possible value of its associated variable (these are also called CSI-trees in Pensar et al. (2016)). The leaves of the tree partition the outcome space of the conditioning variables, and the distribution of the response variable is constant within each partition part. Variables that do not appear in the tree are irrelevant for the induced partition, and thereby independent of the response conditional on the variables included. We say that a CSI-tree is *regular* if every conditioning variable is associated with at least one internal node and (in analogy with the regularity condition for labelled DAGs in Pensar et al. (2015)). Figure 1 in Appendix A illustrates a CSI-tree and the associated CPT.

Formally, we define a Bayesian network with local CSI-structure as the pair $(G_{\mathcal{L}}, \theta)$, where $\mathcal{L} = \{\mathcal{L}_{G_i}\}_{i=1}^n$ and $\mathcal{L}_{G_i}$ is a partition of the parent space $\mathcal{X}_{G_i}$ for each node i, such that the distribution of X_i given its parents is constant within each part $\ell_{G_i} \in \mathcal{L}_{G_i}$, and thus

$$\theta_{X_i|x_{G_i}} = \theta_{X_i|\ell_{G_i}} \quad \text{for all } x_{G_i} \in \ell_{G_i}.$$

The structural component of the model $G_{\mathcal{L}} = (G, \mathcal{L})$ then describes both the global and local independence structure of the model, and the model parameters $\theta = \{\{\theta_{X_i|\ell_{G_i}}\}_{\ell_{G_i} \in \mathcal{L}_{G_i}}\}_{i=1}^n$ are the parameters of every unique CPD in the model.

3. A Bayesian IDA method for Bayesian networks with local structure

In this section, we extend the Bayesian IDA (BIDA) method for categorical data (Pensar et al., 2020; Kvisgaard and Pensar, 2025) to the setting of Bayesian networks with local structure. The main challenges from extending the model with local structure is the increased model flexibility and the vast model space. To reduce the computational complexity, our key idea is to combine a stochastic MCMC scheme for sampling the underlying DAG with a deterministic greedy optimisation routine for inferring the local dependence structure, adapting the approach of Pensar et al. (2015) in their search for the single best scoring labelled DAG. Each DAG is then associated with a single optimised local structure, and the resulting model space has the same dimensions as the space of DAGs. As in the original BIDA setting, we can then decouple inference about the model structure from inference about the causal effects in a given structure. In the following, we consider each of the two components in turn.

3.1. Posterior over the causal structure

Given data D, let $\mathcal{L}^*$ be the optimised local structure for the DAG G. The posterior probability of $G_{\mathcal{L}^*} = (G, \mathcal{L}^*)$ is then given by

$$p(G_{\mathcal{L}^*} \mid D) \propto p(G)p(\mathcal{L}^* \mid G)p(D \mid G_{\mathcal{L}^*}), \tag{6}$$

where

$$p(D \mid G_{\mathcal{L}^*}) = \int_\theta p(D \mid G_{\mathcal{L}^*}, \theta)p(\theta \mid G_{\mathcal{L}^*})d\theta$$

is the marginal likelihood of the data, and $p(G)$, $p(\mathcal{L}^* \mid G)$, and $p(\theta \mid G_{\mathcal{L}^*})$ are priors over the global structure, local structure and parameters of the model, respectively. The right-hand side of (6) is commonly referred to as the *Bayesian score* of $G_{\mathcal{L}^*}$. We assume that the priors satisfy standard modularity constraints, so that the score factorises over the nodes in the model and can be computed efficiently for each node separately.

Specifically, we place a prior on the model structure that encourages sparsity in the global structure by penalising the number of parents of each node:

$$p(G) \propto \prod_{i=1}^{n} (1/\kappa)^{|G_i|}$$

where $\kappa \geq 1$ controls the size of the penalty ($\kappa = 1$ corresponds to a uniform prior). We assume that the CSI-trees generating each partition $\mathcal{L}_{G_i}$ are regular, so that they do not introduce conditional independencies that are in conflict with G. Apart from this, we impose no prior constraints on the trees.

For the model parameters, we use the prior for partitioned CPTs proposed by Chickering et al. (1997), in which the parameter vectors $\{\theta_{X_i|\ell_{G_i}}\}$ for every local CPD are mutually independent and Dirichlet distributed. The marginal likelihood can then be expressed in closed form and factorises over nodes and partition parts:

$$p(D \mid G_{\mathcal{L}^*}) = \prod_{i \in V} \prod_{\ell_{G_i} \in \mathcal{L}^*_{G_i}} \frac{\Gamma(\alpha_{\ell_{G_i}})}{\Gamma(\alpha_{\ell_{G_i}} + N_{\ell_{G_i}})} \prod_{x_i \in \mathcal{X}_i} \frac{\Gamma(\alpha_{x_i|\ell_{G_i}} + N_{x_i|\ell_{G_i}})}{\Gamma(\alpha_{x_i|\ell_{G_i}})}.$$

Here $\alpha_{x_i|\ell_{G_i}}$ is a Dirichlet hyperparameter, $N_{x_i|\ell_{G_i}}$ the corresponding sample count in D, $\alpha_{\ell_{G_i}} = \sum_{x_i} \alpha_{x_i|\ell_{G_i}}$ and $N_{\ell_{G_i}} = \sum_{x_i} N_{x_i|\ell_{G_i}}$. The hyperparameters are set according to a BDeu-type prior with *equivalent sample size* α_0(Heckerman, 1995):

$$\alpha_{x_i|\ell_{G_i}} = \alpha_0 \sum_{x_{G_i} \in \ell_{G_i}} P_0\left(x_i, x_{G_i}\right) = |\ell_{G_i}| \frac{\alpha_0}{|\mathcal{X}_i|\,|\mathcal{X}_{G_i}|}, \tag{7}$$

where $|\ell_{G_i}|$ is the number of parent configurations in partition part ℓ_{G_i}.

To sample DAGs with local structure from the posterior $p(G_{\mathcal{L}^*} \mid D)$, we combine a stochastic MCMC scheme over the space of DAGs with deterministic greedy optimisation of the local structure in a given DAG. This design is adapted from Pensar et al. (2015) and motivated by the fact that a unrealistic global model structure have more severe impact than a suboptimal local CSI-structure: while optimising local independence structure facilitates the discovery of weaker dependencies (edges) that would otherwise be ignored, it cannot compensate for unrealistic global independencies in the underlying DAG. Getting stuck in regions with inferior DAGs is therefore more problematic than failing to find the optimal local structure. Consequently, we let the stochastic part of the algorithm make global changes (jump between neighbouring DAGs), while the local structure is optimised in a simpler, deterministic manner. Because the score decomposes over the nodes, the local structure can be optimised for each node separately.

