Instance-Specific Bayesian Network Structure Learning

Fattaneh Jabbari[1] FAJ5@PITT.EDU
Shyam Visweswaran[1,2] SHV3@PITT.EDU
Gregory F. Cooper [1,2] GFC@PITT.EDU
[1]*Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA*
[2]*Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA*

Abstract

Bayesian network (BN) structure learning algorithms are almost always designed to recover the structure that models *the relationships that are shared by the instances in a population*. While accurately learning such population-wide Bayesian networks is useful, learning Bayesian networks that are specific to each instance is often important as well. For example, to understand and treat a patient (instance), it is critical to understand the specific causal mechanisms that are operating in that particular patient. We introduce an instance-specific BN structure learning method that searches the space of Bayesian networks to build a model that is specific to an instance by guiding the search based on attributes of the given instance (e.g., patient symptoms, signs, lab results, and genotype). The structure discovery performance of the proposed method is compared to an existing state-of-the-art BN structure learning method, namely an implementation of the Greedy Equivalence Search algorithm called FGES, using both simulated and real data. The results show that the proposed method improves the precision of the model structure that is output, when compared to GES, especially for those variables that exhibit context-specific independence.

Keywords: causal Bayesian networks, structure learning, context-specific independence, instance-specific machine learning.

1. Introduction

A Bayesian network (BN) is a well-known graphical model that represents probabilistic relationships among a set of variables. Under assumptions, BNs can be interpreted as causal models and learned from observational data, which has wide applicability (Spirtes et al., 2000; Pearl, 2009; Illari et al., 2011). In this paper, for domain emphasis, we focus on learning causal Bayesian networks (CBNs), although the methods apply to BN structure learning in general. There are two main approaches to learning CBN structures from data: (1) constraint-based and (2) score-based (e.g., Bayesian) approaches, although other methods are also being actively developed and investigated (Peters et al., 2012; Jabbari et al., 2017). A constraint-based approach iteratively performs many statistical independence tests on data to constrain the structures that are consistent with the test results; it then outputs the CBN structure that is most consistent. A Bayesian approach typically uses heuristic search and outputs the most probable CBN structure it can find.

Almost all CBN structure learning algorithms are designed to recover the structure that models the relationships that are shared by the instances in a population. While learning accurate population-wide CBNs is useful, learning CBNs that are specific to a given instance can also be very important. For example, a breast-cancer tumor (instance) in a patient can have a set of causal mechanisms that are different from that of another breast-cancer tumor either in the same patient or in a different patient. However, to determine the most effective treatment for a tumor in the current

patient, it is important to know the particular causal mechanisms that are driving that tumor to be cancerous. In reality, a given tumor usually is a composite of cellular mechanisms that rarely all occur together, yet each individual mechanism may appear relatively commonly in other tumors. A population-wide CBN would at best capture the more common mechanisms operating in breast cancer and not all of the particular mechanisms that are active in the current patient's breast-cancer tumor. The task, then, is to construct the joint set of mechanisms of a given tumor from the individual mechanisms seen in previous tumors. To do so, we use the known features (i.e., the variable values) of the current tumor to help identify and construct the individual mechanisms that compose the set of mechanisms that are jointly driving the current tumor. In the extreme scenario, if the mechanisms in every tumor are completely different from every other, we have little hope of learning its mechanisms from a training set of prior tumors. The reality is that each of several mechanisms that is active in a tumor typically occurs in *some* other tumors, but not in *all* other tumors.

More generally, a given person can be viewed as a joint set of causal mechanisms, where each mechanism is typically shared with many other people, but the joint set is essentially unique to that person. In a given person, the causal learning task is to construct the correct set of mechanisms for that person from the features we know about the person and from a training set of data on many other people. Moreover, this instance-specific causal learning approach is applicable to other causal systems, beyond human biology.

In this paper, we propose a novel, fully Bayesian instance-specific structure learning method that searches the space of CBNs to build a model that is specific to an instance T by guiding the search based on T's attributes. This is a fundamental research problem that has received relatively little attention to date. We hypothesize that such an instance-specific learning approach will model the causal relationships for T better than does a population-wide one. We evaluate this hypothesis using simulated and real data.

2. Related Work

Our method uses CBNs that represent context-specific independence (CSI). (Boutilier et al., 1996) introduced the notion of context-specific independence to capture independence relationships that hold between the parents and a child node in a BN in certain contexts (i.e., when the parent variables take on particular values); in general these types of independencies cannot be captured completely in the structure of standard BNs, wherein the BN structure is invariant to CSI relationships.

A number of greedy search algorithms have been proposed to learn CSI in BNs. (Friedman and Goldszmidt, 1998) introduced a method that incorporates tree-structured conditional probability tables (CPTs) into a BN structure search algorithm using a minimum description length (MDL) score. (Chickering et al., 1997) proposed using decision-graph CPTs that can represent a richer set of independence relationships, compared to tree-structured CPTs. (Chickering et al., 1997) also developed a Bayesian score to evaluate the posterior probability of Bayesian networks that contain decision-graph CPTs. This score is applied along with a greedy search algorithm to learn a global BN structure in which the relationship between each node and its parents is represented using a decision graph. Recently, (Pensar et al., 2015) introduced a method to label the edges of a BN to encode local CSI structures; such graphs are called labeled directed acyclic graphs (LDAGs). Similar to (Chickering et al., 1997), (Pensar et al., 2015) also proposed a LDAG-based Bayesian score and MCMC search to learn a LDAG structure. (Zou et al., 2017) proposed an ordering-based algorithm to learn local structures using Lasso regression (Tibshirani, 1996) on a linear combination of

Boolean functions, where Boolean functions define the interactions among parents of each variable. (Oates et al., 2016) proposed a method that uses integer linear programming to learn multiple DAGs from multiple units of data, where each unit contains a set of data cases.