Several methods for sampling DAGs have been proposed and can readily be employed. In this work, we use a variant of order MCMC (Friedman and Koller, 2003), namely the hybrid scheme presented in Kuipers et al. (2022) and implemented in the `R`-package `BiDAG` (Suter et al., 2023), in which the search space is initialised by a skeleton learned by the PC algorithm (Spirtes et al., 1993).

Regarding the greedy optimisation of CSI-trees, we follow a simple recursive procedure. Starting from a single leaf representing the marginal distribution of the response variable, we find the split variable (among the conditioning variables) that yields the largest improvement in the Bayesian score, and add a new leaf for each possible value of the split variable. In the greediest version, growth stops when no remaining split improves the score sufficiently, or when the branch has reached full depth, meaning that all conditioning variables have been used. While this greedy-stopping strategy is computationally efficient, it may miss higher-scoring splits deeper in the tree. A natural alternative is therefore to grow the tree to full depth first and then prune away subtrees that do not improve the score.

3.2. Posterior over causal effects

Analogously to the standard Bayesian network model presented in Section 2.1, the pairwise causal effect τ_{ij} for any cause-effect pair (i, j) can, conditional on a network with structure $G_{\mathcal{L}^*}$, be expressed by the backdoor formula (2) and the JSD (3). Both the global and local dependence structure can induce independence patterns in the distributions that appear in the formula, parametrised by θ_{ij}.

Even if it ignores these independencies, the local estimator used in the original BIDA, presented in Section 2.2, remains consistent. A straightforward approach to compute the conditional posterior, $p(\tau_{ij} \mid D, G_{\mathcal{L}^*})$, is therefore to proceed exactly as in the case of

standard Bayesian networks without local structure: characterise the adjustment set G_{ij} in G and place Dirichlet priors on every parameter vector of the full CPTs included in θ_{ij}.

Alternatively, one can try to exploit local CSI-structure by inferring structured CPTs for the distributions that appear in the backdoor formula directly from the available data. We propose two local estimators based on this idea: one in which we infer a CSI structure for the conditional distribution of the effect variable, X_j, in (2), and a second one in which we express the backdoor formula in terms of the CPD of the cause, X_i, given the adjustment set $X_{G_{ij}}$, commonly known as the propensity score, and infer a tree-structure for this distribution. In both cases we will use the greedy procedure presented in the previous section to learn CSI-trees, but without the regularity restriction enforcing conditional dependence.

When presenting the two estimators in more detail below, we focus on the case where the adjustment set does not rule out causal relationship between X_i and X_j, that is where $X_j \notin X_{G_{ij}}$. Otherwise, the backdoor formula reduces to $p(x_j \mid do(x_i)) = p(x_j)$, implying a zero causal effect ($\tau_{ij} = 0$).

Estimator 1: Effect tree The first estimator infers a CSI-tree for the CPD of the effect variable, X_j, given the cause, X_i, and the adjustment set, $X_{G_{ij}}$. Let $\mathcal{L}^*_{i,G_{ij}}$ denote the partition of $\mathcal{X}_i \times \mathcal{X}_{G_{ij}}$ produced by the inferred tree, and assume that the conditional distribution is constant within each part, that is $\theta_{X_j|x_i,x_{G_{ij}}} = \theta_{X_j|\ell_{i,G_{ij}}}$ for all $(x_i, x_{G_{ij}}) \in \ell_{i,G_{ij}}$. Conditional on G_{ij} and $\mathcal{L}^*_{i,G_{ij}}$, we place a prior over the set of parameter vectors, $\{\theta_{X_j|\ell_{i,G_{ij}}}\}$, which are assumed to be mutually independent, such that:

$$p(\theta_{ij} \mid G_{ij}, \mathcal{L}^*_{i,G_{ij}}) = p(\theta_{X_{G_{ij}}}) \prod_{\ell_{i,G_{ij}}} p(\theta_{X_j|\ell_{i,G_{ij}}}),$$

where each marginal distribution is a Dirichlet with hyperparameters specified as in (7).

Without any regularity restrictions on the CSI-tree, we can construct trees implying that the cause, X_i, and the effect, X_j, are conditionally independent given $X_{G_{ij}}$, and thereby set $\tau_{ij} = 0$. As such, the search for local structure in the CPD can be seen as a local analogue of a global search for d-separation in the DAG, which is particularly relevant when using parent sets for adjustment (defined locally in the DAG). However, if the adjustment set already rules out conditional independence between X_i and X_j, as is the case, for example, in the class of minimal parent sets, the inferred local structure might contradict information about the global structure encoded in the adjustment set.

Estimator 2: Propensity tree The second estimator infers a tree-structure for the distribution of X_i given $X_{G_{ij}}$, commonly known as the propensity score. Using Bayes' rule, the backdoor formula (2) can be expressed in terms of the propensity score:

$$\sum_{x_{G_{ij}}} \theta_{x_j|x_i,x_{G_{ij}}} \theta_{x_{G_{ij}}} = \sum_{x_{G_{ij}}} \frac{\theta_{x_i,x_j,x_{G_{ij}}}}{\theta_{x_i|x_{G_{ij}}}}.$$

Let $\mathcal{L}^*_{G_{ij}}$ be the tree-induced partition over $\mathcal{X}_{G_{ij}}$. Since $\theta_{x_i|x_{G_{ij}}}$ is assumed constant within each part of $\mathcal{L}^*_{G_{ij}}$, the formula reduces to an average over the partition parts:

$$\sum_{x_{G_{ij}}} \frac{\theta_{x_i,x_j,x_{G_{ij}}}}{\theta_{x_i|x_{G_{ij}}}} = \sum_{\ell_{G_{ij}}} \frac{\sum_{x_{G_{ij}} \in \ell_{G_{ij}}} \theta_{x_i,x_j,x_{G_{ij}}}}{\theta_{x_i|\ell_{G_{ij}}}} = \sum_{\ell_{G_{ij}}} \theta_{x_j|x_i,\ell_{G_{ij}}} \theta_{\ell_{G_{ij}}}.$$

The rightmost expression is on the same form as the original backdoor formula. Thus, if $L_{G_{ij}}$ denotes the random variable that maps $X_{G_{ij}}$ to a part in the partition, then $L_{G_{ij}}$ is a valid adjustment variable. More intuitively, the effect on X_j from interventions on X_i is constant in each partition part, and it thus suffices to average the context-specific effects to adjust for potential confounding.

Again, we assume that the relevant parameter vectors are independently Dirichlet distributed,

$$p(\theta_{ij} \mid G_{ij}, \mathcal{L}^*_{G_{ij}}) = p(\theta_{L_{G_{ij}}}) \prod_{x_i} \prod_{\ell_{G_{ij}}} p(\theta_{X_j|x_i,\ell_{G_{ij}}}),$$

with hyperparameters set by the BDeu rule (7).

Computationally, this second approach is attractive when parent sets are used for adjustment ($G_{ij} = G_i$), because the partition $\mathcal{L}^*_{G_{ij}}$ is then specific to X_i only, and can be reused for every effect $j \in V$. Statistically, however, it cannot detect that X_j is conditionally independent of X_i, or that there are redundant variables in $X_{G_{ij}}$, which can reduce accuracy.

4. Numerical experiments

We evaluated the proposed BIDA method in terms of discovering strong causal effects using data simulated from known Bayesian networks. First, we generated data from synthetic Bayesian networks with local CSI structure, to assess performance when local structure is indeed present. Second, motivated by a more realistic scenario in which one does not know in advance whether local structure is present, we used a collection of benchmark networks from the bnlearn repository (Scutari, 2010). Below, we first describe the general simulation setup and then present the results of each experiment in turn.

4.1. Simulation setup

In both experiments, we produced datasets of different sizes by forward sampling from the corresponding networks and applied the proposed BIDA method to produce point estimates (posterior means) of the pairwise causal effects, τ_{ij}. We considered multiple variants of the method, combining either the *greedy*-stopping or the *pruned*-tree optimisation procedures introduced in Section 3.1 with one of the three parameter estimation methods described in Section 3.2. As a baseline, we also applied the original BIDA without local structure, denoted *full* in the figures below (referring to *full*-grown trees or *full* CPTs). For backdoor adjustment, we focus on results obtained with minimal parent sets, but also compare these to standard parent sets in the Appendix. As evaluation criterion we used the average precision (see e.g. Zhang and Zhang (2009)) in predicting strong pairwise effects (i.e. those exceeding a chosen threshold), which can be interpreted as an approximation of the area under the precision–recall curve. Further implementation details are given in Appendix B.

4.2. Synthetic networks with context-specific independencies

For each simulation run, we generated a Bayesian network with local structure by first sampling a random DAG and then a set of tree-structured CPTs for each node. The DAGs had $n = 20$ nodes and an expected degree of $d = 6$. Each variable was assigned a random

cardinality between 2 and 4. We generated CSI-trees of three different expected depths: minimal regular trees (*Min*); trees of random depth (*Med*); and fully grown trees that yield CPTs without explicit CSIs (*Max*). The exact procedure is described in Appendix B.

Figure 1 shows that, when the ground truth network contains local CSI structures, accounting for such structure improves the precision of the BIDA estimates, as long as the sample size is sufficiently large. When the ground truth distributions are generated by full CPTs, without explicit CSI structure, or when data are limited, the proposed methods perform similarly to the original full-CPT approach.

The observed precision gain stems from a better assessment of the underlying causal DAG structure, which can be seen from the isolated posterior edge probabilities (see Table C.1, Appendix C). The sampled DAGs with local structure are typically denser (more edges) than those without local structure, but still sparser than the ground truth. The pruned-tree approach yields somewhat denser DAGs than greedy-stopping, with higher rates of both true and false positives. Their average precision is, however, very similar. The greediest variant naturally runs faster (see Figure C.3, Appendix C).

To study how the relative performance is affected by the network density and cardinality, we ran an additional experiment in which we employed the MCMC scheme on data from networks in which these two properties varied systematically. In general, except for the setting with the smallest sample, modelling local structure improved inference about the causal DAG structure more when the ground truth contained stronger CSI structures (produced by shallow trees), when the cardinality of the variables was high, or when the underlying DAG was dense (see Figure C.1 in Appendix C). Again, we see that modelling CSI structure only improves precision when the ground truth contains explicit CSIs.

The parameter inference step of BIDA does not appear to benefit from the greedy optimisation of the local structure. When given the same MCMC sample of DAGs, neither of the two tree-based parameter estimation approaches described in Section 3.2 yields more accurate point estimates than the use of full CPTs (see Figure C.2 in Appendix C). In predicting the *strongest* effects (top quintile), both have accuracy similar to the full-CPT baseline. In predicting *positive* effects, however, the relative accuracy depends on the adjustment set: The *effect-tree* approach, optimising a tree-structured CPT for the CPD of the effect X_j, is less accurate than the full-CPT baseline when using minimal parent sets, but performs close to the baseline when using parent sets. As noted in in Section 3.2, the reason is likely that the minimal parent set already encodes information about relevant global conditional (in)dependencies, which is read off the DAG structure, including whether or not the causal effect τ_{ij} is non-zero and, if so, that all variables in the minimal parent set are required in the conditioning set for conditional independence between X_i and X_j to hold. The use of parent sets, on the other hand, generally does not rule out any conditional or marginal independencies between X_i and X_j and may also include redundant variables that are independent of X_j given X_i. The potential to inform the parameter inference step by optimising a CSI structure for the relevant CPD is therefore limited when minimal parent sets are used.

Finally, while the accuracy is not improved, substantially lower runtimes in computing point-estimates counts in favour of the tree-based parameter estimators. The fastest option depends on the tree-optimization method (Figure C.3, Appendix C).