None of the methods in the previous paragraph learns a model that is specific to a given instance (e.g., a given patient), which is the main goal and novel contribution of the current paper. Doing so has two advantages: (1) First and foremost, the learned causal model is specific to the current instance. We dynamically search to define the clusters of cases associated with the test instance T. Importantly, this search occurs at the node level, not at the DAG level. Thus, depending on which node and parents are being scored, we allow for using different clusters of cases when learning a DAG for a test instance T. (2) Given that we seek an instance-specific model, searching for it directly is generally much more efficient than is searching for all (or at least many) possible instance-specific models and then choosing the one that matches the current test instance.

The work in (Cooper et al., 2018) learns tumor-specific causal models from data. However, that method is limited to searching over bipartite causal graphs in which one partition contains causes and the other contains effects. Also, the method assumes there is one and only one cause for each effect. Both assumptions are reasonable for that application, but restrict generality. The current paper describes a general approach for learning unrestricted, instance-specific CBNs.

3. Background

As mentioned earlier, a BN is a graphical model that is often used to represent probabilistic relationships among a set of variables. In general, a BN is composed of a graphical model structure G, which is a directed acyclic graph (DAG), and a set of parameters θ for the DAG. A DAG G consists of nodes that correspond to variables and directed edges that represent the conditional dependence relationships among those variables. A parameter set θ parametrizes the relationships that are present in the DAG. Greedy Equivalence Search (GES) (Chickering, 2003) is a state-of-the-art method for learning a BN structure from observational data. In this section, we provide an overview of GES and the Bayesian Dirichlet equivalent uniform (BDeu) score (Heckerman, 1998), which can be used together with GES to learn a BN structure from discrete data. The GES algorithm and the BDeu score form the infrastructure of our proposed instance-specific BN structure learning method.

3.1 Overview of Greedy Equivalence Search (GES)

(Chickering, 2003) developed GES that identifies a CBN by searching over the equivalence classes of BN structures, i.e., DAGs. The equivalence class of DAGs represents a set of DAGs that have the same d-separation properties and can be represented by *partially directed acyclic graphs* (PDAGs), also known as *patterns*. A PDAG is a mixed graph that contains both directed and undirected edges.

GES is a two-phase score-based algorithm that includes a forward equivalence search (FES) and backward equivalence search (BES). It works as follows. Let ε be the current PDAG during the search. During forward search, let $\boldsymbol{\varepsilon}^{+}(\varepsilon)$ represent the set of PDAGs that are generated by adding a single edge to ε for each legal edge addition (Chickering, 2003, 1995). Similarly, during the BES, $\boldsymbol{\varepsilon}^{-}(\varepsilon)$ is the set of PDAGs that are obtained by deleting each single edge from ε. The forward phase of GES starts with an empty graph (i.e., $\varepsilon = \emptyset$) and replaces the current state with the PDAG in $\boldsymbol{\varepsilon}^{+}(\varepsilon)$ that has the highest score. It continues this phase until no further local improvement can be achieved. The backward phase starts from the local maximum achieved by the forward phase and performs a backward search by replacing ε with the highest scoring PDAG in $\boldsymbol{\varepsilon}^{-}(\varepsilon)$. It stops when

it reaches a local maximum. For more information about this method see (Chickering, 2003). In this paper, we use an efficient implementation of GES called FGES (Ramsey et al., 2017).

Each forward and backward step in GES involves scoring a single node given its parents; therefore, it requires a node-wise decomposable score. The Bayesian information criterion (BIC) score (Schwarz, 1978) is often used to learn a BN structure when variables follow a Gaussian distribution and the BDeu score (Heckerman, 1998) is often used for multinomial variables, although other scores are possible. In the following section, we review the BDeu score since we concentrate on using multinomial variables in this paper.

3.2 Scoring Bayesian Networks

As mentioned earlier, development of a Bayesian approach for learning a BN structure amounts to search for a structure with a high posterior probability on a given dataset. Let D be a dataset containing n discrete variables $\mathcal{X} = \{X_1, X_2, ..., X_n\}$, where each variable X_i can take r_i values and its parents $Pa(X_i)$ can take q_i distinct instantiations. Also, let G be the structure we wish to score. According to Bayes' theorem, the posterior probability of graph G given data D is as follows:

$$P(G|D) = \frac{P(D|G) \cdot P(G)}{P(D)}\,, \tag{1}$$

where $P(D|G)$ is the marginal likelihood of the data, $P(G)$ is the structure prior, and $P(D)$ is the probability of the data. Since $P(D)$ is a normalization constant and independent of the model, we define the score of model G as follows:

$$\text{score}(G) = P(D|G) \cdot P(G)\,, \tag{2}$$

where we can compute $P(D|G)$ by integrating over all unknown parameters θ as follows:

$$P(D|G) = \int_{\theta} P(D|G,\theta) \cdot P(\theta|G)\,. \tag{3}$$

The marginal likelihood of the data has a closed-form solution called the Bayesian Dirichlet (BD) score under the following assumptions: (1) the data are discrete and complete; (2) data samples are independent and identically distributed; (3) parameter priors are represented using Dirichlet distributions that are assumed independent over the model parameters. The BD score is as follows:

$$P(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}\,, \tag{4}$$

where the first product is over all n variables, the second product is over the q_i parent instantiations of variable i, and the third product is over all r_i values of variable X_i. The term N_{ijk} is the number of cases in D in which variable $X_i = k$ and its parent $Pa(X_i) = j$; also, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. The term α_{ijk} is a Dirichlet prior parameter that may be interpreted as representing "pseudo-counts" and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. We may define the pseudo-counts to be evenly distributed, in which case Equation (4) represents the so-called *BDeu score* (Heckerman, 1998):

$$\alpha_{ijk} = \frac{\alpha}{r_i \cdot q_i}\,, \tag{5}$$

where α is a positive constant called the prior equivalent sample size (PESS) (Heckerman et al., 1995). The BDeu score described here is a modular score that is decomposable at node level, as required by the GES algorithm.

4. Instance-Specific GES (IGES)

This section describes a novel algorithm called instance-specific GES (IGES) that takes as input a set D of training instances and an instance $T = \{X_1 = x_1, X_2 = x_2, ..., X_n = x_n\}$ that is not in D, and it returns as output a causal structure G_{IS} for instance T and a (often different) causal structure G_{pop} for the instances in D. The goal of IGES is to find causal structures G_{IS} and G_{pop} that maximize $P(G_{IS}, G_{pop}|D, T)$. To do so, it derives $P(D|T, G_{IS}, G_{pop})$ and $P(G_{IS}, G_{pop})$. Since finding a global optimum for $P(G_{IS}, G_{pop}|D, T)$ is generally not computationally tractable, IGES performs GES-style greedy search.

IGES operates in two phases. The first phase uses GES (as described in Section 3.1) with the BDeu score to find G_{pop} given D. GES uses heuristic search that seeks to find the G_{pop} that optimizes $P(G_{pop}|D)$. The second phase uses GES with a novel, instance-specific Bayesian score called IS-Score (see below) to find G_{IS} given D, T, and G_{pop}; we use the name GES2 to denote this application of GES. GES2 uses heuristic search that seeks to find the instance-specific structure G_{IS} that optimizes $P(G_{IS}|D, T, G_{pop})$. Algorithm 1 shows the high-level procedure [1].

Algorithm 1 IGES(D, T)

Input: dataset D, instance T
Output: an instance-specific model G_{IS} and a population-wide model G_{pop}

1: G_{pop} = GES(D)
2: G_{IS} = GES2(D, T, G_{pop})
3: return G_{IS} and G_{pop}

GES2 is a modification of GES that uses the procedure IS-Score (defined below) to score a node X given its parents $Pa_{IS}(X)$ in G_{IS} and its parents $Pa_{pop}(X)$ in G_{pop}. Let $Pa_{IS}(X) = j$ denote that the variables in vector $Pa_{IS}(X)$ have the values denoted by vector j in instance T. The basic idea behind the IS-Score is to find those cases (samples) in D in which $Pa_{IS}(X) = j$ and use them to score $Pa_{IS}(X) \rightarrow X$ in G_{IS}. In essence, those instances in D form a cluster that are similar to instance T in the context of scoring $Pa_{IS}(X) \rightarrow X$. Since those instances are being used to score G_{IS}, in order to avoid duplicate scoring, they can no longer be used to also score G_{pop}; thus, the score for G_{pop} must be adjusted accordingly. More specifically, let $D_{Pa_{IS}(X)=j}$ denote the instances in D in which $Pa_{IS}(X) = j$; let $D_{Pa_{IS}(X)\neq j}$ denote the remaining instances in D. Using data $D_{Pa_{IS}(X)=j}$, the score for $Pa_{IS}(X) \rightarrow X$ in instance-specific model G_{IS} is as follows:

$$\text{score}_{IS}\left(D_{Pa_{IS}(X)=j}, Pa_{IS}(X) \rightarrow X\right) =$$
$$P(D_{Pa_{IS}(X)=j}|Pa_{IS}(X) \rightarrow X) = \frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j + N_j)} \cdot \prod_{k=1}^{r} \frac{\Gamma(\alpha_{jk} + N_{jk})}{\Gamma(\alpha_{jk})}, \quad (6)$$

where r denotes all the possible instantiations of X, N_{jk} is the number of instances in $D_{Pa_{IS}(X)=j}$ in which X has the value k, and $N_j = \sum_{k=1}^{r} N_{jk}$; the terms α_{jk} and $\alpha_j = \sum_{k=1}^{r} \alpha_{jk}$ are the corresponding Dirichlet priors.

1. To be more comprehensive, the algorithm would return a causal structure for each instance in D; however, doing so would require more computation time, and it is not the main goal of the current paper. We plan to pursue this extension in future work.

Let $Pa_{pop}(X)$ denote the parents of X in the population-wide model G_{pop}, which in general may be different than the parents of X in G_{IS}, as given by $Pa_{IS}(X)$. Using data $D_{Pa_{IS}(X)\neq j}$, the score for $Pa_{pop}(X) \rightarrow X$ in population-wide model G_{pop} is as follows:

$$\begin{aligned} &\text{score}_{pop}\left(D_{Pa_{IS}(X)\neq j}, Pa_{pop}(X) \rightarrow X\right) = \\ &P(D_{Pa_{IS}(X)\neq j} | Pa_{pop}(X) \rightarrow X) = \prod_{i=1}^{q} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + N_i)} \cdot \prod_{k=1}^{r} \frac{\Gamma(\alpha_{ik} + N_{ik})}{\Gamma(\alpha_{ik})}\,, \end{aligned} \tag{7}$$

where r and q are the number of possible instantiations of X and $Pa_{pop}(X)$, respectively. N_{ik} is the number of instances in $D_{Pa_{IS}(X)\neq j}$ for which X takes the value k and its parents $Pa_{pop}(X)$ take value i, and $N_i = \sum_{k=1}^{r} N_{ik}$. The terms α_{ik} and $\alpha_i = \sum_{k=1}^{r} \alpha_{ik}$ are the corresponding Dirichlet priors. We define the parameter priors as follows:

$$\alpha_{jk} = \alpha_{ik} = \frac{\alpha}{r \cdot (q+1)}\,, \tag{8}$$

where $q+1$ is the total number of possible configurations for X's parents in both the instance-specific model (where the number of configurations is equal to 1) and the population-wide model (where there are q configurations). α is a positive constant called the PESS (see Section 3.2).

The overall score for node X is given as the product of the instance-specific score and the population-wide score for X:

$$\begin{aligned} &\text{score}_{overall}(X) = \\ &\quad \text{score}_{IS}(D_{Pa_{IS}(X)=j}, Pa_{IS}(X) \rightarrow X) \cdot \ \text{score}_{pop}(D_{Pa_{IS}(X)\neq j}, Pa_{pop}(X) \rightarrow X)\,. \end{aligned} \tag{9}$$

This score represents the marginal likelihood of X given the instance-specific and population-wide parents of X. Algorithm 2 shows pseudo-code for the IS-Score procedure that derives this marginal likelihood as the overall score for X. It is this procedure that GES2 calls when scoring a node given its parents during forward and backward greedy search.

Algorithm 2 IS-Score(D, T, X, $Pa_{IS}(X)$, $Pa_{pop}(X)$)

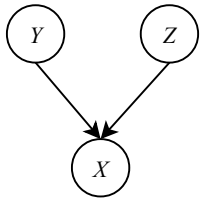
Input: dataset D, instance T , variable X that is being scored, X's instance-specific parent set $Pa_{IS}(X)$, and X's population-wide parent set $Pa_{pop}(X)$

Output: the overall score for X

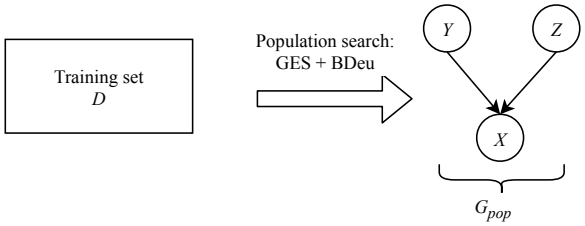
1: derive $D_{Pa_{IS}(X)=j}$ and $D_{Pa_{IS}(X)\neq j}$ from D and the values j of $Pa_{IS}(X)$ in T
2: $\text{score}_{IS} \leftarrow \text{score}_{IS}(D_{Pa_{IS}(X)=j}, Pa_{IS}(X) \rightarrow X)$ ▷ Equation (6)
3: $\text{score}_{pop} \leftarrow \text{score}_{pop}(D_{Pa_{IS}(X)\neq j}, Pa_{pop}(X) \rightarrow X)$ ▷ Equation (7)
4: $\text{score}_{overall} \leftarrow \text{score}_{IS} \cdot \text{score}_{pop}$ ▷ Equation (9)
5: return $\text{score}_{overall}$

Figure 1 shows an example of the IGES procedure. Let Figure 1a represent the ground-truth BN structure and parameters for variable X. In the large sample limit, by applying GES with the BDeu score we expect to learn G_{pop} (Figure 1b), which is the same as the ground-truth structure. However, G_{pop} does not capture the independence of Z and X when $Y = 0$ in the current instance $T = \{X = 1, Y = 0, Z = 1\}$. Figure 1c shows the instance-specific BN (G_{IS}) and the population-wide model (G_{pop}) that would be learned by IGES, in the large sample limit.