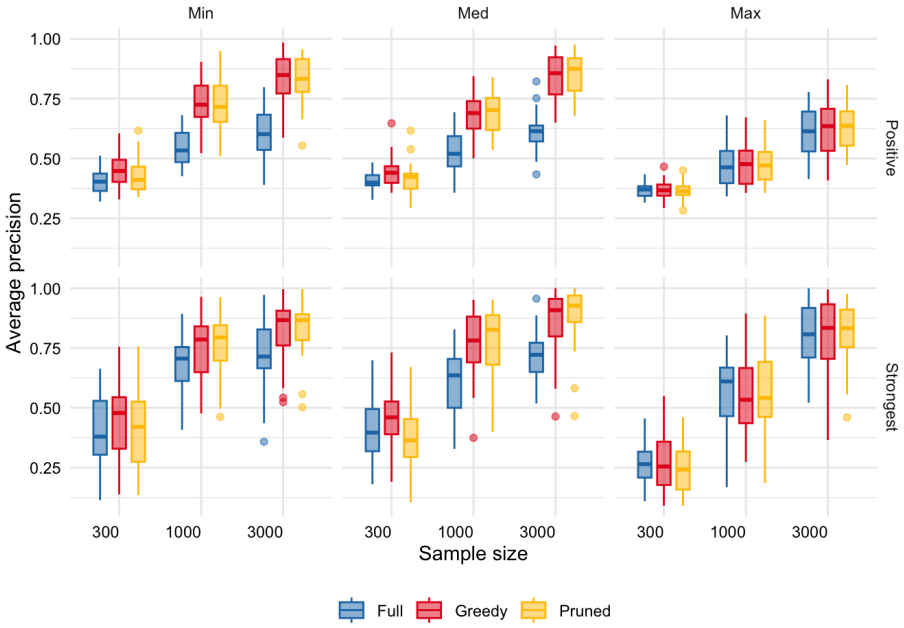


Figure 1: Average precision of the BIDA posterior means in predicting positive or strong effects (rows), across different depths of the ground truth CSI-trees (columns). Strong effects are defined as positive effects (top row) or the top quintile of positive effects (bottom row). Using the propensity-tree estimator (see Section 3.2) paired with different tree-optimisation procedures: *Greedy* refers to trees grown until no improvement can be made, *Pruned* to trees grown to full depth, and then pruned, and *Full* to full CPTs (i.e. the original BIDA without local structures). The boxplots show the distribution over 30 simulation runs.

4.3. Benchmark networks

In the second experiment, we considered only the greedy-stopping procedure for optimising CSI-trees, which performed similarly to pruned trees in the first experiment, but runs faster.

The results are more mixed, as could be expected, because the benchmark networks are not designed to explicitly contain CSI structures. Again, we see that more data are generally required when modelling local structure. For the two largest sample sizes, however, at least one of the proposed BIDA methods is most precise in discovering strong effects in half of the networks (see Table 1). The differences between the methods are relatively small across all networks, with one exception: the *Water2* network, in which the proposed method performs significantly better than the original. This network has by far the most parameters, and the full-CPT method is likely to struggle more with data fragmentation, which explains the performance gain from modelling local structures.

In contrast to what was observed for the synthetic networks, the overall most accurate parameter estimation strategy is now to infer a CSI-tree for the CPD of the effect variable (see Figure D.1, Appendix D). One possible explanation is that the sampled DAGs with local structure are now typically too dense (Table D.2, Appendix D), and that the locally optimised CSI structures infer conditional independencies that compensate for this.

Table 1: Average precision in predicting the strongest pairwise effects (top quintile) in a selection of benchmark networks.[a] Averaged over 30 simulation runs.[b]

N	Structure learning	Parameter learning	Alarm	Asia	Child	Insurance	Sachs	Water2
300	Full	Full	**0.74**	**0.95**	**0.92**	**0.8**	0.08	0.16
300	Greedy	Full	0.62	0.89	0.72	0.64	0.12	**0.34**
300	Greedy	Greedy/Effect	0.61	0.88	0.75	0.73	**0.13**	**0.34**
300	Greedy	Greedy/Propensity	0.61	0.89	0.74	0.66	0.12	**0.34**
1000	Full	Full	**0.89**	**1**	**0.99**	0.83	0.17	0.15
1000	Greedy	Full	0.81	0.98	0.98	0.78	**0.19**	**0.47**
1000	Greedy	Greedy/Effect	0.81	0.98	**0.99**	**0.86**	0.17	**0.47**
1000	Greedy	Greedy/Propensity	0.81	0.98	0.98	0.8	**0.19**	**0.47**
3000	Full	Full	**0.91**	**0.99**	**0.99**	0.86	0.23	0.55
3000	Greedy	Full	0.89	0.98	**0.99**	0.85	**0.27**	**0.65**
3000	Greedy	Greedy/Effect	0.89	0.98	**0.99**	**0.93**	**0.27**	**0.65**
3000	Greedy	Greedy/Propensity	0.89	0.97	**0.99**	0.86	**0.27**	**0.65**

[a] From the bnlearn repository (Scutari, 2010). Key characteristics are shown in Table D.1, Appendix D.
[b] Bold indicates the larges value by sample size (N) and network.

5. Conclusion

We have presented a Bayesian IDA (BIDA) method for estimating causal effects in Bayesian networks with local structure, based on a combination of MCMC-based sampling the of the underlying DAG and a deterministic greedy optimisation routine for inferring local dependence structures in the form of CSI-trees.

In simulation studies, the proposed method improved the discovery of strong causal effects when the ground truth networks contained explicit CSI structures, compared to the original BIDA method. The relative improvement was traced to an better assessment of the underlying causal DAG structure. The two methods performed similarly in the case where no local structure was imposed on the ground truth networks. On a set of benchmark networks, the results were more mixed: modelling of local structure resulted in an improved accuracy in half of the networks when the sample size was sufficiently high, with the largest improvement gained in the most complex network, Water2, which contains around 10,000 parameters. The original BIDA performed overall at a comparable level across all networks, and was overall more accurate for the smallest sample sizes. Relative to the ground truth networks, the sampled DAGs were typically too dense, suggesting overfitting, and that further encouragement of sparsity could be beneficial.

While the role of the structure prior has been highlighted in previous research (Pensar et al., 2015, 2016), a general solution to this issue is still lacking and remains an interesting direction for future research. In a purely probabilistic setting, learning a global structure that is too dense is arguably preferable over learning one that is too sparse, as one cannot compensate for unrealistic global independence assumptions through the CPDs (Friedman and Goldszmidt, 1998; Pensar et al., 2015). In the causal setting, however, where the edges point in the direction of causal influence, the trade-off might be different and, possibly, more involved.

Another natural direction for future research is to investigate how local structure optimization affects the MCMC convergence properties, as restricting updates to greedily optimized local structures may alter the mixing efficiency of the algorithm.

Declaration of the use of generative AI

The authors employed generative AI tools in the preparation of the manuscript, as well as in the production and documentation of the R-code implementing the numerical experiments. Specifically, tools supported by GitHub Copilot were used, including models provided by OpenAI, Anthropic, and Google. The choice of model was automated. In addition, for spelling and grammatical checks, Writefull's model was used via Overleaf.

After using these tools, the authors reviewed and edited all AI-assisted content. They take full responsibility for the content of the published article.

Acknowledgments

This research was supported by the Research Council of Norway under the Centres of Excellence scheme (grant 332645: Integreat - The Norwegian centre for knowledge-driven machine learning).