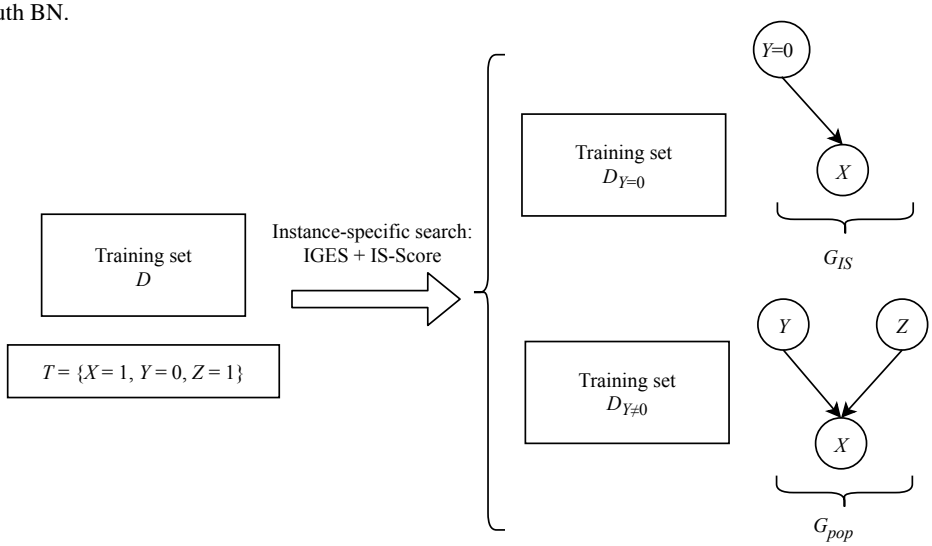
Y Z	$X=0$	$X=1$
0 0	0.9	0.1
0 1	0.9	0.1
1 0	0.23	0.77
1 1	0.52	0.48



(a) The entries in the table on the left represent $P(X|Y,Z)$ and the graph on the right shows the ground-truth BN.



(b) The result of applying GES to the example in the large sample limit.



(c) The results of applying IGES to the example in the large sample limit.

Figure 1: This example illustrates a situation in which the population-wide BN structure learning is not capable of capturing context-specific independence while the instance-specific approach is.

As mentioned, IS-Score derives the marginal likelihood of the data on X, relative to the instance-specific and population-wide parents of X. Assuming parameter independence and parameter modularity (Heckerman et al., 1995) as is commonly done, the marginal likelihood of all the data given T, G_{IS}, and G_{pop} is as follows:

$$P(D|T, G_{IS}, G_{pop}) = \prod_{i=1}^{n} \text{IS-Score}(D, T, X_i, Pa_{IS}(X_i), Pa_{pop}(X_i)) \,, \tag{10}$$

where i iterates over the set of all nodes being modeled. This equation will be used later in Equation (13) to derive an overall BN structure score.

We can also define modular structure priors that are decomposable at the node level to be applied when scoring the parent-child relationship for each node. We use the following structure priors when applying GES to learn the population-wide model (Ramsey et al., 2017):

$$P(G_{pop}) = \prod_{i=1}^{n} \left(\frac{e}{n-1}\right)^{|Pa_{pop}(X_i)|} \cdot \left(1 - \frac{e}{n-1}\right)^{n-1-|Pa_{pop}(X_i)|} , \tag{11}$$

where i iterates over the set of all nodes in G_{pop}, $|Pa_{pop}(X_i)|$ is the number of parents of node X_i in G_{pop}, and e is a prior weight, which we set to be $e = 1$ in this paper. For this structure prior, each node being a parent of another node is modeled as a Bernoulli trial.

To compute the prior probabilities of the instance-specific BN structure G_{IS}, we modify the modular structure prior introduced in (Heckerman et al., 1995) by considering G_{pop} as the prior network:

$$P(G_{IS}) = c \prod_{i=1}^{n} \kappa^{\delta_i} , \tag{12}$$

where c is a normalization constant, i iterates over the set of all nodes, δ_i is the arc difference between instance-specific parents of X_i in G_{IS} (i.e., $Pa_{IS}(X_i)$) and its population-wide parents in G_{pop} (i.e., $Pa_{pop}(X_i)$), and $0 < \kappa \leq 1$ is a penalty factor. We combine Equations (10), (11), and (12) to derive a probability that is proportional to the posterior probability of G_{IS} and G_{pop}:

$$P(G_{IS}, G_{pop}|D, T) \propto P(D|T, G_{IS}, G_{pop}) \cdot P(G_{IS}) \cdot P(G_{pop}) . \tag{13}$$

5. Experimental Results

In this section we investigate the performance of the IGES instance-specific structure discovery algorithm versus a state-of-the-art population-wide method, GES. We applied these two algorithms on both real and simulated datasets. To generate simulated data, we applied the following steps:

- First, we generated random BNs with $V = 50$ nodes and either $E = 50$ or $E = 100$ edges.
- We then parametrized the Bayesian networks to include context-specific independence in the conditional probability tables. We parametrized the CPTs so that each node that has more than one parent includes at least one CSI. In the BNs of size $V = 50$ nodes and $E = 50$ edges, about 15% of the variables (on average) exhibit CSI in each simulated test case T. When we double the density of BNs (i.e., $V = 50$ and $E = 100$) about 30% of the variables (on average) exhibit CSI in each simulated test case T.
- Given the randomly generated BN and its parameters, we simulated a training dataset with 1000 samples. This is dataset D.
- We generated 500 test instances. Each test instance is a case T. The 1000 samples generated in previous step along with each of the 500 test instances are used to learn 500 instance-specific BN structures for each test case T using the proposed IGES algorithm.

We repeated the above steps 10 times and computed the average of the evaluation measures over those runs. We used edge adjacency precision, recall, and F-measure and arrowhead precision, recall, and F-measure (see below) as the primary evaluation measures. We derived specific subtypes of precision, recall, and F-measure for the subset of the nodes that include CSI (P_{IS}, R_{IS}, and F_{IS}), the remaining nodes that do not include CSI (P_{other}, R_{other}, and F_{other}), and over all nodes ($P_{overall}$, $R_{overall}$, and $F_{overall}$). The gold-standard for each node is therefore either an instance-specific structure (which can vary with the instance) or population wide (which does not vary). In the next paragraph, we define how we calculated each of these measures, which are computed at the node level.

Method	P_{IS}	P_{other}	$P_{overall}$	R_{IS}	R_{other}	$R_{overall}$	F_{IS}	F_{other}	$F_{overall}$
IGES ($k = 0.001$)	0.75(0.11)	0.99(0.02)	0.92(0.03)	0.69(0.17)	0.75(0.06)	0.73(0.08)	0.71(0.14)	0.85(0.04)	0.81(0.06)
IGES ($k = 0.1$)	0.81(0.12)	0.98(0.02)	0.93(0.04)	0.72(0.09)	0.75(0.06)	0.74(0.06)	0.76(0.10)	0.85(0.04)	0.82(0.04)
IGES ($k = 0.5$)	0.74(0.13)	0.89(0.04)	0.84(0.05)	0.67(0.11)	0.72(0.08)	0.71(0.07)	0.70(0.11)	0.79(0.06)	0.77(0.05)
IGES ($k = 0.9$)	0.61(0.11)	0.78(0.08)	0.72(0.08)	0.70(0.08)	0.76(0.07)	0.75(0.06)	0.65(0.08)	0.77(0.07)	0.74(0.06)
GES	0.57(0.07)	0.98(0.02)	0.84(0.04)	0.78(0.08)	0.81(0.08)	0.81(0.08)	0.66(0.07)	0.89(0.05)	0.82(0.04)

(a) Adjacency precision (P), recall (R), and F-measure (F)

Method	P_{IS}	P_{other}	$P_{overall}$	R_{IS}	R_{other}	$R_{overall}$	F_{IS}	F_{other}	$F_{overall}$
IGES ($k = 0.001$)	0.23(0.17)	0.91(0.15)	0.72(0.13)	0.44(0.30)	0.60(0.13)	0.59(0.13)	0.29(0.21)	0.71(0.13)	0.64(0.12)
IGES ($k = 0.1$)	0.35(0.17)	0.85(0.13)	0.77(0.14)	0.45(0.18)	0.55(0.08)	0.54(0.07)	0.38(0.17)	0.66(0.08)	0.63(0.08)
IGES ($k = 0.5$)	0.36(0.20)	0.70(0.16)	0.67(0.16)	0.56(0.24)	0.55(0.16)	0.55(0.14)	0.42(0.21)	0.60(0.13)	0.59(0.12)
IGES ($k = 0.9$)	0.22(0.10)	0.55(0.14)	0.49(0.13)	0.50(0.22)	0.59(0.06)	0.58(0.07)	0.30(0.13)	0.56(0.09)	0.52(0.09)
GES	0.16(0.07)	0.81(0.17)	0.58(0.14)	0.68(0.20)	0.68(0.08)	0.68(0.08)	0.24(0.08)	0.73(0.10)	0.61(0.09)

(b) Arrowhead precision (P), recall (R), and F-measure (F)

Method	added	deleted	reversed	log likelihood ratio
IGES ($\kappa = 0.001$)	2.38(0.94)	10.37(2.42)	4.45(2.17)	234.31(146.26)
IGES ($\kappa = 0.1$)	2.25(1.32)	10.09(2.07)	4.17(1.68)	397.47(116.00)
IGES ($\kappa = 0.5$)	5.15(1.71)	10.87(2.26)	3.95(1.83)	438.62(237.32)
IGES ($\kappa = 0.9$)	11.74(4.12)	9.82(2.32)	6.09(2.52)	524.60(229.72)
GES	6.09(1.60)	7.60(2.80)	7.12(2.96)	-

(c) Structural Hamming distance and log likelihood ratio

Table 1: Results for BNs with $V = 50$ nodes and $E = 50$ edges. The numbers in parentheses are standard deviations.

Let G_{output} be G_{IS} when using the IGES algorithm[2] and be G_{pop} when using GES. Also, let G_{truth} be the gold-standard BN structure for a given instance T. To compute precision and recall measures, we first calculated four basic statistics: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Precision was then derived as the ratio $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and recall as the ratio $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. Adjacency precision and recall are defined as follows. TP is the number of adjacencies that are common in both G_{output} and G_{truth} without considering the edge orientation. FP is the number of adjacencies that are present in G_{output} but not in G_{truth}, and FN is the number of adjacencies that are present in G_{truth} but not in G_{output}. For arrowhead precision and recall, TP is the number of edges that are common in both G_{output} and G_{truth} and share the same edge orientation (e.g., $X \to Y$ in both BNs). False orientation in this case would be when $X \to Y$ is present in one BN but there is $Y \to X$, X — Y, or no edge between X and Y in the other one. The F-measure is computed as the ratio $2\mathrm{PR}/(\mathrm{P}+\mathrm{R})$.

Tables 1a and 2a show the adjacency precision and recall for the BNs with $(V = 50, E = 50)$ and $(V = 50, E = 100)$, respectively. Tables 1b and 2b show the arrowhead precision and recall for the BNs with $(V = 50, E = 50)$ and $(V = 50, E = 100)$, respectively. As the tables indicate, using IGES often results in higher precision but lower recall than using the GES search, especially for the nodes with CSI. The F-measure of IGES versus GES is generally higher for F_{IS}, lower for F_{other}, and comparable for $F_{overall}$. In most cases, $\kappa = 0.1$ gives the best results for the IGES method.