References

Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-Specific Independence in Bayesian Networks. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 115–123, Portland, OR, 1996. Morgan Kaufmann Publishers Inc.

David Maxwell Chickering, David Heckerman, and Christopher Meek. A Bayesian Approach to Learning Bayesian Networks with Local Structure. In *13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 80–89, February 1997.

Nir Friedman and Moises Goldszmidt. Learning Bayesian Networks with Local Structure. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 421–459. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-010-6104-9 978-94-011-5014-9. doi: 10.1007/978-94-011-5014-9_15.

Nir Friedman and Daphne Koller. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning 2003 50:1*, 50 (1):95–125, January 2003. ISSN 1573-0565. doi: 10.1023/A:1020249912095.

Paul E. Griffiths, Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight. Measuring Causal Specificity. *Philosophy of Science*, 82(4):529–555, October 2015. ISSN 0031-8248, 1539-767X. doi: 10.1086/682914.

David Heckerman. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20:197–243, 1995.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Number 3. MIT press, 2009. ISBN 978-0-262-01319-2.

Jack Kuipers, Polina Suter, and Giusi Moffa. Efficient Sampling and Structure Learning of Bayesian Networks. *Journal of Computational and Graphical Statistics*, 31(3):639–650, July 2022. ISSN 1061-8600. doi: 10.1080/10618600.2021.2020127.

Vera Kvisgaard and Johan Pensar. Bayesian estimation of causal effects from observational categorical data. *Bayesian Analysis*, pages 1–28, 2025. doi: 10.1214/25-BA1540.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. ISSN 1557-9654. doi: 10.1109/18.61115.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009. ISSN 0090-5364. doi: 10.1214/09-AOS685.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

Johan Pensar, Henrik Nyman, Timo Koski, and Jukka Corander. Labeled directed acyclic graphs: A generalization of context-specific independence in directed graphical models. *Data Mining and Knowledge Discovery*, 29(2):503–533, March 2015. ISSN 1573756X. doi: 10.1007/S10618-014-0355-0/FIGURES/14.

Johan Pensar, Henrik Nyman, Jarno Lintusaari, and Jukka Corander. The role of local partial independence in learning of Bayesian networks. *International Journal of Approximate Reasoning*, 69:91–105, February 2016. ISSN 0888-613X. doi: 10.1016/J.IJAR.2015.11.008.

Johan Pensar, Topi Talvitie, Antti Hyttinen, and Mikko Koivisto. A Bayesian Approach for Estimating Causal Effects from Observational Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5395–5402, 2020. ISSN 2374-3468. doi: 10.1609/AAAI.V34I04.5988.

Marco Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35:1–22, July 2010. ISSN 1548-7660. doi: 10.18637/jss.v035.i03.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1 edition, 1993. ISBN 978-1-4612-7650-0.

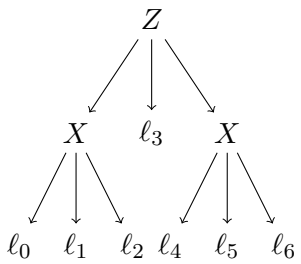
Polina Suter, Jack Kuipers, Giusi Moffa, and Niko Beerenwinkel. Bayesian Structure Learning and Sampling of Bayesian Networks with the R Package BiDAG. *Journal of Statistical Software*, 105:1–31, January 2023. ISSN 1548-7660. doi: 10.18637/jss.v105.i09.

Ethan Zhang and Yi Zhang. Average precision. In LING LIU and M. TAMER ÖZSU, editors, *Encyclopedia of Database Systems*, pages 192–193. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_482.

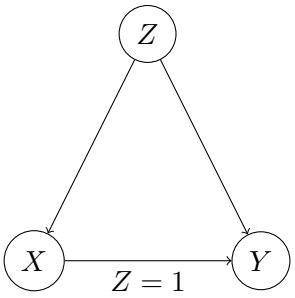
Appendix A. Illustration of context-specific independence and its representations

Z	X	$\theta_{Y\|z,x}$
0	0	$\theta_{Y\|0,0}$
0	1	$\theta_{Y\|0,1}$
0	2	$\theta_{Y\|0,2}$
1	0	$\theta_{Y\|1,\cdot}$
1	1	$\theta_{Y\|1,\cdot}$
1	2	$\theta_{Y\|1,\cdot}$
2	0	$\theta_{Y\|2,0}$
2	1	$\theta_{Y\|2,1}$
2	2	$\theta_{Y\|2,2}$

(a)



(b)



(c)

Figure A.1: Different representations of a CPD with the context-specific independence (CSI) $Y \perp X, z = 1$: (a) the conditional probability table, (b) the CSI-tree partitioning the outcome space of (X, Z), and (c) the *labeled DAG* representation, in which the label $Z = 1$ indicates that the edge is not active in this context (Pensar et al., 2015).

Figure A.1 illustrates the concept of context-specific independence (CSI) and its representations. The CPT in (a) includes the CSI from Definition 1, $(Y \perp X, z = 1)$, which is represented by the CSI-tree in (b). The CSI-tree partitions the outcome space of (X, Z) into 7 parts. This particular CSI can be represented by a *labeled DAG* (Pensar et al., 2015), as shown in (c), in which the edge from X to Y is labeled with the context $Z = 1$, indicating that the edge is not active in this context. While the labeled DAG representation is outside the scope of this article, it is useful for reminding the reader of the causal interpretation of CSIs. A label on an edge in a LDAG indicates contexts in which the edge is not active, with the causal interpretation that a parent is not a cause of its child in those contexts. In our example, X is not a cause of Y when $Z = 1$, as indicated by the label on the edge from X to Y in the LDAG in Figure A.1 (c).

Appendix B. Implementation details

B.1. Implementation of the BIDA method

We implemented our proposed BIDA method in `R`.

First, we implemented the procedure to infer CSI-trees by greedy optimisation of the Bayesian score outlined in Section 3.1. We set the threshold for a sufficient improvement to $\log(\eta)$, with $\eta = 2$.

Next, we combined this deterministic routine for optimising CSI-trees with the hybrid order MCMC scheme implemented in the `BiDAG` package Suter et al. (2023). We set the edge penalty κ equal to 2 in the graph prior. This value is the default value used in `BiDAG` for scoring categorical data, and aligns with the threshold η of including an additional parent as a split variable in the greedy optimisation procedure. In addition, we restricted the maximum size of any parent set to 6. We used an equivalent sample size $\alpha_0 = 1$ in all parameter priors. We ran the default number of MCMC iterations, returning a sample of 1000 DAGs, of which we discarded the first 200 as burn-in.

B.2. Generation of Bayesian networks with local structure

To generate Bayesian networks with local structure, we first drew a DAG with $n = 20$ nodes and expected degree $d = 6$ (number of neighbours). We then randomly selected a variable cardinality k_i between 2 and 4 for each variable.[3]

Next, we generate the CPDs for each node X_i. For any node with at least two parents, local structure in the distribution was generated by recursively growing a CSI-tree, partitioning the parent space $\mathcal{X}_{G_i}$ as follows. Starting from a single leaf, with probability ρ, we performed a split by randomly selecting a parent variable and adding a new leaf for each possible value of that variable. Each parent is involved in at most one split in each branch. To enforce regularity (avoid producing independence between a child and any of its parents), we added a split to the shallowest leaf for every parent absent from the initial tree. Finally, we drew a CPD for each leaf by sampling a set of Dirichlet vectors with mean equal to a random perturbation of $\mu_i \propto (1, 1/2, \ldots, 1/k_i)$ and concentration parameter $\alpha = 10$.

The split probability ρ controls the depth of the CSI-tree. Shallow trees describes relatively simple CSI structures and were expected to be easier to infer. Increasing tree depth allowed more complex CSI patterns. We considered $\rho \in \{0, 1/2, 1\}$; $\rho = 0$ corresponds to the tree where each parent forms exactly one split (by the above post-hoc routine for avoiding independencies) and $\rho = 1$ corresponds to a complete tree (a full CPT). In the plots we refer to the tree depths by *Min, Med* and *Max*, respectively.

3. In a set of preliminary experiments, we systematically varied the density of the DAGs and the cardinality of the variable. We only evaluated the posterior edge probabilities on data sampled from these networks. See Figure 1, Appendix C.

Appendix C. Additional simulation results for the synthetic networks

Table C.1: Evaluation of the posterior sample of DAGs. The grouped columns show figures from running order MCMC with full CPTs (no local structure) and with local structure optimised by the greediest (Greedy) and the pruning (Pruned) procedure. The figures include: the total number of edges (TOT), true positive rate (TPR), false positive rate (FPR), the average precision of the posterior edge probabilities (Avgppv). TOT, TPR and FPR is averaged over the DAGs in the sample. All figures are averaged over 30 simulation runs. The average number of edges in the true DAGs is 60.7.

		Full				Pruned				Greedy			
Depth	N	TOT	TPR	FPR	Avgppv	TOT	TPR	FPR	Avgppv	TOT	TPR	FPR	Avgppv
Min	300	14.22	0.17	0.01	0.40	35.40	0.26	0.06	0.29	27.04	0.26	0.04	0.38
Min	1000	25.58	0.35	0.01	0.48	45.78	0.60	0.03	0.64	41.47	0.58	0.02	0.66
Min	3000	34.42	0.47	0.02	0.51	54.26	0.78	0.02	0.79	52.21	0.77	0.02	0.78
Med	300	13.93	0.17	0.01	0.40	33.88	0.23	0.06	0.29	25.39	0.24	0.03	0.37
Med	1000	24.73	0.33	0.02	0.45	44.37	0.55	0.03	0.60	38.76	0.54	0.02	0.62
Med	3000	32.44	0.44	0.02	0.49	51.37	0.75	0.02	0.78	49.57	0.74	0.02	0.78
Max	300	9.27	0.10	0.01	0.32	22.36	0.13	0.05	0.22	15.58	0.12	0.03	0.27
Max	1000	17.47	0.25	0.01	0.45	31.40	0.33	0.04	0.40	25.55	0.29	0.02	0.43
Max	3000	26.88	0.42	0.01	0.57	40.81	0.53	0.03	0.59	35.97	0.51	0.02	0.60

Figure C.1: Average precision of the posterior edge probabilities from running order MCMC with full CPTs (no local structure) and with local CSI-structure optimised by the greedy-stopping procedure (Greedy) or pruned-tree procedure (Pruned). Across different combinations of the density d of the DAG (expected number of neighbours), variable cardinality k, and depth of CSI-trees (Min, Med, Max). The box-plot show the distribution over 30 simulation runs.