We also computed the edge difference (i.e., the structural Hamming distance) to compare performance of the search procedures on each given instance T. The structural Hamming distance for each G_{output} compared to G_{truth} is composed of three edge modifications: added, deleted, and reversed edges. Tables 1c and 2c show the average results on 500 test cases for the BNs with $(V = 50, E = 50)$ and $(V = 50, E = 100)$, respectively. In these experiments, IGES results in less

2. IGES outputs both G_{IS} and G_{pop} for completeness, but G_{IS} is what it actually learns as the instance-specific BN structure for a given instance T.

Method	P_{IS}	P_{other}	$P_{overall}$	R_{IS}	R_{other}	$R_{overall}$	F_{IS}	F_{other}	$F_{overall}$
IGES ($k = 0.001$)	0.62(0.06)	0.96(0.03)	0.79(0.05)	0.61(0.06)	0.65(0.07)	0.63(0.06)	0.61(0.05)	0.77(0.05)	0.70(0.05)
IGES ($k = 0.1$)	0.66(0.06)	0.95(0.03)	0.82(0.04)	0.57(0.08)	0.66(0.05)	0.63(0.04)	0.61(0.06)	0.78(0.03)	0.71(0.03)
IGES ($k = 0.5$)	0.63(0.06)	0.89(0.04)	0.76(0.03)	0.58(0.04)	0.64(0.04)	0.61(0.03)	0.60(0.04)	0.74(0.03)	0.68(0.03)
IGES ($k = 0.9$)	0.55(0.08)	0.79(0.06)	0.68(0.06)	0.57(0.08)	0.67(0.04)	0.63(0.03)	0.56(0.08)	0.73(0.03)	0.66(0.04)
GES	0.55(0.07)	0.94(0.05)	0.73(0.05)	0.69(0.07)	0.71(0.06)	0.70(0.05)	0.61(0.07)	0.81(0.04)	0.72(0.04)

(a) Adjacency precision (P), recall (R), and F-measure (F)

Method	P_{IS}	P_{other}	$P_{overall}$	R_{IS}	R_{other}	$R_{overall}$	F_{IS}	F_{other}	$F_{overall}$
IGES ($k = 0.001$)	0.36(0.10)	0.78(0.15)	0.59(0.12)	0.50(0.13)	0.50(0.09)	0.50(0.09)	0.41(0.09)	0.60(0.09)	0.53(0.08)
IGES ($k = 0.1$)	0.37(0.08)	0.74(0.14)	0.60(0.11)	0.46(0.07)	0.50(0.05)	0.49(0.05)	0.40(0.05)	0.59(0.07)	0.54(0.07)
IGES ($k = 0.5$)	0.38(0.08)	0.65(0.10)	0.56(0.08)	0.45(0.13)	0.50(0.09)	0.49(0.08)	0.41(0.08)	0.56(0.09)	0.52(0.07)
IGES ($k = 0.9$)	0.30(0.07)	0.57(0.08)	0.49(0.07)	0.45(0.11)	0.54(0.06)	0.53(0.05)	0.36(0.08)	0.56(0.06)	0.51(0.06)
GES	0.24(0.06)	0.76(0.13)	0.52(0.09)	0.59(0.11)	0.61(0.08)	0.61(0.07)	0.34(0.08)	0.67(0.08)	0.56(0.07)

(b) Arrowhead precision (P), recall (R), and F-measure (F)

Method	added	deleted	reversed	log likelihood ratio
IGES ($\kappa = 0.001$)	10.25(3.29)	22.24(3.97)	9.68(2.21)	207.54(190.36)
IGES ($\kappa = 0.1$)	8.70(2.04)	22.80(2.26)	8.37(2.84)	518.43(235.79)
IGES ($\kappa = 0.5$)	12.40(2.26)	24.16(2.69)	8.75(2.18)	638.72(274.38)
IGES ($\kappa = 0.9$)	18.99(4.34)	23.01(1.84)	9.41(2.72)	615.99(232.87)
GES	16.17(4.03)	18.60(2.39)	11.23(3.79)	-

(c) Structural Hamming distance and log likelihood ratio

Table 2: Results for BNs with $V = 50$ nodes and $E = 100$ edges. The numbers in parentheses are standard deviations.

erroneously added and reversed edges but more deleted edges. However, the overall average edge error is lower using IGES.

We also calculated the log likelihood ratio as another performance metric. For each instance T, it is calculated as follows:

$$\text{log likelihood ratio} = \log \frac{P(D|T, G_{IS}, G_{pop})}{P(D|T, G_{pop})}, \tag{14}$$

where we use Equation (10) to compute $P(D|T, G_{IS}, G_{pop})$ in this equation. We also score the denominator $P(D|T, G_{pop})$ using Equation (10) but in this case there is only one model (i.e., G_{pop}) that will be used for all instances. The log likelihood ratio of 0 indicates the algorithms produce models that have the same marginal likelihood (ML) for a given dataset, a positive value indicates that IGES produces a higher ML, and a negative value indicates that GES produces a larger ML. IGES results in higher log likelihood ratios compared to GES, as shown in Tables 1c and 2c.