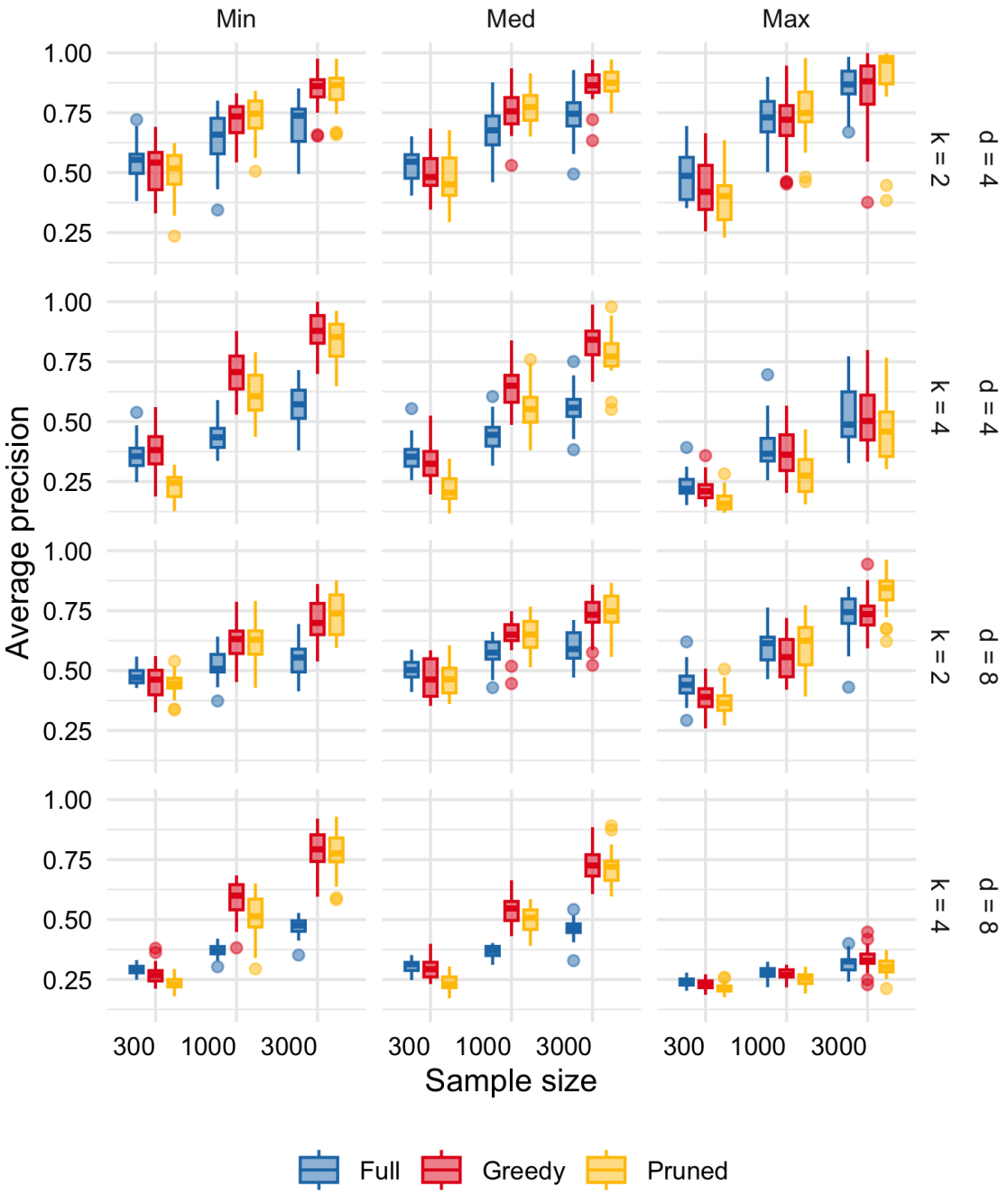
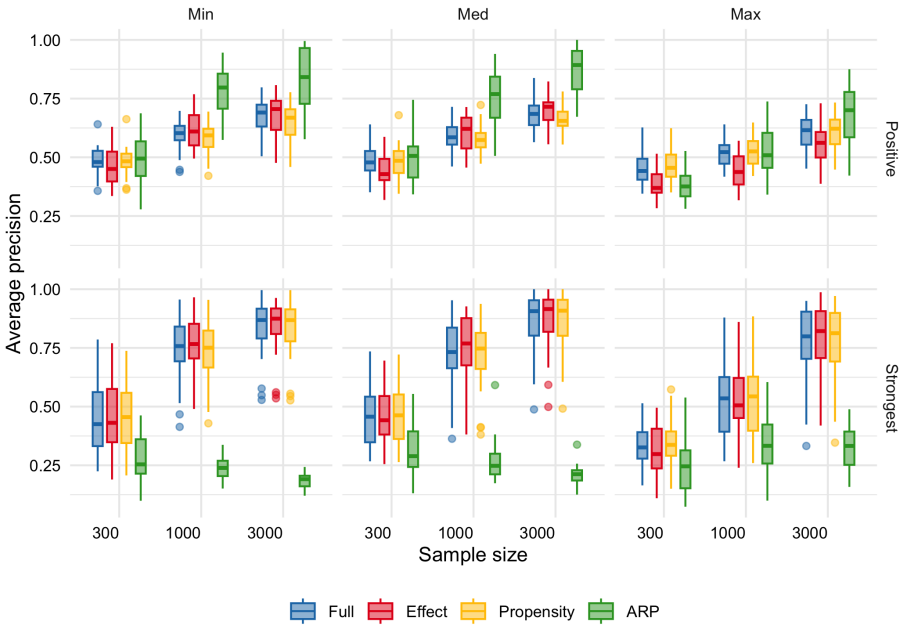


Figure C.2: Average precision of the BIDA posterior means, different conditional parameter estimation strategies: *Full* refers to full CPTs, while *Effect* and *Propensity* refers to inferring a CSI-structured CPD for the effect variable and the propensity score, respectively (see Section 3.2 in the manuscript for details). The CSI-structure is optimised growing greedy trees until no candidate split improves the score, also when sampling DAGs with local structure. The ancestor relation probabilities (ARP) obtained are added as a baseline. Precision in predicting the positive effects and the strongest effects (rows), across different benchmark networks (columns). The boxplots show the distrbution over 30 simulation runs.

(a) Using parent sets for adjustment



(b) Using minimal parent sets for adjustment

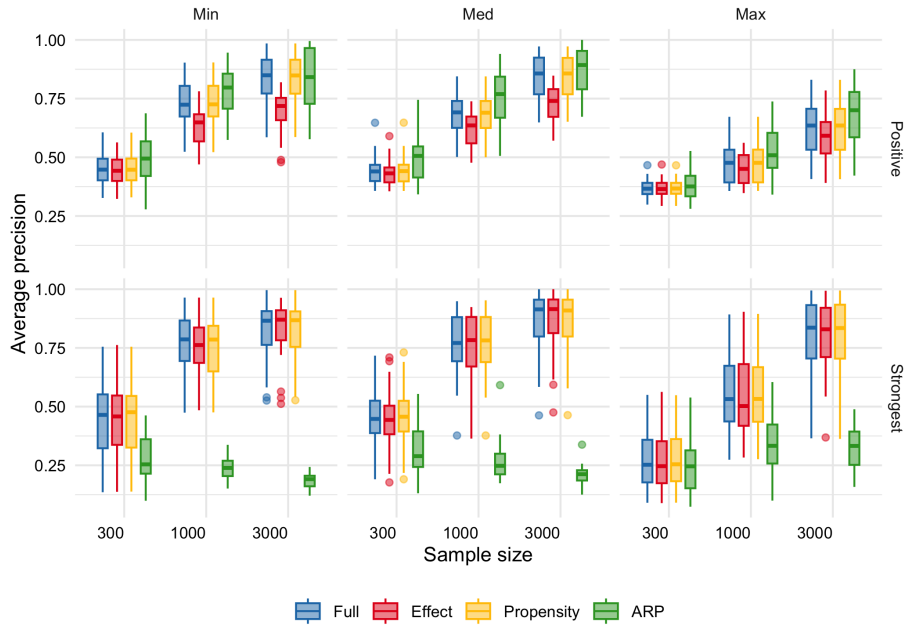
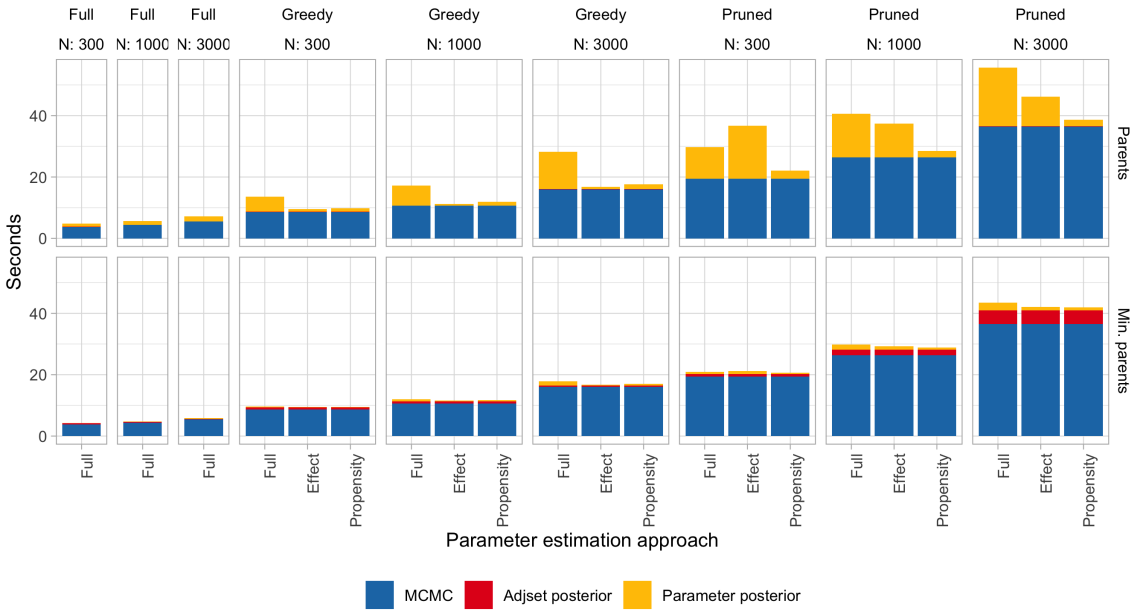


Figure C.3: Decomposed runtimes of the BIDA procedure, assuming no local structure (*Full*) or optimising local CSI-structure with the greedy-stopping procedure (*Greedy*) or pruning procedure (*Pruned*). Across different adjustment sets (rows), and parameter estimation approaches (x-axis): *Full* tables, or optimising a CSI-structured CPD for the *Effect* variable or the *Propensity* score (see Section 3.2 for details). The runtimes are averaged over 30 simulation runs.



Appendix D. Additional simulations results for the benchmark networks

Table D.1: Network characteristics. The networks are downloaded from the online bnlearn repository (https://www.bnlearn.com/bnrepository/). *Water2* is a modified version of the original *Water* network from the bnlearn repository, in which the deterministic distributions of the source nodes in the original network are replaced with uniform distributions.

Network	Nodes	Arcs	Parameters	Average Degree	Maximum In-Degree
Asia	8	8	18	2.00	2
Sachs	11	17	178	3.09	3
Child	20	25	230	2.50	2
Insurance	27	52	1008	3.85	3
Water2	32	66	10083	4.12	5
Alarm	37	46	509	2.49	4

Table D.2: Evaluation of the posterior sample of DAGs. The grouped columns show figures from running order MCMC with full CPTs (no local structure) and with local structure optimised by the greediest procedure (Greedy). The figures include: the total number of edges (TOT), true positive rate (TPR), false positive rate (FPR), the average precision of the posterior edge probabilities (Avgppv). TOT, TPR and FPR are averaged over all DAGs in the sample. All figures are averaged over 30 simulation runs.

			Full				Greedy			
Network	Edges	N	TOT	TPR	FPR	Avgppv	TOT	TPR	FPR	Avgppv
Alarm	46	300	51.38	0.77	0.01	0.82	81.96	0.73	0.04	0.63
Alarm	46	1000	48.61	0.88	0.01	0.91	80.03	0.84	0.03	0.76
Alarm	46	3000	47.16	0.90	0.00	0.94	73.00	0.89	0.02	0.85
Asia	8	300	8.59	0.51	0.09	0.71	8.20	0.52	0.08	0.71
Asia	8	1000	8.71	0.71	0.06	0.89	8.76	0.67	0.07	0.85
Asia	8	3000	8.27	0.78	0.04	0.93	8.27	0.76	0.05	0.91
Child	25	300	19.50	0.66	0.01	0.77	37.46	0.69	0.06	0.58
Child	25	1000	21.64	0.76	0.01	0.82	30.31	0.89	0.02	0.88
Child	25	3000	24.57	0.91	0.01	0.96	27.43	0.93	0.01	0.96
Insurance	52	300	35.57	0.53	0.01	0.64	58.87	0.57	0.04	0.55
Insurance	52	1000	39.17	0.63	0.01	0.71	57.74	0.69	0.03	0.70
Insurance	52	3000	43.38	0.74	0.01	0.78	58.35	0.80	0.03	0.82
Sachs	17	300	11.82	0.34	0.07	0.49	18.25	0.50	0.10	0.43
Sachs	17	1000	15.63	0.66	0.05	0.80	19.19	0.60	0.10	0.49
Sachs	17	3000	17.00	0.74	0.05	0.90	18.43	0.67	0.08	0.64
Water2	66	300	25.43	0.16	0.02	0.21	34.33	0.23	0.02	0.26
Water2	66	1000	28.05	0.20	0.02	0.24	36.81	0.32	0.02	0.37
Water2	66	3000	36.35	0.37	0.01	0.41	47.30	0.47	0.02	0.48

Figure D.1: Average precision of the BIDA posterior means, for different conditional parameter estimation strategies: *Full* refers to full CPTs, while *Effect* and *Propensity* refers to inferring a CSI-structured CPD for the effect variable and the propensity score, respectively (see Section 3.2 in the manuscript for details). The CSI-structure is optimised growing greedy trees until no candidate split improves the score, also when sampling DAGs with local structure. The ancestor relation probabilities (ARP) obtained are added as a baseline. Precision in predicting the positive effects and the strongest effects (rows), across different benchmark networks (columns). The boxplots show the distrbution over 30 simulation runs.

(a) Parent sets



(b) Minimal parent sets



Figure D.2: Decomposed runtimes of the BIDA procedure, assuming *Full* CPTs or optimising local CSI-structure with the *Greedy* procedure in which where trees are grown until no improvement in the score can be made. Across different parameter estimation approaches on the x-axis (explained in Figure D.1). Computing minimal sets are more demanding than parent sets. However, since more effects are set to zero exactly when using minimal sets, it is typically cheaper to sample from the conditional posterior. Of the same reason, the runtimes can be lower when inferring a CSI-structured CPD for the effect variable compared to the propensity score, despite the structure for the propensity score being reused for every effect variable. The sample size is $N = 1000$ and the runtimes are averaged over 30 simulation runs.