We also evaluated the proposed IGES method on a real chronic pancreatitis dataset. The dataset we used was collected as part of the multicenter North American Pancreatitis Study 2 (NAPS2) (Whitcomb et al., 2008). This data consists of 2201 individuals, of whom 980 developed chronic pancreatitis and 1221 were healthy. We discarded data of 2 individuals who had missing values. We split the remainder of the dataset randomly into a training set of 1761 individuals (80%) and a test set of 438 individuals (20%) while preserving the disease distribution. For each individual, the dataset contains 143 variables, of which 142 are single nucleotide variants (SNVs) and one is a binary outcome variable that denotes if the individual developed chronic pancreatitis. Each SNV is a location on the human genome and for each individual takes one of three possible values. To evaluate the performance of IGES versus GES, we computed the log likelihood ratio for the data given the instance-specific model, G_{IS}, and the population-wide model G_{pop}, similar to (Chickering et al., 1997). The results in Table 3 shows that the IGES method had much higher average log likelihoods.

	$\kappa = 0.001$	$\kappa = 0.1$	$\kappa = 0.5$	$\kappa = 0.9$
log likelihood ratio	3323.18(1344.47)	3408.66(1354.24)	3327.98(1212.82)	3116.33(1086.27)

Table 3: Average log likelihood ratio on the test data sampled from the chronic pancreatitis dataset for different κ values. The numbers in parentheses represent standard deviations.

6. Discussion

This paper introduces a Bayesian instance-specific structure learning algorithm called IGES that outputs a Bayesian network structure that is specific to a given instance T (e.g, a patient) by guiding the search based on T's attributes. Although we applied GES-style algorithm in this paper, the proposed method is quite general and can be adopted to any other score-based search method.

The results on simulated data indicate that IGES performs better in terms of adjacency and arrowhead precision (especially when a node exhibits CSI) for discovering the instance-specific BN structure of each test instance T. However, the recall decreases due to more edges being deleted when applying IGES. The structural Hamming distance is lower on average when using IGES (the lower the better). The log likelihood ratio always improves when using IGES for both real and simulated data. A higher log likelihood ratio suggests that the BN structures learned by IGES are more probable and better model the relationships among variables for each instance T.

The IGES method can be extended in numerous ways, including the following: (a) understand better the reason for the relatively lower recall of the instance-specific BN models and try to increase it while retaining precision; (b) extend the IGES algorithm to iteratively learn an instance-specific model for each instance in the training set and use an aggregate of those models to define the population-wide model; (c) attempt to prove that IGES is guaranteed to find the data-generating instance-specific causal model for a test instance in the large sample limit; (d) develop an instance-specific score to learn BN structures that contain other types of variables (e.g., continuous or a mixture of continuous and discrete variables); (e) develop more informative structure and parameter prior probabilities; (f) extend the experimental evaluations. Despite its limitations, the current paper provides support that the proposed IGES method is a promising approach to discover a BN structure that better models the relationships among variables of a given instance T, rather than a population-wide model. The results suggest that further investigation of the approach is warranted.

Acknowledgements

The research reported in this paper was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. This research was also supported by grant #4100070287 from the Pennsylvania Department of Health (DOH). The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Pennsylvania DOH. We thank the PGM reviewers for helpful comments.

References

C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, pages 115–123. Morgan Kaufmann Publishers Inc., 1996.

D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 87–98. Morgan Kaufmann

Publishers Inc., 1995.

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

D. M. Chickering, D. Heckerman, and C. Meek. A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 80–89. Morgan Kaufmann Publishers Inc., 1997.

G. Cooper, C. Cai, and X. Lu. Tumor-specific causal inference (TCI): A Bayesian method for identifying causative genome alterations within individual tumors. *bioRxiv*, (225631), 2018.

N. Friedman and M. Goldszmidt. Learning Bayesian networks with local structure. In *Learning in Graphical Models*, pages 421–459. Springer, 1998.

D. Heckerman. A tutorial on learning with Bayesian networks. In *Learning in Graphical Models*, pages 301–354. Springer, 1998.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

P. Illari, F. Russo, and J. Williamson. *Causality in the Sciences*. OUP Oxford, 2011.

F. Jabbari, J. Ramsey, P. Spirtes, and G. Cooper. Discovery of causal models that contain latent variables through Bayesian scoring of independence constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 142–157. Springer, 2017.

C. J. Oates, J. Q. Smith, S. Mukherjee, and J. Cussens. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, 26(4):797–811, 2016.

J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

J. Pensar, H. Nyman, T. Koski, and J. Corander. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. *Data Mining and Knowledge Discovery*, 29(2):503–533, 2015.

J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2): 121–129, 2017.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

D. C. Whitcomb, D. Yadav, S. Adam, R. H. Hawes, R. E. Brand, M. A. Anderson, M. E. Money, P. A. Banks, M. D. Bishop, J. Baillie, et al. Multicenter approach to recurrent acute and chronic pancreatitis in the United States: the North American Pancreatitis Study 2 (NAPS2). *Pancreatology*, 8(4):520–531, 2008.

Y. Zou, J. Pensar, and T. Roos. Representing local structure in Bayesian networks by Boolean functions. *Pattern Recognition Letters*, 95:73–77, 2017.